Methodos Series 14

S. Michael Gaddis *Editor*

Audit Studies: Behind the Scenes with Theory, Method, and Nuance

Springer

Methodos Series

Methodological Prospects in the Social Sciences

Volume 14

This Book Series is devoted to examining and solving the major methodological problems social sciences are facing. Take for example the gap between empirical and theoretical research, the explanatory power of models, the relevance of multilevel analysis, the weakness of cumulative knowledge, the role of ordinary knowledge in the research process, or the place which should be reserved to "time, change and history" when explaining social facts. These problems are well known and yet they are seldom treated in depth in scientific literature because of their general nature.

So that these problems may be examined and solutions found, the series prompts and fosters the settingup of international multidisciplinary research teams, and it is work by these teams that appears in the Book Series. The series can also host books produced by a single author which follow the same objectives. Proposals for manuscripts and plans for collective books will be carefully examined.

The epistemological scope of these methodological problems is obvious and resorting to Philosophy of Science becomes a necessity. The main objective of the Series remains however the methodological solutions that can be applied to the problems in hand. Therefore the books of the Series are closely connected to the research practices.

More information about this series at http://www.springer.com/series/6279

S. Michael Gaddis
Editor

Audit Studies: Behind the Scenes with Theory, Method, and Nuance

Springer

Editor
S. Michael Gaddis
Department of Sociology
University of California – Los Angeles
Los Angeles, CA, USA

Methodos Series
ISBN 978-3-319-71152-2 ISBN 978-3-319-71153-9 (eBook)
https://doi.org/10.1007/978-3-319-71153-9

Library of Congress Control Number: 2017963885

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Contributors

Stijn Baert Ghent University, Ghent, Belgium

Ryan Beasley SimQuest Solutions, Inc., Boston, MA, USA

Marc Bendick Jr. Bendick and Egan Economic Consultants, Inc., Alexandria, VA, USA

Max Besbris Rice University, Houston, TX, USA

William Carbonaro University of Notre Dame, Notre Dame, IN, USA

Frances Cherry Department of Psychology, Carleton University, Ottawa, Canada

Charles Crabtree University of Michigan, Ann Arbor, MI, USA

Jacob William Faber New York University, New York, NY, USA

S. Michael Gaddis Department of Sociology, University of California – Los Angeles, Los Angeles, CA, USA

Sarah Lageson School of Criminal Justice, Rutgers University, Newark, NJ, USA

Joanna Lahey Texas A&M University and NBER, College Station, TX, USA

David S. Pedulla Department of Sociology, Stanford University, Stanford, CA, USA

Peter Rich Cornell University, Ithaca, NY, USA

Jonathan Schwarz University of Notre Dame, Notre Dame, IN, USA

Patrick Sharkey New York University, New York, NY, USA

Christopher Uggen Department of Sociology, University of Minnesota, Minneapolis, MN, USA

Mike Vuolo Department of Sociology, The Ohio State University, Columbus, OH, USA

Part I
The Theory Behind and History of Audit Studies

Chapter 1
An Introduction to Audit Studies in the Social Sciences

S. Michael Gaddis

Abstract An audit study is a specific type of field experiment primarily used to test for discriminatory behavior when survey and interview questions induce social desirability bas. In this chapter, I first review the language and definitions related to audit studies and encourage adoption of a common language. I then discuss why researchers use the audit method as well as when researchers can and should use this method. Next, I give an overview of the history of audit studies, focusing on major developments and changes in the overall body of work. Finally, I discuss the limitations of correspondence audits and provide some thoughts on future directions.

Keywords Audit studies · Correspondence audits · Discrimination · Field experiments

1.1 Introduction

Since the 1960s, researchers have had a methodological tool at their disposal unlike any other: the audit study.[1] The audit study is a specific type of field experiment that permits researchers to examine difficult to detect behavior, such as racial and gender discrimination, and decision-making in real-world scenarios. Audit studies allow researchers to make strong causal claims and explore questions that are often difficult or impossible to answer with observational data. This type of field experiment has exploded in popularity in recent years, particularly to examine different types of discrimination, due to the rise of online applications for housing and employment and easy access to decision makers across many contexts via email.

[1] These types of studies are known by a variety of names, often depending on the decade of publication, the context and method used for testing, discipline, or country. Audits are also sometimes referred to as correspondence tests or situation tests. For now, I refer to all this research as "audit studies." Later in this chapter, I define and clarify these terms.

S. M. Gaddis (✉)
Department of Sociology, University of California – Los Angeles, Los Angeles, CA, USA
e-mail: mgaddis@soc.ucla.edu

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_1

However, the learning curve for designing and implementing these experiments can be quite steep, despite appearing to be a simple and quick method for examining discrimination. Thus, we have written this book to help scholars design, conduct, and analyze their own audits. This book draws upon the knowledge of a variety of social scientists and other experts who combined have implemented dozens of in-person and correspondence audits to examine a variety of research questions. These experienced scholars share insights from both their successes and failures and invite you, the reader, "behind the scenes" to examine how you might construct your own audit study and improve upon this method in the future. We write this book with a wide audience in mind and hope that you will find this book useful whether you have already fielded your own audit study, are just thinking about how you might design an audit study, or just want to learn more about the method to better understand research using audits.

In this introductory chapter, I approach the subject as one might with a lay audience. However, even experienced researchers with in-depth knowledge of the audit method should find this chapter useful. I mostly focus on the aspects of audit studies related to research rather than those related to activism or law and policy.[2] I begin this chapter with the basics – a discussion of the language and definitions related to audit studies. Significant differences in language persist between studies, researchers, and disciplines, and I hope that this part will help readers understand these differences as well as encourage researchers to adopt a common language. Next, I give a succinct overview of why researchers began using audits to examine discrimination. The audit method is a powerful tool to answer certain types of questions and I attempt to outline when researchers can and should use this method. I then give an overview of the history of audit studies. Although others have written superb reviews of this body of literature in the past (Baert, Chap. 3 of this volume; Oh and Yinger 2015; Riach and Rich 2002; Zschirnt and Ruedin 2016), I focus on the forest rather than the trees in this part and provide a narrative of the arc of audit studies over time.[3] Finally, I close this chapter with a succinct discussion of the limitations of correspondence audits and thoughts on how we might improve this method, which complements the closing chapter of this book (Pedulla, Chap. 9 of this volume).

Readers looking for additional information on audit studies should consult two resources. First, we have created a website – www.auditstudies.com – to go along with the release of this volume. There you will find a comprehensive database of audits, information about subscribing to an audit method listserv, as well as additional information. Second, at the end of this chapter I provide a brief recommended reading list of important comprehensive works, reviews, and other methods-based articles and books.

[2] For an excellent chapter on the connections to activism, see Cherry and Bendick (Chap. 2 of this volume) and for an excellent, although a bit outdated, chapter on the links between audits and law and policy, see Fix et al. (1993).

[3] Some of the work in this section stems from and expands upon work I did to examine the signals of race conveyed by names in correspondence audits (Gaddis 2017a, b, c, d).

Beyond this introductory chapter, several accomplished scholars present their expert knowledge about audit studies. In the first part – The Theory Behind and History of Audit Studies – the authors cover a wide range of history, explain why we should conduct audit studies, examine the connections between audit studies and activism, and outline what researchers have uncovered about labor market processes using audit studies in the past decade. In the second part – The Method of Audit Studies: Design, Implementation, and Analysis – the experts provide guidance on designing your own audit study, discuss the challenges and best practices regarding email, review extensive issues of validity, and consider the technical setup of matching procedures. In the final part – Nuance in Audit Studies: Context, Mechanisms, and the Future – the authors focus on more nuanced aspects of audit studies and address limitations and challenges, examine the use of context to explore mechanisms, and consider the value of variation. I return to a brief discussion of the rest of this book at the end of this chapter.

1.2 The Basics of Audit Studies: Language and Definitions

Field experiments encompass a wide range of studies and ideas and describe the highest level of the hierarchy I focus on here. Audit studies are one type of field experiment. At their core, field experiments in the social sciences attempt to mimic the experiments of the natural sciences by implementing a randomized research design in a field setting (as opposed to a lab or survey setting). Although many may think of psychology as the disciplinary home to social science experiments, researchers in economics, political science, and sociology have ramped up the quantity and quality of field experiments conducted in these disciplines over the past few decades. Although not the only reason for the increase in field experiments across these disciplines, audit studies do represent a major part of the heightened activity.

Audit studies generally refer to a specific type of field experiment in which a researcher randomizes one or more characteristics about individuals (real or hypothetical) and sends these individuals out into the field to test the effect of those characteristics on some outcome. Historically, audit studies have focused on race and ethnicity (Daniel 1968; Bertrand and Mullainathan 2004; Wienk et al. 1979) and gender (Ayres and Siegelman 1995; Levinson 1975; Neumark et al. 1996). In recent years, researchers have expanded the manipulated characteristics to include age (Ahmed et al. 2012; Bendick et al. 1997; Farber et al. 2017; Lahey 2008; Neumark et al. 2016; Riach 2015; Riach and Rich 2010), criminal record (Baert and Verhofstadt 2015; Evans 2016; Evans and Porter 2015; Furst and Evans 2016; Pager 2003), disability (Ameri et al. forthcoming; Baert 2014a; Ravaud et al. 1992; Turner et al. 2005; Verhaeghe et al. 2016), educational credentials (Carbonaro and Schwarz, Chap. 7 of this volume; Darolia et al. 2015; Deming et al. 2016; Deterding and Pedulla 2016; Gaddis 2015, 2017e; Jackson 2009), immigrant assimilation or generational status (Gell-Redman et al. 2017; Ghoshal and Gaddis 2015; Hanson and Santas 2014), mental health (Baert et al. 2016a), military service (Baert and Balcaen

2013; Figinski 2017; Kleykamp 2009), parental status (Bygren et al. 2017; Correll et al. 2007; Petit 2007), physical appearance (Bóo et al. 2013; Galarza and Yamada 2014; Maurer-Fazio and Lei 2015; Patacchini et al. 2015; Ruffle and Shtudiner 2015; Stone and Wright 2013), religious affiliation (Adida et al. 2010; Pierné 2013; Wallace et al. 2014; Wright et al. 2013), sexual orientation (Ahmed et al. 2013; Baert 2014b; Bailey et al. 2013; Drydakis 2009, 2011a, 2014; Mishel 2016; Tilcsik 2011; Weichselbaumer 2015), social class (Heylen and Van den Broeck 2016; Rivera and Tilcsik 2016), and spells of unemployment and part-time employment (Birkelund et al. 2017; Eriksson and Rooth 2014; Kroft et al. 2013; Pedulla 2016), among other characteristics (Baert and Omey 2015; Drydakis 2010; Kugelmass 2016; Tunstall et al. 2014; Weichselbaumer 2016).

The "individuals" sent into the field may be actual people in an in-person audit or simply applicants or emails from hypothetical people in correspondence audits (more below). The outcomes may be an offer to interview for a job (Bertrand and Mullainathan 2004; Darolia et al. 2015; Deming et al. 2016; Gaddis 2015), a job offer (Bendick et al. 1994, 2010; Pager et al. 2009a, b; Turner et al. 1991a), the order in which applicants are contacted (Duguet et al. 2015), a response to a housing inquiry (Ahmed and Hammarstedt 2008; Bengtsson et al. 2012; Carlsson and Ericksson 2014; Carpusor and Loges 2006; Ewens et al. 2014; Feldman and Weseley 2013; Hogan and Berry 2011; Van der Bracht et al. 2015), the types of housing shown (Galster 1990a; Turner et al. 2002, 2013), information about the availability of a house for purchase or rent (Galster 1990b, Turner et al. 2002, 2013; Yinger 1986), an offer of different housing than requested or racial steering (Galster and Godfrey 2005; Turner et al. 1990), a response to a mortgage application or request for information (Hanson et al. 2016; Smith and Cloud 1996; Smith and DeLair 1999), a response to a roommate request (Gaddis and Ghoshal 2015, 2017; Ghoshal and Gaddis 2015), an offer to schedule a doctor's appointment (Kugelmass 2016; Sharma et al. 2015), a response from a politician or other public official (Broockman 2013; Butler and Broockman 2011; Chen et al. 2016; Distelhorst and Hou 2014; Einstein and Glick 2017; Hemker and Rink forthcoming; Janusz and Lajevardi 2016; McClendon 2016; Mendez and Grose 2014; White et al. 2015), a response from a professor (Milkman et al. 2012, 2015; Zhao and Biernat 2017), the price paid or bargained for during economic transactions for goods (Anagol et al. 2017; Ayres 1991; Ayres and Siegelman 1995; Besbris et al. 2015; Doleac and Stein 2013), or a number of other outcomes (Allred et al. 2017; Edelman et al. 2017; Giulietti et al. 2015; Ridley et al. 1989; Wallace et al. 2012; Wissoker et al. 1998; Wright et al. 2015).

Two main variations of audits exist: in-person audits and correspondence audits. In-person audits rely on trained assistants to conduct the experiment. Early audit studies almost exclusively referred to the research subjects posing as legitimate applicants for employment or housing as testers or auditors. This is due, in part, to the fact that the language for such research was adopted from early testing for legal violations for enforcement rather than research purposes (see Boggs et al. 1993 and Fix and Turner 1999 for an in-depth discussion of differences between paired testing for enforcement purposes versus research). However, as correspondence audits overtook in-person audits as the norm and real individuals posing as subjects were

not required, researchers shifted their language to refer to applicants, candidates, constituents, prospective tenants, etc. In other words, the language should match what the audit context dictates. Although the language identifying testers, auditors, or applicants may vary due to the nature of the study, we recommend that researchers adopt a common language of "in-person audits" to identify field cases using live human beings and "correspondence audits" to identify online, telephone, or by mail audits using hypothetical individuals or recorded messages in the case of some audits by telephone.

Although most audit studies include paired (or sometimes triplet) testing with comparisons of two (or three) testers or applicants, not all do (for example, see Hipes et al. 2016; Lauster and Easterbrook 2011; Rivera and Tilcsik 2016). Paired testing, also referred to as matched testing, is a design in which the subject or organization being audited (e.g., employer, real estate agent, etc.) receives applications or emails from two or more of testers with different characteristics. Conversely, non-paired testing is a design in which the subject or organization being audited only ever receives a single tester application or email. For example, a paired test design might send both a black couple and a white couple to each real estate agent's office in the sample whereas a non-paired test design would send only one of the two couples (randomly) to each real estate agent's office in the sample. There can be statistical advantages to paired testing, however, in some cases it may be necessary to implement a non-paired test design to reduce suspicion and avoid experiment discovery (Vuolo et al. 2016, Chap. 6 of this volume; Weichselbaumer 2015, 2016).

1.3 The Need for Audit Studies[4]

Not coincidentally, the rise of audit studies by researchers corresponds with the public policy of the civil rights era aimed to stop racial discrimination and reduce, if not eliminate, racial inequality. Prior to the 1960s, racial discrimination in the United States occurred openly in public, was relatively common, had minimal stigma attached to it, was shaped by open prejudicial attitudes and beliefs, and arguably was informed by a conscious or active racial prejudice. Individual employers, real estate agents, and landlords could discriminate with impunity and often made public their beliefs and actions. In the United States, the Civil Rights Act of 1964 intended to change these behaviors, if not beliefs and attitudes, by outlawing discrimination on the basis of race, color, religion, sex, or national origin. The Equal Employment Opportunity Commission (EEOC) gained the ability to litigate discrimination cases following the passage of the Equal Employment Opportunity Act in 1972. Title VII of the Civil Rights Act of 1964 finally could be enforced.

[4] In this section, I discuss audits from the perspective of racial discrimination. However, the need for and use of audits is similar across other types of discrimination as well as some non-discrimination-based domains of inquiry.

However, we can imagine and, indeed do live in, a world where the Civil Rights Act may have changed the *act* of discrimination without changing the *amount* of discrimination, *intentions* behind discrimination, or an individual's *desire* to discriminate. Although not a sharp change overnight, discrimination of all types has changed in response to the Civil Rights Act. Modern discrimination has become more covert, uncommon, and stigmatized, while being shaped by private prejudicial attitudes and beliefs, and, perhaps, informed by an unconscious or latent racial prejudice. Individuals may fear litigation for engaging in discrimination or have a social desirability bias to not acknowledge discriminatory actions. This makes it difficult for researchers to document and examine discrimination.

Thus, two traditional methods of social science inquiry are difficult, if not impossible, to employ to examine discrimination in the post-civil rights era. First, pointed interviews and survey questions asking perpetrators about racial discrimination are unlikely to elicit truthful responses. To my knowledge, the most recent research project to successfully elicit clearly truthful responses from employers about engaging in racial discrimination occurred in the late 1980s (Kirschenman and Neckerman 1991). Moreover, surveys and interviews do not document actions, but rather self-reported beliefs, attitudes, recollections of past actions, or predictions of future actions. Due to respondents' fear and social desirability bias, and the sometimes unconscious nature of racial prejudice, direct questions about discrimination through interviews and surveys exhibit low construct validity.

Second, statistical analyses using secondary data that do not have explicit questions about discrimination also fail to adequately capture discrimination. To understand the difficulty of this process, let's first consider a definition of discrimination. In a 2004 book stemming from the Committee on National Statistics' Panel on Methods for Assessing Discrimination, panelists defined racial discrimination as "differential treatment on the basis of race that disadvantages a racial group" (Blank et al. 2004: 39). Although researchers can document the second (race) and third parts (disadvantage) of the definition with secondary data, directly capturing the first part (differential treatment) is impossible. Thus, secondary data analysis must use indirect residual attribution to suggest that, after including a litany of control variables that affect the dependent variable of interest on which blacks and whites differ, any remaining coefficient for race represents discrimination (Blank et al. 2004; Lucas 2008; Neumark forthcoming). However, this method is unlikely to correctly attribute the true amount of racial discrimination (Quillian 2006), due to omitted variable bias, among other issues (Altonji and Blank 1999; Blank et al. 2004; Farkas and Vicknair 1996; Lucas 2008).

Researchers developed the audit method as a means of catching individuals and organizations in the act of discrimination. Generally, experiments *can* be done when a presumed cause is manipulable and *should* be done when it is otherwise difficult to prove non-spuriousness. Many, if not all, types of discrimination are great candidates for examination through experimental means because the presumed cause often is manipulable in many contexts and, as discussed earlier, traditional methods of social science inquiry have been unable to directly document discrimination or

rule out a spurious relationship. If we consider the previously stated definition of racial discrimination – "differential treatment on the basis of race that disadvantages a racial group" (Blank et al. 2004: 39) – we see that audit studies manipulate the second part (race) to directly capture the first part (differential treatment) of the definition. Thus, by carefully controlling and counterbalancing all other variables in the experimental process, audit studies provide strong causal evidence of discrimination.

1.4 A History of Audit Studies

1.4.1 The Early Years: The First In-Person and Correspondence Audits

In-person audits began in the 1940s and 1950s by means of activists and private organizations with some assistance from academic researchers. One of the earliest media mentions of audits occurred in the New York Times in 1956 (Rowland). In Chap. 2, Frances Cherry and Marc Bendick Jr. (Chap. 2 of this volume) do an excellent job of covering some of this early work, so I leave discussion of that part of the history of audit studies to them.

The earliest known published audit study of significant scope and scale was conducted in England in the late 1960s. With the Race Relations Acts of 1965, Parliament passed the first legislation addressing racial discrimination in the United Kingdom in public domains. The following year, the U.K. Parliament created the Race Relations Board, which was tasked with reviewing complaints falling under the Race Relations Act. However, the Race Relations Act did not cover employment and housing discrimination until 1968, so in tandem with the National Committee for Commonwealth Immigrants, the Race Relations Board commissioned a study on racial discrimination in employment, housing, and other contexts. Along with surveys and interviews, the study implemented the audit method to extensively examine discrimination (Daniel 1968).

Described as "situation tests," the audits were born when Daniel and the research team had doubts over whether surveys and interviews would give them an accurate portrayal of the state of discrimination. Moreover, the team was unsure if the "findings would appear conclusive to those people who are strongly passionate or committed about the subject on one side or the other" (1968: 20). That doubt led them "not to depend entirely on what people told us in interviews, but to put the matter to the test in a way that would provide objective evidence" (ibid). These tests were conducted with triplets of candidates – usually white English, white immigrant, and black applicants – in the domains of housing (both rental and purchase), employment, and other services. The tests consistently uncovered discrimination against blacks and immigrants.

At the time, this commissioned study of racial discrimination was monumentally important. Along with the hard work of researcher William Wentworth Daniel, results from this study led to the revised Race Relations Act of 1968 outlawing racial discrimination in employment and housing (Smith 2015). However, this study often has been overlooked or forgotten by academics; at the time of this writing, Google Scholar reports that the resulting book by Daniel (1968) has garnered fewer than 500 citations in nearly 50 years. Still, *Racial Discrimination in England*'s use of the audit method in government-sponsored research marks the beginning of a series of high profile in-person audits conducted to examine racial discrimination.

Just a few years later, in 1969, the first-ever correspondence audit was conducted in the United Kingdom. Published by two researchers from the non-profit institute Social and Community Planning Research, this study sought to examine racial discrimination among employers looking to hire white-collar workers (Jowell and Prescott-Clarke 1970). The authors chose to conduct a correspondence audit through the mail because "postal applications were possible and, in many cases, necessary" to apply for employment (1970: 399). The authors matched British-born whites with four different immigrant groups to test for racial discrimination across an ambitious-for-the-time 128 job postings (256 total applicants) and noted the importance of both realism in the application and controlling for all differences between candidates including aspects such as handwriting. Again, although this study has collected few citations in nearly 50 years (fewer than 150 at the time of this writing), it remains an incredibly important entry in the annals of the audit method because it introduced the world to correspondence audits.

1.4.2 The First Wave: The Early 1970s Through the Mid 1980s

In the United States, a number of non-academic-based audits followed the two UK studies. Private fair housing audits rose to prominence in the late 1960s and 1970s in the United States following passage of the Civil Rights Act of 1968 (also known as the Fair Housing Act), which provided federal enforcement of anti-discrimination housing law through an office of the U.S. Department of Housing and Urban Development (HUD). These audits were often conducted in partnership with academic researchers (often local) and often focused on one major city, such as Akron, Ohio (Saltman 1975), Chicago (as reported in Cohen and Taylor 2000), Detroit (Pearce 1979), Los Angeles (Johnson et al. 1971), and New York (as reported in Purnell 2013). Additionally, organizations often produced method-based manuals and guides for the practice of auditing (Kovar 1974; Leadership Council for Metropolitan Open Communities 1975; Murphy 1972).

However, the largest, and arguably most important, audit on housing discrimination during this era, the Housing Market Practices Survey (HMPS), occurred in 1977 (Wienk et al. 1979). This first large-scale housing audit was commissioned by HUD to test for discrimination against blacks in both the sale and rental housing

markets. HUD paired with local fair housing organizations and other organizations to recruit and train testers to conduct the in-person audits. This research included 3264 audits across 40 metro areas, with a plurality of the audits occurring in five metro areas. The HMPS found discrimination against blacks in reported housing availability, treatment by real estate agents, reported terms and conditions, and the types and levels of information requested by real estate agents. This research was critically important in leading the way for future audits, including three additional national housing audits commissioned by HUD (Turner and James 2015; Turner et al. 2002, 2013; Turner et al. 1991b; Yinger 1991, 1993), several smaller local audits (see below), and the Urban Institute employment audits a decade later (Cross et al. 1990; Mincy 1993; Turner et al. 1991a). Arguably, four aspects of the HMPS were important in shaping future audits. First, the HMPS showed that large-scale audits for discrimination in the United States were possible. Second, this research essentially gave auditing a gold seal of approval from an arm of the federal government (for more details on audits and the courts, see Boggs et al. 1993; Fix et al. 1993; Pager 2007a). Third, it was the first research to show the extent to which racial discrimination was widespread across many cities. Finally, the HMPS showed creativity in expanding the outcomes examined by audits.

Other one-off in-person and correspondence audits conducted during the 1970s and early 1980s examined housing and employment discrimination in the United Kingdom (McIntosh and Smith 1974), housing discrimination in France (Bovenkerk et al. 1979) and the United States (Feins and Bratt 1983; Galster and Constantine 1991[5]; Hansen and James 1987; James et al. 1984; Newburger 1984; Roychoudhury and Goodman 1992, 1996[6]), and employment discrimination in the United States (Hitt et al. 1982; Jolson 1974; Levinson 1975; McIntyre et al. 1980; Newman 1978), Canada (Adam 1981; Henry and Ginzberg 1985), Australia (Riach and Rich 1987, 1991), and England (Brown and Gay 1985; Firth 1981; Hubbuck and Carter 1980). Additionally, George Galster (1990a, 1990b) reviewed several fair housing audits conducted in the 1980s that were mostly unpublished and analyzed data from 71 separate audits.

During this period, researchers also began to expand the domains in which they investigated discrimination. As early as 1985, Galster and Constantine (1991) investigated housing discrimination based on parental and relationship status among women. Ayres (1991 and Ayres and Siegelman 1995) examined racial and gender discrimination in bargaining for new car prices, while Ridley et al. (1989) examined racial discrimination in hailing a taxi. Other research from this period examined discrimination based on disability (Fry 1986; Graham et al. 1990; Ravaud et al. 1992). The first wave of audits conducted in the 1970s and 1980s filled in a number of gaps in our knowledge about the extent and geography of discrimination, conditions under which discrimination occurred, and variations in outcomes that were affected by discrimination, particularly in housing and, to some degree, employment.

[5] Conducted in 1985

[6] Conducted throughout the 1980s.

1.4.3 The Second Wave: The Late 1980s Through the Late 1990s

Beginning with the last part of the 1980s and continuing throughout the 1990s, a second wave of audits was ushered in with the second iteration of the HUD housing audit (Turner Micklensons and Edwards 1991; Yinger 1991, 1995) and a series of large-scale employment audits conducted by the Urban Institute (Cross et al. 1990; Mincy 1993; Turner et al. 1991a), in part, aided by guidelines for adapting housing audits to hiring situations (Bendick 1989). The HUD housing audit in 1989, known as the Housing Discrimination Study (HDS) 1989, was conducted in partnership with the Urban Institute. The HDS 1989 varied from and improved on the HMPS in 1977 in many ways. First, the former included Hispanic testers paired with whites for some audits to examine discrimination against Hispanics as well (Ondrich et al. 1998; Page 1995), something that was only done in an extension of the HMPS and only in Dallas (Hakken 1979). Second, in the HDS 1989 auditors focused on specific advertised housing units, whereas in the HMPS auditors approached agents about more general housing options fitting certain criteria. Thus, the HDS 1989 could more accurately examine racial steering. Third, the HDS 1989 examined fewer metro areas (25 instead of 40), but conducted more audits (3800 instead of 3264). Overall, the HDS 1989 replicated the general finding of the HMPS that housing discrimination against blacks was prevalent and widespread. However, there was no strong evidence suggesting that discrimination increased or decreased between the two data collection periods (Elmi and Mickelsons 1991).

The first of the Urban Institute employment audits was conducted in Chicago and San Diego in 1989 and examined discrimination against Hispanics (Cross et al. 1990). Researchers sampled newspaper advertisements and matched pairs successfully applied to almost 300 entry-level jobs in the two cities. The study found that Hispanics faced discrimination at both the application and interview phases, which lead to fewer interviews and fewer job offers when compared with their white counterparts. In 1990, the Urban Institute conducted a similar employment audit in Chicago and Washington, D.C. to examine discrimination against African Americans (Turner et al. 1991a). Matched pairs successfully completed nearly 450 audits in the two cities. The study found that employers discriminated against blacks in accepting their applications, inviting them to interview, and offering them a job. Black applicants were also more likely to be steered toward lower quality jobs rather than the advertised position to which they responded. Additionally, whites were treated more favorably in a number of respects, including waiting time, length of interview, and positive comments.

The Urban Institute studies were the first large-scale true employment audits conducted in the U.S. Researchers and staff went to great lengths to make the study as methodologically sound as possible and paid close attention to detail in sampling, creating matched pairs, and standardizing procedures for the audits (Mincy 1993). Although these studies provided a meticulous model for subsequent researchers to follow when conducting employment audits, others have extensively critiqued the

Urban Institutes studies and the in-person audit method more broadly (Heckman 1998; Heckman and Siegelman 1993). However, by moving development and knowledge of the method forward and by providing extensive guidance (along with Bendick 1989) for the numerous employment audits that followed them, the Urban Institute audits were clearly of great importance.

Following the HDS 1989 and the Urban Institute employment audits, a wave of audits examining employment, housing, and other forms of discrimination occurred. Many audits were conducted in Europe through the International Labour Office (ILO) based on guidelines developed by Frank Bovenkerk (1992). Studies in the U.S. (Bendick et al. 1991, 1994; James and DelCastillo 1992; Nunes and Seligman 1999) and Europe (Arrijn et al. 1998; Bovenkerk et al. 1995; de Prada et al. 1996; Esmail and Everington 1993, 1997; Goldberg et al. 1995; Smeesters and Nayer 1998) focused on race and ethnic discrimination. Researchers conducted sex discrimination employment audits in the U.S. (Neumark et al. 1996; Nunes and Seligman 2000) and Europe (Weichselbaumer 2000), as well as age and disability-based discrimination employment audits in the U.S. (Bendick et al. 1999) and Europe (Graham et al. 1990; Gras et al. 1996). This period also included the continuation of telephone-based (Bendick et al. 1999; Massey and Lundy 2001; Purnell et al. 1999) and written correspondence audits (Bendick et al. 1997; Gras et al. 1996; Weichselbaumer 2000). Still, the cost-prohibitive nature of in-person audits and labor-intensive nature of correspondence audits during the 1990s meant that use of the audit method was relatively rare.

1.4.4 The Third Wave: The Early 2000s Through the Late 2000s

Until the early 2000s, most audits were conducted in-person and relied on trained assistants to physically participate in the process. With housing and employment applications increasingly taking place over the internet, researchers began conducting more correspondence audits. However, some important audits in the early 2000s were still in-person, including the second iteration of HUD and the Urban Institute's Housing Discrimination Study (HDS 2000: Bavan 2007; Ross and Turner 2005; Turner et al. 2002). Devah Pager was the first to examine the effects of a criminal record using an audit study (2003) and produced an incredibly strong body of work during this period consisting of in-person audits as well as examinations of the method (Pager 2007a, b; Pager et al. 2009a, b; Pager and Quillian 2005; Pager and Shepherd 2008).

The 2000s brought about significant changes in the audit method and the importance of this era is highlighted by the fact that the two most cited audit studies of all time both occurred in the early 2000s. Devah Pager's (2003) in-person audit study of race and criminal record in the low-wage labor market in Milwaukee has garnered over 2000 citations according to Google Scholar. Marianne Bertrand and

Sendhil Mullainathan's (2004) correspondence audit study of race in labor markets in Boston and Chicago has over 3100 citations at the time of this writing. Both studies have been incredibly important in shaping our understanding of racial discrimination, however, the differences between them are stark and mark a major turning point in the history of audit studies.

Bertrand and Mullainathan's 2004 study, published in *The American Economic Review*, is the most influential correspondence audit study of the past two decades. In total, the authors applied to over 1300 job advertisements, compared to Pager's 350 jobs (2003), listed in newspapers in Boston and Chicago via fax and mail. Additionally, the authors used birth record data and a small convenience sample pretest to select names to convey race on each resume. Rather than send two applicants per job, the authors often used four resumes to examine both race and resume quality simultaneously and obtained a final sample size of 4870. Bertrand and Mullainathan found that white applicants were about 50% more likely than black applicants to receive a callback. Moreover, black applicants benefited less than white applicants from higher resume quality.

Bertrand and Mullainathan's (2004) landmark study ushered in a new era of correspondence audits. Arguably, this study paved the way for the increase in audits that followed for at least three reasons. First, the research showed that a large-scale audit – in particular, a correspondence audit – could be undertaken by a small team of academic researchers, compared to past audits conducted by larger teams such as those at HUD and the Urban Institute. Although Bertrand and Mullainathan applied via fax and mail, the timing was ripe for the switch to applications over the internet which further expanded the possibilities of correspondence audits. Second, the study opened a dialogue about signaling race through correspondence audits. Because the authors conducted a small pretest and used a moderate number of names – 36 in total – the plurality of studies that followed used the same names to signal race (see Gaddis 2017d).[7] Although over a decade would pass before scholars began to seriously question these signals (Butler and Homola 2017; Gaddis 2017a, b, c, d; Weichselbaumer 2017), Bertrand and Mullainathan were the first to truly investigate them. Finally, this study showed that it was possible to successfully manipulate several characteristics simultaneously. Beyond race and gender, the authors varied other resume characteristics such as education, experience, and skills. These manipulations likely sparked ideas among researchers about mechanisms and interactions that would follow in future studies.

The vast majority of the studies that followed Bertrand and Mullainathan during the 2000s were conducted via the correspondence method. A few notable exceptions are the previously mentioned studies by Devah Pager (2003; Pager et al. 2009a) and three studies carried out by the International Labour Office (ILO) in Italy (Allasino et al. 2004), Sweden (Attström 2007), and France (Cediey and Foroni 2008), although the ILO studies used a mix of in-person and correspondence methods.

[7] Although credit should also be given to Lodder, McFarland, and White (2003) who pre-tested names in a small employment correspondence audit in Chicago before Bertrand and Mullainathan (2004).

Additionally, two in-person studies examined discrimination in market transactions: baseball card sales (List 2004) and auto repair quotes (Gneezy and List 2004).

During this time, correspondence audits examining employment discrimination based on race and ethnicity expanded to cover more countries and race/ethnicities such as Albanians in Greece (Drydakis and Vlassis 2010) and Turks in Germany (Kaas and Manger 2012), and a variety of other groups in Australia (Booth et al. 2012), Canada (Oreopoulos 2011), Denmark (Hjarnø and Jensen 2008), France (Duguet et al. 2010), Great Britain (Wood et al. 2009), Ireland (McGinnity and Lunn 2011), Sweden (Bursell 2007; Carlsson 2010; Carlsson and Rooth 2007; Rooth 2010), and the U.S. (Jacquemet and Yannelis 2012; Thanasombat and Trasviña 2005; Widner and Chicoine 2011). Additionally, researchers examined employment discrimination on the basis of gender and family status in France (Petit 2007) and the U.S. (Correll et al. 2007), gender in England (Riach and Rich 2006a), Spain (Albert et al. 2011) and Sweden (Arai et al. 2016),[8] age in England (Riach and Rich 2010), France (Riach and Rich 2006b), Spain (Albert et al. 2011; Riach and Rich 2007), and the U.S. (Lahey 2008), sexual orientation in Austria (Weichselbaumer 2003), Greece (Drydakis 2009, 2011a) and the U.S. (Tilcsik 2011), race and criminal record in the U.S. (Galgano 2009), race and military status in the U.S. (Kleykamp 2009), educational credentials in the United Kingdom (Jackson 2009), caste in India (Siddique 2011), caste and religion in India (Banerjee et al. 2009), and physical attractiveness and obesity in Sweden (Rooth 2009). One additional study of note during this period is Philip Oreopoulos' correspondence audit in Toronto, which included six different racial/ethnic/immigrant groups. He applied to over 3200 job postings using 13,000 different resumes to create one of the most ambitious correspondence audits of its time.

The expansion of audit research during the 2000s included housing discrimination studies as well. The HDS 2000 expanded to include Asians and Pacific Islanders as well as Native Americans (Turner and Ross 2003a, b) and examined housing discrimination on the basis of disability (Turner et al. 2005). Correspondence audits examined housing discrimination based on race and ethnicity in Canada (Hogan and Berry 2011), Greece (Drydakis 2011b), Italy (Baldini and Federici 2011), Spain (Bosch et al. 2010), Sweden (Ahmed et al. 2010; Ahmed and Hammarstedt 2008), and the United States (Carpusor and Loges 2006; Friedman et al. 2010; Hanson and Hawley 2011; Hanson et al. 2011). Additional research examined housing discrimination on the basis of sexual orientation (Ahmed and Hammarstedt 2008, 2009).

Beyond the major expansion of correspondence audits during this time, the period is marked by the beginning of researchers' exploration of mechanisms of discrimination, intentions behind discrimination, and conditions under which discrimination occurs rather than simply documenting the existence of discrimination. At least four studies during this period attempted to uncover greater detail related to these issues. First, two studies followed up with employers after submitting them to an audit to examine bias in more detail. In one study, Devah Pager and Lincoln Quillian (2005) conducted a telephone survey to follow up with employers who had

[8] Conducted in 2006 and 2007.

unknowingly participated months earlier in an in-person audit study. When given a vignette scenario that mimicked the audit scenario they were subjected to, employers suggested they would be much more likely to hire individuals than the callback rates suggested. In fact, the results of the vignette survey showed no differences between white and black applicants, suggesting the existence of social desirability bias. In another study, Dan-Olof Rooth (2010) administered the Implicit Association Test (IAT) to test whether discriminatory behavior in a prior correspondence audit was associated with IAT scores. He found a strong positive correlation between discrimination against Arab-Muslims[9] and IAT scores but no correlation with a separate explicit measure of bias. These results could suggest that individuals are engaging in discrimination only due to implicit bias (without having a true explicit bias) or could suggest the existence of social desirability bias.

The second set of studies attempted to distinguish between statistical discrimination and taste-based discrimination. In one study, Joanna Lahey (2008) designed a computerized method of creating resumes to examine many values of many variables rather than the often-binary choice sets of resumes prior to her study (see also Lahey and Beasley 2009). Using this revision of the correspondence audit, she could test if employers were less likely to call back older workers due to judgments and assumptions about human capital (statistical discrimination) or due to a general preference for younger workers (taste-based discrimination). She found some evidence for statistical but not taste-based age discrimination. Importantly, her computerized method of creating resumes has also been used to develop several large-scale correspondence audits (e.g., Deming et al. 2016; Oreopoulos 2011). In another study, Leo Kaas and Christian Manger (2012) conducted a correspondence audit in Germany in which they found that Turkish applicants were less likely to receive a callback than German applicants. However, they submitted some applications with two reference letters that included information on personality and work ethic. The authors found that among applications that included these reference letters, there were no statistical differences between the callback rates for German and Turkish applicants, suggesting that employers in Germany engage in statistical discrimination against Turkish applicants. These four studies highlight an important shift in audit studies from simply documenting discrimination to exploring the process in more detail. This trend would continue throughout the following decade and shape the focus and contributions of future audit studies.

1.4.5 The Current Wave: The Early 2010s to Present

Since the early 2010s, the number of audit studies appearing in journals and working paper form has grown exponentially. By my count, the number of audit studies conducted between 2010 and 2017 is already quadruple the number conducted between 2000 and 2009. For that reason alone, it would be incredibly difficult to

[9] Rooth makes a distinction that he is specifically testing the combined category.

cover all of these studies with any detail in this part. With apologies to those not covered here, I focus on what I consider to be the most significant developments during the past 7 years. However, it is also important to note that researchers have continued to expand the domains of study to areas such as healthcare (Kugelmass 2016; Sharma et al. 2015; Shin et al. 2016), politics and public service (Butler and Broockman 2011; Einstein and Glick 2017; Giulietti et al. 2015; Hughes et al. 2017; McClendon 2016; White et al. 2015), religious organizations (Wallace et al. 2012; Wright et al. 2015), eBay and Craigslist transactions (Besbris et al. 2015; Doleac and Stein 2013; Nunley et al. 2011), and new sharing economy market transactions such as Airbnb and Uber (Cui et al. 2017; Edelman et al. 2017; Ge et al. 2016). Additionally, researchers have expanded the countries of study to include Argentina (Bóo et al. 2013), Belgium (Baert 2016; Baert and Verhofstadt 2015), Brazil (de Leon and Kim 2016), China (Maurer-Fazio 2012; Maurer-Fazio and Lei 2015; Zhou et al. 2013), the Czech Republic (Bartoš et al. 2016), Ghana (Michelitch 2015), Israel (Ariel et al. 2015; Ruffle and Shtudiner 2015; Zussman 2013), Malaysia (Lee and Khalid 2016), Mexico (Arceo-Gomez and Campos-Vazquez 2014; Campos-Vazquez and Arceo-Gomez 2015), Norway (Andersson et al. 2012), Peru (Galarza and Yamada 2014, 2017), and Poland (Wysienska-Di Carlo and Karpinski 2014). HUD has also continued to conduct audit studies with a new iteration of the HDS in 2012 (Turner et al. 2013).

I believe there have been at least four major developments in audit research during the most recent period: (1) continued attempts to adjudicate among types of discrimination, (2) an increased focus on context and the conditions under which discrimination occurs, (3) an increased focus on methodological issues in audit design, and (4) the inclusion of additional data from outside the audit itself. These developments are not mutually exclusive; many studies incorporate two or more of these developments.

Adjudicating Among Types of Discrimination

Scholars have long sought to understand the reasons for discrimination and to better adjudicate among types of discrimination (Aigner and Cain 1977; Altonji and Blank 1999; Arrow 1972; Becker 1957; Dymski 2006; Guryan and Charles 2013). Discrimination research has often focused on whether decision makers discriminate based on a general dislike of a certain group (taste-based discrimination) or based on assumptions about the average characteristics of an individual from that group (statistical discrimination).[10] Recent audits have attempted to adjudicate between taste-based and statistical discrimination by varying multiple characteristics and examining differences in response rates between types of characteristics (more or less susceptible to taste-based discrimination) and examining interactions with characteristics that might provide information to overcome statistical discrimination

[10] David Neumark (forthcoming) provides an excellent review of these and other types of discrimination, so I do not go into more detailed explanation here.

(Agerström et al. 2012; Ahmed et al. 2010; Auspurg et al. 2017; Baldini and Federici 2011; Bosch et al. 2010; Capéau et al. 2012; Carlsson and Ericksson 2014; Drydakis 2014; Edo et al. 2013; Ewens et al. 2014; Gneezy et al. 2012; Hanson and Hawley 2014; Hanson and Santas 2014). The results from these studies are somewhat mixed as to whether taste-based or statistical discrimination occurs more often (or some combination of the two). These mixed findings likely stem from the variety of locations and characteristics studied.

Two studies related to taste-based versus statistical discrimination stand out among the rest (Bartoš et al. 2016; Pager 2016). In the first, the authors examined how both an individual characteristic, in this case race, and the type of market can lead to "attention discrimination," or the differential use of available information. The authors set up audits in rental housing and labor markets and found that in the first market, decision makers selected more applicants overall and more often examined additional information from minority applicants. In the later market, decision makers selected fewer applicants overall and more often examined additional information from majority applicants. Thus, discrimination in acquiring information about candidates occurred at the initial stage of selection and varied by the selectivity of the market. We should be cautious to consider how these types of processes – overall response or selection rates in a given market and the differential use of available information – might influence future audits.

In the second, Devah Pager (2016) examined whether firms that discriminated in a previous audit are still in business 6 years later. Economists suggest that an efficient market should eventually weed out taste-based discrimination since not all employers exhibit that type of discrimination and those who do will pay a penalty for inefficient hiring (Arrow 1973; Becker 1957). Using additional data on firm failure, Pager found that prior discrimination is associated with a firm going out of business. Although other factors may explain this relationship, the findings are at least consistent with taste-based discrimination.

Context and Conditions Under Which Discrimination Occurs

Another major development during this period has been researchers' increased focus on context and the conditions under which discrimination occurs. Two aspects of context – geographic location and occupation or market characteristics – have played a significant role in recent audits. Those audits that have taken geographic variation into account often examine differences by neighborhood characteristics such as racial, ethnic, immigrant, and SES composition (Acolin et al. 2016; Carlsson and Ericksson 2014, 2015; Carlsson et al. 2017; Galster et al. Forthcoming; Ghoshal and Gaddis 2015; Hanson and Hawley 2011; MacDonald et al. forthcoming). Others have examined geography in more detail by tying discrimination- or prejudice-based theories into the analysis (Besbris et al. 2015, Chap. 8 of this volume; Gaddis and Ghoshal 2015; Hanson and Hawley 2014; Phillips 2016a). A second strand of research has considered if levels of discrimination are influenced by the types or composition of occupations (Albert et al. 2011; Andriessen et al. 2012; Booth and

Leigh 2010; Bursell 2014; Carlsson 2011; Derous et al. 2012; Zhou et al. 2013), whether a job is a promotion (Baert et al. 2016a), whether an applicant is overqualified (Baert and Verhaest 2014; Verhaest et al. forthcoming), or market tightness or slackness (Baert et al. 2015; Carlsson et al. 2015; Farber et al. 2017; Vuolo et al. 2017).

Some researchers have varied multiple individual characteristics simultaneously and examined interactions to try to capture a broader spectrum of the decision-making process. In particular, recent audits have focused on interactions between race/ethnicity and educational credentials (Carbonaro and Schwarz, Chap. 7 of this volume; Darolia et al. 2015; Deming et al. 2016; Gaddis 2015, 2017e; Lee and Khalid 2016; Nunley et al. 2015), race/ethnicity and criminal record (Ahmed and Lang 2017; Decker et al. 2015; Uggen et al. 2014), race/ethnicity and sexual orientation (Mazziotta et al. 2015) and various combinations of personal characteristics and human capital characteristics (Andersson et al. 2012; Baert and Vujic 2016; Baert et al. 2016b, 2017; Johnson and Lahey 2011; Namingit et al. 2017; Neumark et al. 2015; Nunley et al. 2016, 2017; Oreopoulos and Dechief 2012; Pedulla 2016; Phillips 2017).

Some of the most interesting research to examine context and conditions has focused on the effects of policies. In one such study, a team of researchers examined whether discrimination against individuals with a disability varied by whether a company was subject to the Americans with Disabilities Act (ADA) (Ameri et al. forthcoming). The authors found that the ADA reduced discrimination against disabled applicants among employers that were covered under the law. A second study used audit and non-audit data to examine differences in age discrimination across states by differences in anti-discrimination policies (Neumark et al. 2017). The authors found no strong relationship between the strength of state laws and discrimination rates. Finally, a third study used a difference-in-differences design with an audit, multiple time points, and a policy change (Agan and Starr 2016). The authors tested the effect of ban-the-box policies, which prevent an employer from collecting information on criminal record, on levels of racial discrimination in hiring. They found that after ban-the-box policies went into effect, levels of racial discrimination increased. The authors suggest that when employers cannot ask about criminal history, they may engage in statistical discrimination and assume that black applicants have a criminal record.

Methodological Issues in Audit Design

In recent years, scholars have considered at least three methodological issues in audit design: (1) paired vs nonpaired audits, (2) indirect signals of race, and (3) the Heckman critique of unobserved differences between groups. First, in my experience, the question of paired versus non-paired audit design is often a concern during IRB submission and subsequent discussions. A paired audit design opens the research up to an increased chance of experiment discovery because decision

makers can potentially see two applicants or inquiries that are very similar. However, conventional wisdom suggests that the paired design is more statistically efficient, decreases the amount of time required for data collection, and can lead to a larger sample size (Lahey and Beasley, Chap. 4 of this volume). In at least two cases, fear of experiment discovery preemptively led to a non-paired audit design (Weichselbaumer 2015, 2016). Additionally, researchers have raised concerns that paired designs may influence findings of discrimination because researchers insert fake applicants into the applicant pool without knowing the composition of that applicant pool (Phillips 2016b; Weichselbaumer 2015). Employers compare applicants to each other and by inserting more than one applicant into a particular pool, researchers may influence the process. In fact, Phillips (2016b) developed a method to test these effects and found that "adjusting for applicant pool composition increases measured discrimination by 20% on average" (2016b: 1). Moreover, proper power analysis suggests that paired audits are not needed as often as researchers think (Vuolo et al. 2016, Chap. 6 of this volume).

I have devoted considerable time and effort to a second methodological concern – the indirect signaling of race through names (Gaddis 2017a, b, c d). With correspondence audits, researchers lose the ability to directly convey race through appearance and must rely on an indirect signal, such as a name, to signal race. Although prior research occasionally raised some concerns about the signal of names (e.g. Bertrand and Mullainathan 2004), only 17.5% of the studies I reviewed used pretests to examine the perception of names used in an audit (Gaddis 2017a). My work has shown that racial perceptions of white and black names are often linked with social class (Gaddis 2017a), Hispanic names are strongly identified (Gaddis 2017b), immigrant generational status can be discerned through names (Gaddis 2017c), and, perhaps most importantly, audit findings are strongly linked to the names researchers use (Gaddis 2017d). Still, more needs to be done to examine the signals we use in audit studies (see next part).

The final area of methodological inquiry concerns the Heckman critique of unobserved differences between groups and has received the most scholarly attention of the three issues discussed here (Heckman 1998; Heckman and Siegelman 1993). James Heckman's critique is that scholars using the audit design assume that unobservable characteristics have equal means across groups, yet scholars cannot confirm that. Heckman suggests that multiple components could enter into the decision-making process – some controlled for by audit design and others unknown to designers but known to the decision makers. In other words, characteristics that researchers do not include on a resume or in an email. These components combine to place a candidate above or below the threshold to receive a response. If the two groups being studied have different variances on these important unobserved components, audit studies may over or underestimate discrimination or detect an effect when there is not one. David Neumark (forthcoming) provides a more detailed discussion of this critique and has devised a method to produce an unbiased estimate of discrimination and avoid this critique (Neumark 2012). Neumark (2012) reanalyzed Bertrand and Mullainathan's (2004) original audit data using this method to account

for the variance of unobservables and found stronger evidence of racial discrimination. Two individual studies have implemented Neumark's method, with no clear pattern regarding bias (Baert 2015; Neumark et al. 2016). Two other studies have re-analyzed data from multiple audits and suggest that employment audits appear to be susceptible to the Heckman critique (Carlsson et al. 2014; Neumark and Rich 2016). The authors of these two studies advise that scholars still have a lot of work to do in improving the audit method by more directly addressing this critique.

Including Additional Data from Outside the Audit

A final major development in recent audit research is the inclusion of additional data from outside the audit itself, something done by many of the studies already mentioned in this part. Several researchers have included geographic data on neighborhood and city characteristics to supplement audits (e.g. Acolin et al. 2016; Carlsson and Ericksson 2014, 2015; Ghoshal and Gaddis 2015; Hanson and Hawley 2011). Others have included other types of available data, such as firm closure (Pager 2016), mortgage lender transactions (Hanson et al. 2017), and existing survey data on racial/ethnic attitudes and beliefs (Carlsson and Ericksson 2017; Carlsson and Rooth 2012).

One of the most promising avenues of inquiry into discrimination is the combination of audits with other methods of data collection. Following in the footsteps of Pager and Quillian (2005), researchers are increasingly obtaining a second round of information from the same individuals who previously participated in an audit. Some researchers have followed-up with employers to administer implicit association tests (IATs) to examine the connection between implicit bias and discrimination (Agerström and Rooth 2011; Rooth 2010). Other researchers have followed-up with surveys or interviews after an audit to attempt to better understand the reasons behind discriminatory actions (Bonnet et al. 2016; Midtbøen 2014, 2015, 2016; Zussman 2013). Although institutional review boards (IRBs) may be hesitant to allow researchers to engage in multiple points of contact with audit participants, some researchers have successfully shown that additional methods of data collection do not necessarily need to follow up with the original audit participants (Gaddis and Ghoshal 2017; Kang et al. 2016).

I believe that researchers should continue in the direction of the trends discussed above – adjudicating among types of discrimination, focusing on context and the conditions under which discrimination occurs, focusing on methodological issues in audit design, and including additional data from outside audits. In particular, researchers should try to include geographic data in audits, given the wide availability of geographic data and the relative simplicity and usefulness of including such data in analyzing audit outcomes. Next, in the final part, I outline some limitations of correspondence audits and return to the issues discussed in this part with additional thoughts on continuing to improve correspondence audits.

1.5 Limitations of and Ways to Improve Correspondence Audits

Despite the rapid advancement of correspondence audits over the past two decades, several serious limitations exist that scholars must continue to address. Limitations of in-person audits have been covered by others in detail, particularly James Heckman (1998; Heckman and Siegelman 1993), and I draw upon that work here. However, correspondence audits often have their own unique quirks and limitations. By no means is this part intended to be an exhaustive list of all the limitations of correspondence audits, but instead some areas where I see the biggest problems and/or new potential solutions. I highly recommend the reader turn to David Pedulla's chapter (Chap. 9 of this volume) for a more extensive and detailed discussion of these and other issues.

Perhaps most important is the general limitation of audit studies in uncovering mechanisms rather than simply documenting the existence of discrimination. As discussed in the previous part, recent work has started to expand our knowledge in this area in increasingly innovative ways. Not all questions will lend themselves to design tricks built into studies to help discover mechanisms, nor can researchers always implement complex factorial designs to test potential mechanisms. My recommendation is that researchers should be more open to collecting survey experiment data side-by-side with field data from audit studies (e.g. Diehl et al. 2013; Gaddis and Ghoshal 2017). The deception of the audit study may allow us to document discrimination but a similar scenario presented as a survey experiment may allow us to explore potential mechanisms with the right questions. Moreover, the rise of Amazon's Mechanical Turk (MTuk) makes collecting survey experiment data relatively quick and cheap (Campbell and Gaddis 2017; Porter et al. 2017). In ongoing work combining an audit with a survey experiment, I find that roommate discrimination against many different racial and ethnic groups is driven by issues of cultural fit. However, blacks face higher levels of discrimination than others due to negative perceptions about financial stability and courteousness, despite respondents receiving the same information about all racial/ethnic groups (Gaddis and Ghoshal 2017). These findings would not have come to light if we had implemented a correspondence audit or survey experiment alone.

A second major limitation of correspondence audits is indirect signaling of characteristics. Correspondence audits often require signals to be sent through names, statements, lists, or other text embedded in communications. In my own research, I have worked to understand how names can be used to signal race, ethnicity, and immigrant status (Gaddis 2017a, b, c) and have found that signals of race are conflated with social class and that conflation explains differences in response rates across previous correspondence audits (Gaddis 2017d). Still, more work needs to be done to ensure that construct validity is high when we need to indirectly signal characteristics in correspondence audits. At a minimum, researchers should pretest their signals in a scientific manner to help increase construct validity. Additionally,

more work is needed to explore the possibility of alternate signals since there is often more than one way to indirectly signal a characteristic.

The signaling of characteristics is also related to the way we can conduct correspondence audits and the level of external validity of those audits. A characteristic such as race or gender may convey different things depending on how it is signaled and the context in which it is signaled. Not only are correspondence audits only as good as the signals they use to convey key characteristics, but audit studies also only tell us about a specific avenue of correspondence with a specific signal. For example, real job seekers may use any combination of online job sites, personal and professional networks, alumni resources, headhunters, and employment events. How race is conveyed and the meaning of race likely vary across these different means of searching for a job. Static, written signals – such as name, professional affiliations, or even checking a box for race – may cue stereotypes about race. Dynamic, interpersonal signals – such as a discussion with a reference or interaction with the individual – may permit more flexibility in thoughts about race. Although others have raised concerns about how audits begin with a narrow sampling frame (e.g., jobs or housing posted in newspapers or on websites) and limit generalizability to the entire job or housing search process (Friedman 2015; Gaddis 2015; Heckman and Siegelman 1993; Pitingolo and Ross 2015), I suggest that the narrow sampling frame also limits our knowledge of discrimination processes only to those that can be conveyed through certain static and often indirect signals.

Although in-person audits have occasionally examined multiple outcomes at various stages of the processes they study (Bendick et al. 1994; Pager et al. 2009a; Turner et al. 1991a), correspondence audits have been almost entirely limited to studying outcomes at the initial contact phase. Critics have pointed out that we do not know whether the disparities witnessed at the initial contact phase lead to disparities at later phases (Heckman 1998; Heckman and Siegelman 1993). Others have used nationally representative data to simulate the effect of employer callback disparities on wages (Lanning 2013). Still, as my own research shows, we should use all the information possible to expand the outcomes examined by audit studies. Additional information in both employment (Gaddis 2015) and housing advertisements (Gaddis and Ghoshal 2015, 2017; Ghoshal and Gaddis 2015) should be used to our advantage.

Furthermore, we should consider additional ways that audits might be tweaked to examine other outcomes. In employment audits, do human resources staff visit LinkedIn or Facebook pages, contact references, or attempt multiple contacts with applicants at different rates? Some recent articles provide excellent examples of the directions audits might continue to go in the future (Acquisti and Fong 2015; Baert forthcoming; Bartoš et al. 2016; Blommaert et al. 2014; Butler and Crabtree forthcoming; and see Crabtree, Chap. 5 of this volume for more discussion). Additionally, is it possible to return to the strategies of earlier audits and use a sub-sample with real humans to proceed deeper into processes, such as sending trained assistants into in-person or Skype interviews? I believe that future waves of audit studies will need to be creative and incorporate more variety in outcomes to push this method forward.

1.6 This Volume and Online Resources

This volume is organized into three broad parts: (1) The Theory Behind and History of Audit Studies, (2) The Method of Audit Studies: Design, Implementation, and Analysis, and (3) Nuance in Audit Studies: Context, Mechanisms, and the Future. You are reading the first chapter of the first part and, hopefully, you already have a better understanding of audit studies. In the second chapter, Fran Cherry and Marc Bendick discuss the historical connections between activism and scholarship through audits. Their chapter highlights the potential power of audit studies to not just document discrimination but reduce it as well. The authors advocate for a return to scholar-activism and outline four characteristics that will help facilitate that path. In the third chapter, Stijn Baert provides an excellent overview of labor market correspondence audits conducted since Bertrand and Mullainathan's groundbreaking study. Baert organizes these studies across two major dimensions: discrimination treatment characteristic, which includes nine federally-banned (U.S.) and five state-banned discrimination grounds, and country of analysis. Overall, the author provides information on 90 labor market correspondence audits across 24 countries.

The chapters in the second part give the reader a "behind-the-scenes" look at the nuts and bolts of audit studies, as well as serve as a guide for designing and implementing your own audit studies. In the fourth chapter, Joanna Lahey and Ryan Beasley outline a number of technical aspects related to designing and conducting a correspondence audit. They cover issues of validity, participant selection, timing, technical design of correspondence, matching, sample size, and analysis, among other issues. Their chapter serves as a terrific starting point for anyone needing more information on creating their own audit. In the fifth chapter, Charles Crabtree extends this discussion by providing a detailed overview of designing and implementing an email correspondence audit. He provides information on sample selection, collecting email addresses, sending emails, and collecting outcomes. This chapter is particularly useful in thinking about automating an audit design using programming scripts. A coding appendix for this chapter will be available at auditstudies.com. In the sixth chapter, Mike Vuolo, Christopher Uggen, and Sarah Lageson offer an extensive consideration of matched versus non-matched audit designs. They provide statistical guidelines for when matching is appropriate and show that non-matched audit designs can be more efficient. Additionally, they raise some important substantive points for researchers to think about when deciding to use a matched or non-matched design.

Finally, the chapters in the third part provide even deeper insight into the audit process by discussing more design considerations and nuance. In the seventh chapter, William Carbonaro and Jonathan Schwarz outline their thought process in selecting cities in which to conduct an audit, the difficulties of using a small city, the unknowns of the employer side of an audit, and the choice of jobs for a sample. This chapter shares important "lessons learned" from experienced researchers. Although scholars cannot think through all of the possible variables involved in designing and fielding an audit in advance, I think this chapter serves as a great example of how auditing is an incredibly difficult and nuanced process. In the eighth chapter, Max

Besbris, Jacob William Faber, Peter Rich, and Patrick Sharkey show how an audit can be designed to investigate a non-individual-level treatment. They use an audit to examine the mechanism of place-based stigma in the relationship between neighborhoods and outcomes for residents of those neighborhoods. Their audit, the discussion of thinking about signaling characteristics, and the theory-based use of geography provide a strong example of what future audits might looks like. In the ninth and final chapter, David Pedulla explores how audits might change and develop in the coming years. He highlights research that identifies mechanisms, examines when and where discrimination happens, and scrutinizes issues of representativeness. David's chapter serves as a terrific bookend to this volume and should be read closely by anyone wishing to implement an audit of their own.

On behalf of the other contributors, we hope you find this volume informative and useful. We have a number of overarching goals for this book: (1) to create a go-to guide for anyone looking to conduct an audit study, (2) to provide resources for using the audit method, both within this book and online, and (3) to record the history of audits. For more information on audits, please consult our website at www.auditstudies.com and take a look at the recommend reading list below.

1.7 Recommended Reading

1.7.1 Comprehensive Articles and Books on Audits

"Situation Testing for Employment Discrimination in the United States." 2007. By Marc Bendick Jr. *Horizons Stratégiques*, 3:17–39.

Clear and Convincing Evidence: Measurement of Discrimination in America. 1993. Edited by Michael Fix and and Raymond J. Struyk. Washington, DC: The Urban Institute.

"Experimental Research on Labor Market Discrimination." Forthcoming. By David Neumark. *Journal of Economic Literature.*

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

1.7.2 Reviews of Audits and Discrimination Research

"What Have We Learned from Paired Testing in Housing Markets?" 2015. By Sun Jung Oh and John Yinger. *Cityscape: A Journal of Policy Development and Research*, 17(3):15–59.

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

"Field Experiments of Discrimination in the Market Place." 2002. By Peter A. Riach and Judith Rich. *The Economic Journal*, 112:F480-F518.

"What Do Field Experiments of Discrimination in Markets Tells Us? A Meta-Analysis of Studies Conducted Since 2000." 2014. By Judith Rich. Available at SSRN: https://ssrn.com/abstract=2517887

"A Multidisciplinary Survey on Discrimination Analysis." 2013. By Andrea Romei and Salvatore Ruggieri. *The Knowledge Engineering Review*, 29(5):582–638.

1.7.3 Meta-Analyses of Audits

"Meta-Analysis of Field Experiments Shows no Change in Racial Discrimination in Hiring over Time." 2017. By Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn Midtbøen. *Proceedings of the National Academy of Sciences.*

"Ethnic Discrimination in Hiring Decisions: A Meta-Analysis of Correspondence Tests 1990–2015." 2016. By Eva Zschirnt and Didier Ruedin. *Journal of Ethnic and Migration Studies*, 42(7):1115–34.

1.7.4 Articles and Books on the Methodology of Audits, Discrimination, and Field Experiments

Field Experiments (General)

"Field Experiments Across the Social Sciences." 2017. By Delia Baldassarri and Maria Abascal. *Annual Review of Sociology*, 43:41–73.

Field Experiments: Design, Analysis,and Interpretation. 2012. By Alan S. Gerber and Donald P. Green. New York, NY: W.W. Norton.

"The Principles of Experimental Design and Their Application in Sociology." 2013. By Michelle Jackson and D. R. Cox. *Annual Review of Sociology*, 39:27–49.

Audits (General)

Audit Studies: Behind the Scenes with Theory, Method, and Nuance. 2018. Edited by S. Michael Gaddis. Switzerland: Springer International Publishing.

Discrimination (General)

Measuring Racial Discrimination. 2004. By Rebecca N. Blank, Marilyn Dabady, and Constance F. Citro. Washington, DC: The National Academies Press.

Automating Resume Creation for Audits

"Computerizing Audit Studies." 2009. By Joanna N. Lahey and Ryan A. Beasley. *Journal of Economic Behavior & Organization*, 70(3):508–14.

Critiques of Audits and Solutions

"Detecting Discrimination." 1998. By James J. Heckman. *Journal of Economic Perspectives*, 12(2):101–16.

"The Urban Institute Audit Studies: Their Methods and Findings." 1993. By James J. Heckman and Peter Siegelman. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by M. Fix and R. J. Struyk, 187–258. Washington, DC: The Urban Institute Press.

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

"Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-Examination of the Evidence." 2016. By David Neumark and Judith Rich. Available at NBER: http://www.nber.org/papers/w22278

Signaling Characteristics in Audits

"How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." 2017. By S. Michael Gaddis. *Sociological Science,* 4:469–489.

"Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." 2017. By S. Michael Gaddis. *Socius*, 3:1–11.

"Assessing Immigrant Generational Status from Names: Scientific Evidence for Experiments." 2017. By S. Michael Gaddis. Available at SSRN: https://ssrn.com/abstract=3022217

"Auditing Audit Studies: The Effects of Name Perception and Selection on Social Science Measurement of Racial Discrimination." 2017. By S. Michael Gaddis. Available at SSRN: https://ssrn.com/abstract=3022207

Statistical Analysis of Audits

"Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests with Nominal Outcomes." 2016. By Mike Vuolo, Christopher Uggen, and Sarah Lageson. *Sociological Methods & Research*, 45(2):260–303.

1.7.5 Theoretical Articles and Books on Discrimination

"Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots." 2013. By Jonathan Guryan and Kerwin Kofi Charles. *The Economic Journal*, 123:F417–32.

Theorizing Discrimination in an Era of Contested Prejudice: Discrimination in the United States, Volume 1. 2008. By Samuel Roundfield Lucas. Philadelphia, PA: Temple University Press.

Acknowledgments Many scholars have played important roles in sharpening my thoughts on this method. I cannot name them all here but I want to express my thanks to Devah Pager, Bill Carbonaro, Joanna Lahey, and David Pedulla for support in helping this volume come together. Additionally, I would like to thank fellow panelists and audience members at sessions on audits at the 2014 annual meeting of the Association for Public Policy Analysis and Management in Albuquerque, NM and the 2015 annual meeting of the American Sociological Association in Chicago, IL. Finally, thanks to a host of other people who helped make this volume happen: the anonymous reviewers of the chapters, the editors and staff at Springer, and my lovely wife who has always been incredibly supportive of my sometimes chaotic academic endeavors.

References

Acolin, A., Bostic, R., & Painter, G. (2016). A field study of rental market discrimination across origins in France. *Journal of Urban Economics, 95*, 49–63.

Acquisti, A., & Fong, C. M. (2015). *An experiment in hiring discrimination via online social networks*. Available at SSRN: https://ssrn.com/abstract=2031979

Adam, B. D. (1981). Stigma and employability: Discrimination by sex and sexual orientation in the Ontario legal profession. *Canadian Review of Sociology, 18*(2), 216–221.

Adida, C. L., Laitin, D. D., & Valfort, M.-A. (2010). Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences, 107*(52), 22384–22390.

Agan, A. Y., & Starr, S. B. (2016). *Ban the box, criminal records, and statistical discrimination: A field experiment*. Available at SSRN: https://ssrn.com/abstract=2795795

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96*(4), 790–805.

Agerström, J., Björklund, F., Carlsson, R., & Rooth, D.-O. (2012). Warm and competent Hassan = cold and incompetent Eric: A harsh equation of real-life hiring discrimination. *Basic and Applied Social Psychology, 34*(4), 359–366.

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the internet. *Journal of Urban Economics, 64*(2), 362–372.

Ahmed, A. M., & Hammarstedt, M. (2009). Detecting discrimination against homosexuals: Evidence from a field experiment on the internet. *Economica, 76*(303), 588–597.

Ahmed, A. M., & Lang, E. (2017). The employability of ex-offenders: A field experiment in the Swedish labor market. *IZA Journal of Labor Policy, 6*(6), 1–23.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2008). Are lesbians discriminated against in the rental housing market? Evidence from a correspondence testing experiment. *Journal of Housing Economics, 17*(3), 234–238.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2010). Can discrimination in the housing market be reduced by increasing the information about the applicants? *Land Economics, 86*(1), 79–90.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2012). Does age matter for employability? A field experiment on ageism in the Swedish labour market. *Applied Economics Letters, 19*(4), 403–406.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2013). Are gay men and lesbians discriminated against in the hiring process? *Southern Economic Journal, 79*(3), 565–585.

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review, 30*(2), 175–187.

Albert, R., Escot, L., & Fernández-Cornejo, J. A. (2011). A field experiment to study sex and age discrimination in the Madrid labour market. *The International Journal of Human Resource Management, 22*(2), 351–375.

Allasino, E., Reyneri, E., Venturini, A., & Zincone, G. (2004). *Labour market discrimination against migrant workers in Italy*. Geneva: International Labour Office.

Allred, B. B., Findley, M. G., Nielson, D., & Sharman, J. C. (2017). Anonymous shell companies: A global audit study and field experiment in 176 countries. *Journal of International Business Studies, 48*(5), 596–619.

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3C, pp. 3143–3259). New York: Elsevier.

Ameri, M., Schur, L., Adya, M., Scott Bentley, F., McKay, P., & Kruse, D. (Forthcoming). The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review*. https://doi.org/10.1177/0019793917717474.

Anagol, S., Cole, S., & Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the Indian life insurance market. *The Review of Economics and Statistics, 99*(1), 1–15.

Andersson, L., Jakobsson, N., & Kotsadam, A. (2012). A field experiment of discrimination in the Norwegian housing market: Gender, class, and ethnicity. *Land Economics, 88*(2), 233–240.

Andriessen, I., Nievers, E., Dagevos, J., & Faulk, L. (2012). Ethnic discrimination in the Dutch labor markets: Its relationship with job characteristics and multiple group membership. *Work and Occupations, 39*(3), 237–269.

Arai, M., Bursell, M., & Nekby, L. (2016). The reverse gender gap in ethnic discrimination: Employer stereotypes of men and women with Arabic names. *International Migration Review, 50*(2), 385–412.

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2014). Race and marriage in the labor market: A discrimination correspondence study in a developing country. *The American Economic Review, 104*(5), 376–380.

Ariel, B., Tobby-Alimi, I., Cohen, I., Ezra, M. B., Cohen, Y., & Sosinski, G. (2015). Ethnic and racial employment discrimination in low-wage and high-wage markets: Randomized controlled trials using correspondence tests in Israel. *The Law & Ethics of Human Rights, 9*(1), 113–139.

Arrijn, P., Feld, S., & Nayer, A. (1998). *Discrimination in access to employment on grounds of foreign origin: The case of Belgium*. Geneva: International Labour Office, Conditions of Work Branch.

Arrow, K. (1972). Some mathematical models of race discrimination in the labor market. In A. H. Pascal (Ed.), *Racial discrimination in economic life* (pp. 187–204). Lexington: D.C. Heath.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor market* (pp. 3–33). Princeton: Princeton University Press.

Attström, K. (2007). *Discrimination against native swedes of immigrant origin in access to employment*. Geneva: International Labour Office.
Auspurg, K., Hinz, T., & Schmid, L. (2017). Contexts and conditions of ethnic discrimination: Evidence from a field experiment in a German housing market. *Journal of Housing Economics, 35*, 26–36.
Ayers, I. (1991). Fair driving: Gender and race discrimination in retail car negotiations. *Harvard Law Review, 104*(4), 817–872.
Ayers, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *The American Economic Review, 85*(3), 304–321.
Baert, S. (2014a). Wage subsidies and hiring chances for the disabled: Some causal evidence. *The European Journal of Health Economics, 17*(1), 71–86.
Baert, S. (2014b). Career lesbians. Getting hired for not having kids? *Industrial Relations Journal, 45*(6), 543–561.
Baert, S. (2015). Field experimental evidence on gender discrimination in hiring: Biased as Heckman and Siegelman predicted? *Economics: The Open-Access, Open-Assessment E-Journal, 9*(25), 1–11.
Baert, S. (2016). Wage subsidies and hiring chances for the disabled: Some causal evidence. *The European Journal of Health Economics, 17*(1), 71–86.
Baert, S. (2018). Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Baert, S. (Forthcoming). Facebook profile picture appearance affects recruiters' first hiring decisions. *New Media & Society*. https://doi.org/10.1177/1461444816687294.
Baert, S., & Balcaen, P. (2013). The impact of military work experience on later hiring chances in the civilian labour market. Evidence from a field experiment. *Economics: The Open-Access, Open-Assessment E-Journal, 7*(37), 1–17.
Baert, S., & Omey, E. (2015). Hiring discrimination against pro-union applicants: The role of union density and firm size. *De Economist, 163*(3), 263–280.
Baert, S., & Verhaest, D. (2014). *Unemployment or overeducation: Which is a worse signal to employers?* Available at SSRN: https://ssrn.com/abstract=2468488
Baert, S., & Verhofstadt, E. (2015). Labour market discrimination against former juvenile delinquents: Evidence from a field experiment. *Applied Economics, 47*(11), 1061–1072.
Baert, S., & Vujic, S. (2016). Immigrant volunteering: A way out of labour market discrimination? *Economics Letters, 146*, 95–98.
Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2015). Is there less discrimination in occupations where recruitment is difficult? *ILR Review, 68*(3), 467–500.
Baert, S., de Pauw, A.-S., & Deschacht, N. (2016a). Do employer preferences contribute to sticky floors? *ILR Review, 69*(3), 714–736.
Baert, S., de Visschere, S., Schoors, K., Vandenberge, D., & Omey, E. (2016b). First depressed, then discriminated against? *Social Science & Medicine, 170*, 247–254.
Baert, S., Norga, J., Thuy, Y., & van Hecke, M. (2016c). Getting grey hairs in the labour market. An alternative experiment on age discrimination. *Journal of Economic Psychology, 57*, 86–101.
Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination? *Economics Letters, 155*, 35–38.
Bailey, J., Wallace, M., & Wright, B. (2013). Are gay men and lesbians discriminated against when applying for jobs? A four-city, internet-based field experiment. *Journal of Homosexuality, 60*(6), 873–894.
Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology, 43*, 41–73.
Baldini, M., & Federici, M. (2011). Ethnic discrimination in the Italian rental housing market. *Journal of Housing Economics, 20*(1), 1–14.
Banerjee, A., Bertrand, M., Datta, S., & Mullainathan, S. (2009). Labor market discrimination in Delhi: Evidence from a field experiment. *Journal of Comparative Economics, 37*(1), 14–27.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *The American Economic Review, 106*(6), 1437–1475.

Bavan, M. (2007). Does housing discrimination exist based on the 'color' of an individual's voice? *City, 9*(1), 93–107.

Becker, G. S. (1957). *The economics of discrimination*. Chicago: The University of Chicago Press.

Bendick, M., Jr. (1989). *Auditing race discrimination in employment: A research design*. Washington, DC: The Urban Institute.

Bendick, M., Jr. (2007). Situation testing for employment discrimination in the United States. *Horizons Stratégiques, 3*, 17–39.

Bendick, M., Jr., Jackson, C. W., Reinoso, V. A., & Hodges, L. E. (1991). Discrimination against Latino job applicants: A controlled experiment. *Human Resource Management, 30*(4), 469–484.

Bendick, M., Jr., Jackson, C. W., & Reinoso, V. A. (1994). Measuring employment discrimination through controlled experiments. *The Review of Black Political Economy, 23*(1), 25–48.

Bendick, M., Jr., Jackson, C. W., & Horacio Romero, J. (1997). Employment discrimination against older workers: An experimental study of hiring practices. *Journal of Aging & Social Policy, 8*(4), 25–46.

Bendick, M., Jr., Brown, L. E., & Wall, K. (1999). No foot in the door: An experimental study of employment discrimination against older workers. *Journal of Aging & Social Policy, 10*(4), 5–23.

Bendick, M., Jr., Rodriguez, R. E., & Jayaraman, S. (2010). Employment discrimination in upscale restaurants: Evidence from matched pair testing. *The Social Science Journal, 47*(4), 802–818.

Bengtsson, R., Iverman, E., & Hinnerich, B. T. (2012). Gender and ethnic discrimination in the rental housing market. *Applied Economics Letters, 19*(1), 1–5.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013.

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2015). Effect of neighborhood stigma on economic transactions. *Proceedings of the National Academy of Sciences, 112*(16), 4994–4998.

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2018). The geography of stigma: Experimental methods to identify the penalty of place. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Birkelund, G. E., Heggebø, K., & Rogstad, J. (2017). Additive or multiplicative disadvantage? The scarring effects of unemployment for ethnic minorities. *European Sociological Review, 33*(1), 17–29.

Blank, R. M., Dabady, M., & Citro, C. F. (2004). *Measuring racial discrimination*. Washington, DC: The National Academies Press.

Blommaert, L., Coenders, M., & van Tubergen, F. (2014). Discrimination of Arabic-named applicants in the Netherlands: An internet-based field experiment examining different phases in online recruitment procedures. *Social Forces, 92*(3), 957–982.

Boggs, R. V. O., Sellers, J. M., & Bendick, M., Jr. (1993). Use of testing in civil rights enforcement. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 345–375). Washington, DC: The Urban Institute Press.

Bonnet, F., Lalé, E., Safi, M., & Wasmer, E. (2016). Better residential than ethnic discrimination! Reconciling audit and interview findings in the Parisian housing market. *Urban Studies, 53*(13), 2815–2833.

Bóo, F. L., Rossi, M., & Urzúa, S. S. (2013). The labor market return to an attractive face: Evidence from a field experiment. *Economics Letters, 118*(1), 170–172.

Booth, A. L., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters, 107*(2), 236–238.

Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics, 74*(4), 547–573.

Bosch, M., Angeles Carnero, M., & Farré, L. (2010). Information and discrimination in the rental housing market: Evidence from a field experiment. *Regional Science and Urban Economics, 40*(1), 11–19.

Bovenkerk, F. (1992). *Testing discrimination in natural experiments: A manual for international comparative research on discrimination on the grounds of 'race' and ethnic origin*. Geneva: International Labour Office.

Bovenkerk, F., Kilborne, B., Raveau, F., & Smith, D. (1979). Comparative aspects of research on discrimination against non-white citizens in great Britain, France, and the Netherlands. In J. Berting, F. Geyer, & R. Jurkovich (Eds.), *Problems in international comparative research in the social science* (pp. 105–122). Oxford: Pergamon Press.

Bovenkerk, F., Gras, M. J. I., Ramsoedh, D., Dankoor, M., & Havelaar, A. (1995). *Discrimination against migrant workers and ethnic minorities in access to employment in the Netherlands*. Geneva: International Labour Office, Employment Department.

Broockman, D. E. (2013). Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science, 57*(3), 521–536.

Brown, C., & Gay, P. (1985). *Racial discrimination: 17 years after the act*. London: Policy Studies Institute.

Bursell, M. (2007). *What's in a name? A field experiment test for the existence of ethnic discrimination in the hiring process*. Stockholm: The Stockholm University Linnaeus Center for Integration Studies, University of Stockholm.

Bursell, M. (2014). The multiple burdens of foreign-named men – Evidence from a field experiment on gendered ethnic hiring discrimination in Sweden. *European Sociological Review, 30*(3), 399–409.

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science, 55*(3), 463–477.

Butler, D. M., & Crabtree, C. (Forthcoming). Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. *Journal of Experimental Political Science, 4*(1), 57–67.

Butler, D. M., & Homola, J. (2017). An empirical justification for the use of racially distinctive names to signal race in experiments. *Political Analysis, 25*(1), 122–130.

Bygren, M., Erlandsson, A., & Gähler, M. (2017). Do employers prefer fathers? Evidence from a field experiment testing the gender by parenthood interaction effect on callbacks to job applications. *European Sociological Review, 33*(3), 337–348.

Campbell, C., & Gaddis, S. M. (2017). I don't agree with giving cash': A survey experiment examining support for public assistance. *Social Science Quarterly, 98*(5), 1352–1373.

Campos-Vazquez, R. M., & Arceo-Gomez, E. O. (2015). *How does explicit discrimination in job ads interact with discrimination in callbacks? Evidence from a correspondence study in Mexico City*. Working paper available at: http://repositorio-digital.cide.edu/handle/11651/845

Capéau, B., Eeman, L., Groenez, S., & Lamberts, M. (2012). *Two concepts of discrimination: inequality of opportunity versus unequal treatment of equals*. Working Paper. Available at: https://www.researchgate.net/profile/Steven_Groenez/publication/254407466_Two_concepts_of_discrimination_inequality_of_opportunity_versus_unequal_treatment_of_equals/links/5492c1340cf209fc7e9f7e49.pdf

Carbonaro, W., & Schwarz, J. (2018). Opportunities and challenges in designing and conducting a labor market resume study. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Carlsson, M. (2010). Experimental evidence of discrimination in the hiring of first-and second-generation immigrants. *Labour, 24*(3), 263–278.

Carlsson, M. (2011). Does hiring discrimination cause gender segregation in the Swedish labor market? *Feminist Economics, 17*(3), 71–102.

Carlsson, M., & Ericksson, S. (2014). Discrimination in the rental market for apartments. *Journal of Housing Economics, 23*, 41–54.

Carlsson, M., & Ericksson, S. (2015). Ethnic discrimination in the London market for shared housing. *Journal of Ethnic and Migration Studies, 41*(8), 1276–1301.
Carlsson, M., & Ericksson, S. (2017). Do attitudes expressed in surveys predict ethnic discrimination? *Ethnic and Racial Studies, 40*(10), 1739–1757.
Carlsson, M., & Rooth, D.-O. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics, 14*(4), 716–729.
Carlsson, M., & Rooth, D.-O. (2012). Revealing taste-based discrimination in hiring: A correspondence testing experiment with geographic variation. *Applied Economics Letters, 19*(18), 1861–1864.
Carlsson, M., Fumarco, L., & Rooth, D.-O. (2014). Does the design of correspondence studies influence the measurement of discrimination? *IZA Journal of Migration, 3*, 11.
Carlsson, M., Fumarco, L., & Rooth, D.-O. (2015). *Does Labor Market Tightness Affect Ethnic Discrimination in Hiring?* Working paper available at: https://lnu.se/globalassets/lmd-swp20151.pdf
Carlsson, M., Reshid, A. A., & Rooth, D.-O. (2017). *Neighborhood signaling effects, commuting time, and employment*. Working paper available at: https://lnu.se/globalassets/feh/lmd-swp201703.pdf
Carpusor, A. G., & Loges, W. E. (2006). Rental discrimination and ethnicity in names. *Journal of Applied Social Psychology, 36*(4), 934–952.
Cediey, E., & Foroni, F. (2008). *Discrimination in access to employment on grounds of foreign origin in France*. Geneva: International Labour Office.
Chen, J., Pan, J., & Yiqing, X. (2016). Sources of authoritarian responsiveness: A field experiment in China. *American Journal of Political Science, 60*(2), 383–400.
Cherry, F., & Bendick, M., Jr. (2018). Making it count discrimination auditing and the activist scholar tradition. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Cohen, A., & Taylor, E. (2000). *American pharaoh: Mayor Richard J. Daley – His battle for Chicago and the nation*. Boston: Back Bay Books.
Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology, 112*(5), 1297–1338.
Crabtree, C. (2018). An introduction to conducting email audit studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Cross, H., Kenney, G. M., Mell, J., & Zimmermann, W. (1990). *Employer hiring practices: Differential treatment of Hispanic and Anglo job seekers*. Washington, DC: The Urban Institute Press.
Cui, R., Li, J., & Zhang, D. J. (2017). *Discrimination with incomplete information in the sharing economy: Field evidence from Airbnb*. Available at SSRN: https://ssrn.com/abstract=2882982
Daniel, W. W. (1968). *Racial discrimination in England*. Baltimore: Penguin Books.
Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2015). Do employers prefer workers who attend for-profit colleges? Evidence from a field experiment. *Journal of Policy Analysis and Management, 34*(4), 881–903.
de Leon, F. L. L., & Kim, S-H. (2016). *In-Group and Out-Group Biases in the Marketplace: A Field Experiment during the World Cup*. Available at SSRN: https://ssrn.com/abstract=2584414
de Prada, M. A., Actis, W., Pereda, C., & Perez Molina, R. (1996). *Labour market discrimination against migrant workers in Spain*. Geneva: International Labour Office, Employment Department.
Decker, S. H., Ortiz, N., Spohn, C., & Hedberg, E. (2015). Criminal stigma, race, and ethnicity: The consequences of imprisonment for employment. *Journal of Criminal Justice, 43*(2), 108–121.
Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *The American Economic Review, 106*(3), 778–806.

Derous, E., Ryan, A. M., & Nguyen, H.-H. D. (2012). Multiple categorization in resume screening: Examining effects on hiring discrimination against Arab applicants in field and lab settings. *Journal of Organizational Behavior, 33*(4), 544–570.

Deterding, N. M., & Pedulla, D. S. (2016). Educational authority in the 'open door' marketplace: Labor market consequences of for-profit, nonprofit, and fictional educational credentials. *Sociology of Education, 89*(3), 155–170.

Diehl, C., Andorfer, V. A., Khoudja, Y., & Krause, K. (2013). Not in my kitchen? Ethnic discrimination and discrimination intentions in shared housing among university students in Germany. *Journal of Ethnic and Migration Studies, 39*(10), 1679–1697.

Distelhorst, G., & Hou, Y. (2014). Ingroup bias in official behavior: A national field experiment in China. *Quarterly Journal of Political Science, 9*(2), 203–230.

Doleac, J. L., & Stein, L. C. D. (2013). The visible hand: Race and online market outcomes. *The Economic Journal, 123*, F469–F492.

Drydakis, N. (2009). Sexual orientation discrimination in the labour market. *Labour Economics, 16*(4), 364–372.

Drydakis, N. (2010). Labour discrimination as a symptom of HIV: Experimental evaluation – The Greek case. *Journal of Industrial Relations, 52*(2), 201–217.

Drydakis, N. (2011a). Women's sexual orientation and labor market outcomes in Greece. *Feminist Economics, 17*(1), 89–117.

Drydakis, N. (2011b). Ethnic discrimination in the Greek housing market. *Journal of Population Economics, 24*(4), 1235–1255.

Drydakis, N. (2014). Sexual orientation discrimination in the Cypriot labour market. Distastes or uncertainty? *International Journal of Manpower, 35*(5), 720–744.

Drydakis, N., & Vlassis, M. (2010). Ethnic discrimination in the Greek labour market: Occupational access, insurance coverage and wage offers. *The Manchester School, 78*(3), 201–218.

Duguet, E., Leandri, N., L'horty, Y., & Petit, P. (2010). Are young French jobseekers of ethnic immigrant origin discriminated against? A controlled experiment in the Paris area. *Annals of Economics and Statistics/Annales d'Économie et de Statistique, 99/100*, 187–215.

Duguet, E., Du Parquet, L., L'horty, Y., & Petit, P. (2015). New evidence of ethnic and gender discriminations in the French labor market using experimental data: A ranking extension of responses from correspondence tests. *Annals of Economics and Statistics/Annales d'Économie et de Statistique, 117*(118), 21–39.

Dymski, G. A. (2006). A discrimination in the credit and housing markets: Findings and challenges. In W. N. Rodgers III (Ed.), *Handbook on the economics of discrimination* (pp. 215–259). Northampton: Edward Elgar.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*(2), 1–22.

Edo, A., Jacquemet, N., & Yannelis, C. (2013). *Language Skills and Homophilous Hiring Discrimination: Evidence from Gender- and Racially-Differentiated Applications*. Documents de travail du Centre d'Economie de la Sorbonne. Available at: https://hal.archives-ouvertes.fr/halshs-00877458/

Einstein, K. L., & Glick, D. M. (2017). Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *American Journal of Political Science, 61*(1), 100–116.

Elmi, A., & Mickelsons, M. (1991). *Housing discrimination study: Replication of 1977 study measures using current data*. Washington, DC: U.S Department of Housing and Urban Development.

Eriksson, S., & Rooth, D.-O. (2014). Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. *The American Economic Review, 104*(3), 1014–1039.

Esmail, A., & Everington, S. (1993). Racial discrimination against doctors from ethnic minorities. *BMJ: British Medical Journal, 306*(6879), 691–692.

Esmail, A., & Everington, S. (1997). Asian doctors are still being discriminated against. *BMJ: British Medical Journal, 314*(7094), 1619.

Evans, D. N. (2016). The effect of criminal convictions on real estate agent decisions in new York City. *Journal of Crime and Justice, 39*(3), 363–379.
Evans, D. N., & Porter, J. R. (2015). Criminal history and landlord rental decisions a New York quasi-experimental study. *Journal of Experimental Criminology, 11*(1), 21–42.
Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *The Review of Economics and Statistics, 96*(1), 119–134.
Farber, H. S., Silverman, D., & von Wachter, T. (2017). Factors determining callbacks to job applications by the unemployed: An audit study. *RSF: The Russell Sage Foundation Journal of the Social Science, 3*(3), 168–201.
Farkas, G., & Vicknair, K. (1996). Appropriate tests of racial wage discrimination require controls for cognitive skill: Comment on Cancio, Evans, and Maume. *American Sociological Review, 61*(4), 557–560.
Feins, J. D., & Bratt, R. G. (1983). Barred in Boston: Racial discrimination in housing. *Journal of the American Planning Association, 49*(3), 344–355.
Feldman, M. E., & Weseley, A. J. (2013). Which name unlocks the door? The effect of tenant race/ethnicity on landlord response. *Journal of Applied Social Psychology, 43*(S2), E416–E425.
Figinski, T. F. (2017). The effect of potential activations on the employment of military reservists: Evidence from a field experiment. *ILR Review, 70*(4), 1037–1056.
Firth, M. (1981). Racial discrimination in the British labor market. *Industrial and Labor Relations Review, 34*(2), 265–272.
Fix, M., & Struyk, R. J. (1993). *Clear and convincing evidence: Measurement of discrimination in America*. Washington, DC: The Urban Institute Press.
Fix, M., & Turner, M. A. (1999). Measuring racial and ethnic discrimination in America. In M. Fix & M. A. Turner (Eds.), *A national report card on discrimination in America: The role of testing* (pp. 7–26). Washington, DC: The Urban Institute.
Fix, M., Galster, G. C., & Struyk, R. J. (1993). An overview of auditing for discrimination. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 1–67). Washington, DC: The Urban Institute Press.
Friedman, S. (2015). Commentary: Housing discrimination research in the 21st century. *City, 17*(3), 143–150.
Friedman, S., Squires, G. D., & Galvan, C. (2010). *Cybersegregation in Boston and Dallas: Is neil a more desirable tenant than Tyrone or Jorge?* Available online at: http://mumford.albany.edu/mumford/cybersegregation/friedmansquiresgalvan.May2010.pdf
Fry, E. (1986). *An equal chance for disabled people? A study of discrimination in employment*. London: The Spastics Society, Campaigns and Parliamentary Department.
Furst, R. T., & Evans, D. N. (2016). Renting apartments to felons: Variations in real estate agent decisions due to stigma. *Deviant Behavior, 38*(6), 698–708.
Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.
Gaddis, S. M. (2017a). How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*(19), 469–489.
Gaddis, S. M. (2017b). Racial/ethnic perceptions from Hispanic names: Selecting names to test for discrimination. *Socius, 3*, 1–11.
Gaddis, S. M. (2017c). *Assessing immigrant generational status from names: Scientific evidence for experiments*. Available at SSRN: https://ssrn.com/abstract=3022217
Gaddis, S. M. (2017d). *Auditing audit studies: The effects of name perception and selection on social science measurement of racial discrimination*. Available at SSRN: https://ssrn.com/abstract=3022207
Gaddis, S. M. (2017e). *A field experiment on associate degrees and certificates: Statistical discrimination, stigma, signal boost, and signal saturation*. Available at SSRN: https://ssrn.com/abstract=3022203
Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. *The Annals of the American Academy of Political and Social Science, 660*(1), 282–299.

Gaddis, S. M., & Ghoshal, R. (2017). *Why do millennials engage in racial discrimination? (When they say they won't)*. Available at SSRN: https://ssrn.com/abstract=3022208

Galarza, F. B., & Yamada, G. (2014). Labor market discrimination in lima, Peru: Evidence from a field experiment. *World Development, 58*, 83–94.

Galarza, F. B., & Yamada, G. (2017). Triple penalty in employment access: The role of beauty, race, and sex. *Journal of Applied Economics, 20*(1), 29–47.

Galgano, S. W. (2009). Barriers to reintegration: An audit study of the impact of race and offender status on employment opportunities for women. *Social Thought & Research, 30*, 21–37.

Galster, G. (1990a). Racial steering in urban housing markets: A review of the audit evidence. *Review of Black Political Economy, 18*(3), 105–129.

Galster, G. (1990b). Racial discrimination in housing markets in the 1980s: A review of the audit evidence. *Journal of Planning Education and Research, 9*(3), 165–175.

Galster, G., & Constantine, P. (1991). Discrimination against female-headed households in rental housing: Theory and exploratory evidence. *Review of Social Economy, 49*(1), 76–100.

Galster, G., & Godfrey, E. (2005). By words and deeds: Racial steering by real estate agents in the U.S. in 2000. *Journal of the American Planning Association, 71*(3), 251–268.

Galster, G., MacDonald, H., & Nelson, J. (Forthcoming). What explains the differential treatment of renters based on ethnicity? New evidence from Sydney. *Urban Affairs Review*. https://doi.org/10.1177/1078087416679735.

Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). *Racial and gender discrimination in transportation network companies*. Available at NBER: http://www.nber.org/papers/w22776

Gell-Redman, M., Visalvanich, N., Crabtree, C., & Fariss, C. J. (2017). *It's all about race: How state legislators respond to immigrant constituents*. Available at SSRN: https://ssrn.com/abstract=2999173

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: W.W. Norton.

Ghoshal, R., & Gaddis, S. M. (2015). *Finding a roommate on craigslist: Racial discrimination and residential segregation*. Available at SSRN: https://ssrn.com/abstract=2605853

Giulietti, C., Tonin, M., & Vlassopoulos, M. (2015). *Racial discrimination in local public services: A field experiment in the U.S.* Available at SSRN: https://ssrn.com/abstract=2681054

Gneezy, U., & List, J. (2004). Are the disabled discriminated against in product markets? Evidence from field experiments. *American Economic Association Annual Meeting*.

Gneezy, U., List, J., & Price, M. K. (2012). *Toward an understanding of why people discriminate: Evidence from a series of natural field experiments*. Available at NBER: http://www.nber.org/papers/w17855

Goldberg, A., Mourinho, D., & Kulke, U. (1995). *Labour market discrimination against foreign workers in Germany*. Geneva: International Labour Office, mployment Department.

Graham, P., Jordan, A., & Lamb, B. (1990). *An equal chance or no chance? A study of discrimination against disabled people in the labour market*. London: The Spastics Society.

Gras, M., Bovenkerk, F., Gorter, K., Kruiswijk, P., & Ramsoedh, D. (1996). *Een schijn van kans: Twee empirische onderzoekingen naar discriminatie op grond van handicap en etnische afkomst*. Deventer: Gouda Quint.

Guryan, J., & Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economics Journal, 123*, F417–F432.

Hakken, J. (1979). *Discrimination against Chicanos in the Dallas rental housing market: An experimental extension of the housing market practices survey*. Washington, DC: U.S. Department of Housing and Urban Development.

Hansen, J. L., & James, F. J. (1987). Housing discrimination in small cities and nonmetropolitan areas. In G. A. Tobin (Ed.), *Divided neighborhoods: Changing patterns of racial segregation* (pp. 181–207). Thousand Oaks: Sage.

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics, 70*(2), 99–114.

Hanson, A., & Hawley, Z. (2014). Where does racial discrimination occur? An experimental analysis across neighborhood and housing unit characteristics. *Regional Science and Urban Economics, 44*, 94–106.

Hanson, A., & Santas, M. (2014). Field experiment tests for discrimination against Hispanics in the U.S. rental housing market. *Southern Economic Journal, 81*(1), 135–167.
Hanson, A., Hawley, Z., & Taylor, A. (2011). Subtle discrimination in the rental housing market: Evidence from E-mail correspondence with landlords. *Journal of Housing Economics, 20*(4), 276–284.
Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics, 92*, 48–65.
Hanson, A., Hawley, Z., & Martin, H. (2017). *Does differential treatment translate to differential outcomes for minority borrowers? Evidence from matching a field experiment to loan-level data*. Available at SSRN: https://ssrn.com/abstract=2945493
Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives, 12*(2), 101–116.
Heckman, J. J., & Siegelman, P. (1993). The urban institute audit studies: Their methods and findings. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 187–258). Washington, DC: The Urban Institute Press.
Hemker, J., & Rink, A. (Forthcoming). Multiple dimensions of bureaucratic discrimination: Evidence from German Welfare Offices. *American Journal of Political Science, 61*(4), 765–1022. https://doi.org/10.1111/ajps.12312.
Henry, F., & Ginzberg, E. (1985). *Who gets the work? A test of racial discrimination in employment*. Toronto: The Urban Alliance on Race Relations and the Social Planning Council of Metropolitan Toronto.
Heylen, K., & Van den Broeck, K. (2016). Discrimination and selection in the Belgian private rental market. *Housing Studies, 31*(2), 223–236.
Hipes, C., Lucas, J., Phelan, J. C., & White, R. C. (2016). The stigma of mental illness in the labor market. *Social Science Research, 56*, 16–25.
Hitt, M. A., Zikmund, W. G., & Pickens, B. A. (1982). Discrimination in industrial employment: An investigation of race and sex bias among professionals. *Work and Occupations, 9*(2), 217–231.
Hjarnø, J., & Jensen, T. (2008). *Discrimination in employment against immigrants in Denmark: A situation testing survey*. Geneva: International Labour Office.
Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & Community, 10*(4), 351–372.
Hubbock, J., & Carter, S. (1980). *Half a chance? A report on job discrimination against young blacks in Nottingham*. London: Commission for Racial Equality.
Hughes, D. A., Gell-Redman, M., Crabtree, C., Krishnaswami, N., Rodenberger, D., & Monge, G. (2017). *Who gets to vote? New evidence of discrimination among local election officials*. Working paper available online at http://cess.nyu.edu/wp-content/uploads/2017/02/Who-Gets-to-Vote.pdf
Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology, 60*(4), 669–692.
Jackson, M., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology, 39*, 27–49.
Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics, 19*(6), 824–832.
James, F., & DelCastillo, S. (1992). Measuring job discrimination: Hopeful evidence from recent audits. *Harvard Journal of African American Public Policy, 1*, 33–53.
James, F., McCummings, B., & Tynan, E. (1984). *Minorities in the sunbelt*. New Burnswick: Rutgers Center for Urban Policy Research.
Janusz, A., & Lajevardi, N. (2016). *The political marginalization of Latinos:Evidence from three field experiments*. Available at SSRN: https://ssrn.com/abstract=2799043
Johnson, E., & Lahey, J. (2011). The resume characteristics determining job interviews for middle-aged women seeking entry-level employment. *Journal of Career Development, 38*(4), 310–330.
Johnson, D. A., Porter, R. J., & Mateljan, P. L. (1971). Racial discrimination in apartment rentals. *Journal of Applied Social Psychology, 1*(4), 364–377.
Jolson, M. A. (1974). Employment barrier in marketing. *Journal of Marketing, 38*(2), 67–69.

Jowell, R., & Prescott-Clarke, P. (1970). Racial discrimination and white-collar workers in Britain. *Race, 11*(4), 397–417.
Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review, 13*(1), 1–20.
Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. (2016). Whitened resumes: Race and self-presentation in the labor market. *Administrative Science Quarterly, 61*(3), 469–502.
Kirschenman, J., & Neckerman, K. M. (1991). 'We'd love to hire them, but…': The meaning of race for employers. In C. Jencks & P. E. Peterson (Eds.), *The urban underclass* (pp. 203–232). Washington, DC: The Brookings Institution.
Kleykamp, M. (2009). A great place to start? The effect of prior military service on hiring. *Armed Forces & Society, 35*(2), 266–285.
Kovar, L. J. (1974). *Auditing real estate practices: A manual*. Philadelphia: National Neighbors.
Kroft, K., Lange, F., & Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics, 128*(3), 1123–1167.
Kugelmass, H. (2016). 'Sorry, I'm not accepting new patients': An audit study of access to mental health care. *Journal of Health and Social Behavior, 57*(2), 168–183.
Lahey, J. N. (2008). Age, women, and hiring: An experimental study. *Journal of Human Resources, 43*(1), 30–56.
Lahey, J. N., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior & Organization, 70*(3), 508–514.
Lahey, J. N., & Beasley, R. A. (2018). Technical aspects of correspondence studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Lanning, J. A. (2013). Opportunities denied, wages diminished: Using search theory to translate audit-pair study findings into wage differentials. *The B.E. Journal of Economic Analysis & Policy, 13*(2), 921–958.
Lauster, N., & Easterbrook, A. (2011). No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. *Social Problems, 58*(3), 389–409.
Leadership Council for Metropolitan Open Communities. (1975). *Investigation and auditing in fair housing cases*. Chicago: Leadership Council for Metropolitan Open Communities.
Lee, H.-A., & Khalid, M. A. (2016). Discrimination of high degrees: Race and graduate hiring in Malaysia. *Journal of the Asia Pacific Economy, 21*(1), 53–76.
Levinson, R. M. (1975). Sex discrimination and employment practices: An experiment with unconventional job inquiries. *Social Problems, 22*(4), 533–543.
List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics, 119*(1), 49–89.
Lodder, L. A., McFarland, S., & White, D. (2003). *Racial preference and suburban employment opportunities*. Chicago: Legal Assistance Foundation of Chicago.
Lucas, S. R. (2008). *Theorizing discrimination in an era of contested prejudice: Discrimination in the United States* (Vol. 1). Philadelphia: Temple University Press.
MacDonald, H., Galster, G., & Dufty-Jones, R. (Forthcoming). The geography of rental housing discrimination, segregation, and social exclusion: New evidence from Sydney. *Journal of Urban Affairs*. https://doi.org/10.1080/07352166.2017.1324247.
Massey, D. S., & Lundy, G. (2001). Use of black English and racial discrimination in urban housing markets: New methods and findings. *Urban Affairs Review, 36*(4), 452–469.
Maurer-Fazio, M. (2012). Ethnic discrimination in China's internet job board labor market. *IZA Journal of Migration, 1*(1), 12.
Maurer-Fazio, M., & Lei, L. (2015). 'As rare as a panda': How facial attractiveness, gender, and occupation affect interview callbacks at Chinese firms. *International Journal of Manpower, 36*(1), 68–85.
Mazziotta, A., Zerr, M., & Rohmann, A. (2015). The effects of multiple stigmas on discrimination in the German housing market. *Social Psychology, 46*(6), 325–334.

McClendon, G. H. (2016). Race and responsiveness: An experiment with south African politicians. *Journal of Experimental Political Science, 3*(1), 60–74.
McGinnity, F., & Lunn, P. D. (2011). Measuring discrimination facing ethnic minority job applicants: An Irish experiment. *Work, Employment and Society, 25*(4), 693–708.
McIntosh, N., & Smith, D. J. (1974). *The extent of racial discrimination*. London: PEP, The Social Science Institute.
McIntyre, S., Moberg, D. J., & Posner, B. Z. (1980). Preferential treatment in preselection decisions according to sex and race. *Academy of Management Journal, 23*(4), 738–749.
Mendez, M. S., & Grose, C. R. (2014). *Doubling down: Inequality in responsiveness and the policy preferences of elected officials*. Available at SSRN: https://ssrn.com/abstract=2422596
Michelitch, K. (2015). Does electoral competition exacerbate interethnic or Interpartisan economic discrimination? Evidence from a field experiment in market price bargaining. *American Political Science Review, 109*(1), 43–61.
Midtbøen, A. H. (2014). The invisible second generation? Statistical discrimination and immigrant stereotypes in employment processes in Norway. *Journal of Ethnic and Migration Studies, 40*(10), 1657–1675.
Midtbøen, A. H. (2015). The context of employment discrimination: Interpreting the findings of a field experiment. *The British Journal of Sociology, 66*(1), 193–214.
Midtbøen, A. H. (2016). Discrimination of the second generation: Evidence from a field experiment in Norway. *Journal of International Migration and Integration, 17*(1), 253–272.
Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science, 23*(7), 710–717.
Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology, 100*(6), 1678–1712.
Mincy, R. B. (1993). The urban institute audit studies: Their research and policy context. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 165–186). Washington, DC: The Urban Institute Press.
Mishel, E. (2016). Discrimination against queer women in the U.S. workforce: A resume audit study. *Socius: Sociological Research for a Dynamic World*. https://doi.org/10.1177/2378023115621316.
Murphy, J. K. (1972). *Audit handbook: Procedures for determining the extent of racial discrimination in apartment rentals*. Palo Alto: Mid-peninsula Citizens for Fair Housing.
Namingit, S., Blankenau, W., & Shwab, B. (2017). *Sick and tell: A field experiment analyzing the effects of an illness-related employment gap on the callback rate*. Working paper available at: https://economics.ku.edu/sites/economics.ku.edu/files/files/Seminar/papers1617/March31.pdf
Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *The Journal of Human Resources, 47*(4), 1128–1157.
Neumark, D. (Forthcoming). Experimental research on labor market discrimination. *Journal of Economic Literature*.
Neumark, D., & Rich, J. (2016). *Do field experiments on labor and housing markets overstate discrimination? A re-examination of the evidence*. Available at NBER: http://www.nber.org/papers/w22278
Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex discrimination in hiring in the restaurant industry: An audit study. *Quarterly Journal of Economics, 111*(3), 915–942.
Neumark, D., Burn, I., & Button, P. (2015). *Is it harder for older workers to find jobs? New and improved evidence from a field experiment*. Available at NBER: http://www.nber.org/papers/w21669
Neumark, D., Burn, I., & Button, P. (2016). Experimental age discrimination evidence and the Heckman critique. *American Economic Review: Papers and Proceedings, 106*(5), 303–308.
Neumark, D., Burn, I., Button, P., & Chehras, N. (2017). *Do state laws protecting older workers from discrimination laws reduce age discrimination in hiring? Experimental (and nonexperimental) evidence*. Available at SSRN: https://ssrn.com/abstract=2900439

Newburger, H. (1984). *Recent evidence on discrimination in housing*. Washington, DC: US Department of Housing and Urban Development, Office of Policy Development and Research.

Newman, J. M. (1978). Discrimination in recruitment: An empirical analysis. *Industrial and Labor Relations Review, 32*(1), 15–23.

Nunes, A. P., & Seligman, B. (1999). Treatment of Caucasian and African-American applicants by San Francisco Bay Area employment agencies: Results of a study utilizing 'testers'. In *The testing project of the impact fund*. San Francisco: The Impact Fund.

Nunes, A. P., & Seligman, B. (2000). A study of the treatment of female and male applicants by San Francisco Bay Area auto service shops. In *The testing project of the impact fund*. San Francisco: The Impact Fund.

Nunley, J. M., Owens, M. F., & Stephen Howard, R. (2011). The effects of information and competition on racial discrimination: Evidence from a field experiment. *Journal of Economic Behavior & Organization, 80*(3), 670–679.

Nunley, J. M., Pugh, A., Romero, N., & Alan Seals, R., Jr. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The B.E. Journal of Economic Analysis & Policy, 15*(3), 1093–1125.

Nunley, J. M., Pugh, A., Romero, N., & Alan Seals, R., Jr. (2016). College major, internship experience, and employment opportunities: Estimates from a resume audit. *Labour Economics, 38*, 37–46.

Nunley, J. M., Pugh, A., Romero, N., & Alan Seals, R., Jr. (2017). The effects of unemployment and underemployment on employment opportunities: Results from a correspondence audit of the labor market for college graduates. *ILR Review, 70*(3), 642–669.

Oh, S. J., & Yinger, J. (2015). What have we learned from paired testing housing markets? *Cityscape: A Journal of Policy Development Research, 17*(3), 15–59.

Ondrich, J., Stricker, A., & Yinger, J. (1998). Do real estate brokers choose to discriminate? Evidence from the 1989 housing discrimination study. *Southern Economic Journal, 64*(4), 880–901.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy, 3*(4), 148–171.

Oreopoulos, P., & Dechief, D. (2012). *Why do some employers prefer to interview Matthew, but not Samir? New evidence from Toronto, Montreal, and Vancouver*. Available at SSRN: https://ssrn.com/abstract=2018047

Page, M. (1995). Racial and ethnic discrimination in urban housing markets: Evidence from a recent audit study. *Journal of Urban Economics, 38*(2), 183–206.

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108*(5), 937–975.

Pager, D. (2007a). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science, 609*, 104–133.

Pager, D. (2007b). *Marked: Race, crime, and finding work in an era of mass incarceration*. Chicago: The University of Chicago Press.

Pager, D. (2016). Are firms that discriminate more likely to go out of business? *Sociological Science, 3*, 849–859.

Pager, D., & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review, 70*(3), 355–380.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181–209.

Pager, D., Western, B., & Bonikowski, B. (2009a). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review, 74*(5), 777–799.

Pager, D., Western, B., & Sugie, N. (2009b). Sequencing disadvantage: Barriers to employment facing young black and white men with criminal records. *The Annals of the American Academy of Political and Social Science, 623*, 195–213.

Patacchini, E., Ragusa, G., & Zenou, Y. (2015). Unexplored dimensions of discrimination in Europe: Homosexuality and physical appearance. *Journal of Population Economics, 28*(4), 1045–1073.

Pearce, D. M. (1979). Gatekeepers and Homeseekers: Institutional patterns in racial steering. *Social Problems, 26*(3), 325–342.
Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*(2), 262–289.
Pedulla, D. S. (2018). Emerging frontiers in audit study research: Mechanisms, variation, and representativeness. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector. *Labour Economics, 14*(3), 371–391.
Phillips, D. C. (2016a). *Do low-wage employers discriminate against applicants with long commutes? Evidence from a correspondence experiment*. Working paper. Available at: https://sites.google.com/site/davidcphillipseconomics/research
Phillips, D. C. (2016b). *Do comparisons of fictional applicants measure discrimination when search externalities are present? evidence from existing experiments*. Working paper. Available at: https://sites.google.com/site/davidcphillipseconomics/research
Phillips, D. C. (2017). Landlords avoid tenants who pay with vouchers. *Economics Letters, 151*, 48–52.
Pierné, G. (2013). Hiring discrimination based on national origin and religious closeness: Results from a field experiment in the Paris area. *IZA Journal of Labor Economics, 2*(4), 1–15.
Pitingolo, R., & Ross, S. L. (2015). Housing discrimination among available housing units in 2012: Do paired-testing studies understand housing discrimination? *City, 17*(3), 61–85.
Porter, N. D., Verdery, A. M., & Gaddis, S. M. (2017). *Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities*. Available at SSRN: https://ssrn.com/abstract=2844155
Purnell, B. (2013). *Fighting Jim crow in the county of kings: The congress of racial equality in Brooklyn*. Lexington: University Press of Kentucky.
Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology, 18*(1), 10–30.
Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology, 32*, 299–328.
Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. (2017). The persistence of racial discrimination: A meta-analysis of field experiments in hiring over time. *Proceedings of the National Academy of Sciences, 114*(41), 10870–10875.
Ravaud, J.-F., Madiot, B., & Ville, I. (1992). Discrimination towards disabled people seeking employment. *Social Science & Medicine, 35*(8), 951–958.
Riach, P. A. (2015). A field experiment investigating age discrimination in four European labour markets. *International Review of Applied Economics, 29*(5), 608–619.
Riach, P. A., & Rich, J. (1987). Testing for sexual discrimination in the labour market. *Australian Economic Papers, 26*(49), 165–178.
Riach, P. A., & Rich, J. (1991). Testing for racial discrimination in the labour market. *Cambridge Journal of Economics, 15*(3), 239–256.
Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal, 112*, F480–F518.
Riach, P. A., & Rich, J. (2006a). An experimental investigation of sexual discrimination in hiring in the English labor market. *The B.E. Journal of Economic Analysis & Policy, 6*(2), Article 1.
Riach, P. A., & Rich, J. (2006b). *An experimental investigation of age discrimination in the french labour market*. Available at SSRN: https://ssrn.com/abstract=956389
Riach, P. A., & Rich, J. (2007). *An experimental investigation of age discrimination in the spanish labour market*. Available at SSRN: https://ssrn.com/abstract=970498
Riach, P. A., & Rich, J. (2010). An experimental investigation of age discrimination in the English labor market. *Annals of Economics and Statistics, 99/100*, 169–185.
Rich, J. (2014). *What do field experiments of discrimination in markets tells us? A meta-analysis of studies conducted since 2000*. Available at SSRN: https://ssrn.com/abstract=2517887

Ridley, S., Bayton, J. A., & Outtz, J. H. (1989). *Taxi service in the District of Columbia: Is it influenced by Patron's race and destination?* Washington, DC: Lawyer's Committee for Civil Rights Under the Law.

Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review, 81*(6), 1097–1131.

Romei, A., & Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review, 29*(5), 582–638.

Rooth, D.-O. (2009). Obesity, attractiveness, and differential treatment in hiring: A field experiment. *Journal of Human Resources, 44*(3), 710–735.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523–534.

Ross, S. L., & Turner, M. A. (2005). Housing discrimination in metropolitan America: Explaining changes between 1989 and 2000. *Social Problems, 52*(2), 152–180.

Rowland, S., Jr. (1956). Race bias easing in churches here: Many congregations employ 'open door' policy—Housing discrimination persists. *New York Times, February, 12*, 1.

Roychoudhury, C., & Goodman, A. C. (1992). An ordered Probit model for estimating racial discrimination through fair housing audits. *Journal of Housing Economics, 2*(4), 358–373.

Roychoudhury, C., & Goodman, A. C. (1996). Evidence of racial discrimination in different dimensions of owner-occupied housing search. *Real Estate Economics, 24*(2), 161–178.

Ruffle, B. J., & Shtudiner, Z.'e. (2015). Are good-looking people more employable? *Management Science, 61*(8), 1760–1776.

Saltman, J. (1975). Implementing open housing laws through social action. *Journal of Applied Behavioral Science, 11*(1), 39–61.

Sharma, R., Mitra, A., & Stano, M. (2015). Insurance, race/ethnicity, and sex in the search for a new physician. *Economics Letters, 137*, 150–153.

Shin, R. Q., Smith, L. C., Welch, J. C., & Ezeofor, I. (2016). Is Allison more likely than Lakisha to receive a callback from counseling professionals? A racism audit study. *The Counseling Psychologist, 44*(8), 1187–1211.

Siddique, Z. (2011). Evidence on caste based discrimination. *Labour Economics, 18*(S1), S146–S159.

Smeesters, B., & Nayer, A. (1998). *La Discrimination a L'acces a L'emploi en Raison de L'origine Etrangere: Le Cas de le Belgique*. Geneva: International Labour Office, Conditions of Work Branch.

Smith, D. J. (2015). W. W. Daniel Obituary. *The guardian*, November 10. Available online at: https://www.theguardian.com/education/2015/nov/10/ww-daniel

Smith, S. L., & Cloud, C. (1996). The role of private, nonprofit fair housing enforcement organizations in lending testing. In J. Goering & R. Wienk (Eds.), *Mortgage lending, racial discrimination, and Federal Policy* (pp. 589–610). Washington, DC: Urban Institute Press.

Smith, R., & DeLair, M. (1999). New evidence from lender testing: Discrimination at the pre-application stage. In M. A. Turner & F. Skidmore (Eds.), *Mortgage lending discrimination: A review of existing evidence* (pp. 23–41). Washington, DC: Urban Institute Press.

Stone, A., & Wright, T. (2013). When your face doesn't fit: Employment discrimination against people with facial disfigurements. *Journal of Applied Social Psychology, 43*(3), 515–526.

Thanasombat, S., & Trasviña, J. (2005). Screening names instead of qualifications: Testing with emailed resumes reveals racial preferences. *AAPI Nexus: Policy, Practice, and Community, 3*(2), 105–115.

Tilcsik, A. (2011). Price and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology, 117*(2), 586–626.

Tunstall, R., Green, A., Lupton, R., Watmough, S., & Bates, K. (2014). Does poor neighborhood reputation create a neighbourhood effect on employment? The results of a field experiment in the UK. *Urban Studies, 51*(4), 763–780.

Turner, M. A., & James, J. (2015). Discrimination as an object of measurement. *Cityscape: A Journal of Policy Development Research, 17*(3), 3–14.

Turner, M. A., & Ross, S. L. (2003a). *Discrimination in metropolitan housing markets: Phase 2 – Asians and Pacific islanders final report*. Washington, DC: The Urban Institute Press.

Turner, M. A., & Ross, S. L. (2003b). *Discrimination in metropolitan housing markets: Phase 3 – Native Americans*. Washington, DC: The Urban Institute Press.
Turner, M. A., Mikelsons, M., & Edwards, J. (1990). *Analysis of steering in the housing discrimination study*. Washington, DC: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.
Turner, M. A., Fix, M., & Struyk, R. J. (1991a). *Opportunities denied, opportunities diminished: Racial discrimination in hiring*. Washington, DC: The Urban Institute Press.
Turner, M. A., Struyk, R. J., & Yinger, J. (1991b). *Housing discrimination study: Synthesis*. Washington, DC: U.S. Department of Housing and Urban Development.
Turner, M. A., Ross, S. L., Galster, G. C., & Yinger, J. (2002). *Discrimination in metropolitan housing markets: National results from Phase 1 HDS 2000*. Washington, DC: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.
Turner, M. A., Herbig, C., Kaye, D., Fenderson, J., & Levy, D. (2005). *Discrimination against persons with disabilities: Barriers at every step*. Washington, DC: The Urban Institute Press.
Turner, M. A., Santos, R., Levy, D. K., Wissoker, D., Aranda, C., & Pitingolo, R. (2013). *Housing discrimination against racial and ethnic minorities 2012*. Washington, DC: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.
Uggen, C., Vuolo, M., Lageson, S., Ruhland, E., & Whitham, H. K. (2014). The edge of stigma: An experimental audit of the effects of low-level criminal records on employment. *Criminology, 52*(4), 627–654.
Van der Bracht, Koen, A. C., & Van de Putte, B. (2015). The not-in-my-property syndrome: The occurrence of ethnic discrimination in the rental housing market in Belgium. *Journal of Ethnic and Migration Studies, 41*(1), 158–175.
Verhaeghe, P.-P., Van der Bracht, K., & Van de Putte, B. (2016). Discrimination of tenants with a visual impairment on the housing market: Empirical evidence form correspondence tests. *Disability and Health Journal, 9*(2), 234–238.
Verhaest, D., Bogaert, E., Dereymaeker, J., Mestdagh, L., & Baert, S. (Forthcoming). Do employers prefer overqualified graduates? A field experiment. *Industrial Relations*.
Vuolo, M., Uggen, C., & Lageson, S. (2016). Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes. *Sociological Methods & Research, 45*(2), 260–303.
Vuolo, M., Uggen, C., & Lageson, S. (2017). Race, recession, and social closure in the low-wage labor market: Experimental and observational evidence. *Research in the Sociology of Work, 30*, 141–183.
Vuolo, M., Uggen, C., & Lageson, S. (2018). To match or not to match? Statistical and substantive considerations in audit design and analysis. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Wallace, M., Wright, B. R. E., Zozula, C., Missari, S., Donnelly, C. M., & Wisnesky, A. S. (2012). A new approach for studying stratification and religion: Early results from a national internet-based field experiment Study of U.S. Churches. *Research in the Sociology of Work, 23*, 369–397.
Wallace, M., Wright, B. R. E., & Hyde, A. (2014). Religious affiliation and hiring discrimination in the American south: A field experiment. *Social Currents, 1*(2), 189–207.
Weichselbaumer, D. (2000). *Is it sex or personality? The impact of sex-stereotypes on discrimination in applicant selection*. University of Linz Economics Working Paper No. 0021. Available at SSRN: https://ssrn.com/abstract=251249
Weichselbaumer, D. (2003). Sexual orientation discrimination in hiring. *Labour Economics, 10*(6), 629–642.
Weichselbaumer, D. (2015). Testing for discrimination against lesbians of different marital status: A field experiment. *Industrial Relations: A Journal of Economy and Society, 54*(1), 131–161.
Weichselbaumer, D. (2016). *Discrimination against female migrants wearing headscarves*. Available at SSRN: https://ssrn.com/abstract=2842960
Weichselbaumer, D. (2017). Discrimination against migrant job applicants in Austria: An experimental study. *German Economic Review, 18*(2), 237–265.

White, A. R., Nathan, N. L., & Faller, J. K. (2015). What do I need to vote? Bureaucratic discretion and discrimination by local election officials. *American Political Science Review, 109*(1), 129–142.

Widner, D., & Chicoine, S. (2011). It's all in the name: Employment discrimination against Arab Americans. *Sociological Forum, 26*(4), 806–823.

Wienk, R. E., Reid, C. E., Simonson, J. C., & Eggers, F. J. (1979). *Measuring racial discrimination in American housing markets: The housing market practices survey*. Washington, DC: Department of Housing and Urban Development, Office of Policy Development and Research.

Wissoker, D. A., Zimmermann, W., & Galster, G. (1998). *Testing for discrimination in home insurance*. Washington, DC: The Urban Institute.

Wood, M., Hales, J., Purdon, S., Sejersen, T., & Hayllar, O. (2009). *A test for racial discrimination in recruitment practice in British cities*. Norwich: Department for Work and Pensions.

Wright, B. R. E., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. *Research in Social Stratification and Mobility, 34*, 111–126.

Wright, B. R. E., Wallace, M., Wisnesky, A. S., Donnelly, C. M., Missari, S., & Zozula, C. (2015). Religion, race, and discrimination: A field experiment of how American churches welcome newcomers. *Journal for the Scientific Study of Religion, 54*(2), 185–204.

Wysienska-Di Carlo, K., & Karpinski, Z. (2014). *Discrimination facing immigrant job applicants in Poland – results of a field experiment*. Presented at the XVIII ISA World Congress of Sociology.

Yinger, J. (1986). Measuring racial discrimination with fair housing audits: Caught in the act. *The American Economic Review, 76*(5), 881–893.

Yinger, J. (1991). Acts of discrimination: Evidence from the 1989 housing discrimination study. *Journal of Housing Economics, 1*, 318–346.

Yinger, J. (1993). Access denied, access constrained: Results and implications of the 1989 housing discrimination study. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 69–112). Washington, DC: The Urban Institute Press.

Yinger, J. (1995). *Closed doors, opportunities lost: The continuing costs of housing discrimination*. New York: Russell Sage Foundation.

Zhao, X., & Biernat, M. (2017). Welcome to the U.S. but change your name? Adopting Anglo names and discrimination. *Journal of Experimental Social Psychology, 70*, 59–68.

Zhou, X., Zhang, J., & Song, X. (2013). *Gender discrimination in hiring: Evidence from 19,130 resumes in China*. Available at SSRN: https://ssrn.com/abstract=2195840

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*(7), 1115–1134.

Zussman, A. (2013). Ethnic discrimination: Lessons from the Israeli online market for used cars. *The Economic Journal, 123*, F433–F468.

Chapter 2
Making It Count: Discrimination Auditing and the Activist Scholar Tradition

Frances Cherry and Marc Bendick Jr.

Abstract Discrimination auditing can usefully be viewed as part of a tradition of social science activist scholarship since World War II. This perspective suggests that the single-minded pursuit of methodological rigor, especially when reflected in exclusive reliance on documents-based audits, often sacrifices other characteristics historically associated with auditing's unique contributions to societal and scientific advancement. This chapter advocates and illustrates a balanced research agenda in which the most rigorous auditing studies are paralleled by others more directly in the activist scholar tradition. The hallmarks of that tradition are: in-person testers, the lived experience of discrimination, researcher-community partnerships, and goals beyond academic ones.

Keywords Situation testing · Participatory action research · Prejudice · Employment · Community organizing

2.1 Introduction

> The next day, Dorothy parked her 1946 Plymouth on Palmerston Boulevard. As she walked with Langston up the steps to the house, Dorothy noticed the red and white For Rent sign still on the door….
>
> "Not a good sign," Langston said. He rang the bell. Watson opened the door and stepped out onto the porch….
>
> "Well, we're here," Dorothy said. "We'd like to sign the contract, pay you, and bring our things in from the car."
>
> Langston watched the man open his mouth, close it, stop, pause.… Langston instantly knew that they would not get the flat. The coming refusal was as certain as the sunset – but Langston sensed that it would come in a distinct way….

F. Cherry (✉)
Department of Psychology, Carleton University, Ottawa, Canada
e-mail: francherry@cunet.carleton.ca

M. Bendick Jr.
Bendick and Egan Economic Consultants, Inc., Alexandria, VA, USA
e-mail: marc@bendickegan.com

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_2

> "I'm so sorry," Watson said, looking only at Dorothy, "and I hope you haven't been overly inconvenienced, but I have made other arrangements. A retired couple came by yesterday, after you left. They needed a quiet place, and they were prepared to take out a two-year lease, and I'm sorry, but I couldn't refuse them."
>
> "Yes, you could have," Dorothy shot back.
>
> Lawrence Hill, *Any Known Blood*

This passage by novelist Lawrence Hill fictionalizes his parents' experience as an interracial couple in post-World War II Toronto. His father, Daniel Hill, was completing his doctorate in sociology, after which he became the first full-time Director of the Ontario Human Rights Commission in 1962. His mother, Donna Bender Hill, worked for the Toronto Labour Committee for Human Rights documenting discrimination in employment, housing, and restaurants to promote anti-discrimination legislation. They thus simultaneously experienced and studied the segregated, racialized daily life of North America in the 1950s.

Any Known Blood goes on to describe auditing in the Hills' response to this encounter. They persuade a white couple to apply for the apartment they had been refused, and the landlord promptly offers the property to the new applicants. The landlord reassures the white couple that he knows of no black neighbors and that he would "draw the line there."

Writ large, such grassroots responses emerged as the American Civil Rights Movement of the 1950s and 1960s. Ordinary persons' lived experience of discrimination in real situations was a key resource mobilized by that movement to recruit activists, sway public opinion, secure anti-discrimination laws, and support their enforcement. Auditing was one of the alliances between civil rights advocates and professional researchers generating that resource.

This chapter examines the emergence, growth, and evolution of those efforts from the end of World War II through the present. It describes the work of "activist scholars" (Cherry 2004, 2008; Cherry and Borshuk 1998; Torre and Fine 2011; Torre et al. 2012) from multiple academic disciplines and their partnerships with a range of community members and advocacy organizations.

This history provides important guidance for today. As other chapters in this volume document, discrimination auditing in the Twenty-First Century often embodies considerable methodological rigor, especially when documents-based audits are conducted rather than audits involving live testers (Crabtree 2018; Gaddis 2018; Lahey and Beasley 2018). The creative search for rigor has undoubtedly enhanced the method's credibility and power in some ways. However, single-minded pursuit of rigor risks sacrificing other considerations historically associated with auditing's unique contributions to both society and science. This chapter calls for a more balanced research agenda in which the most rigorous auditing studies are paralleled by others more directly in the tradition of their historical precedents. The hallmarks of that tradition are: in-person testers, the lived experience of discrimination, researcher-community partnerships, and goals beyond academic publication.

2.2 Auditing in the Era of Gradualist Persuasion

The history of scholar activism underlying this call begins around the end of World War II. Throughout the late 1940s and 1950s, literally hundreds of civic, religious and educational organizations sought to improve inter-racial and inter-religious relations in the United States (Giles and Van Til 1946; Watson 1947; Williams 1947). These groups included state and local race relations committees (e.g., mayors' unity committees and state fair housing commissions), national race relations organizations (e.g., the National Association for the Advancement of Colored People, NAACP, and the National Urban League), faith-based organizations (e.g., the American Friends Service Committee and the National Conference of Christians and Jews), and educational institutions (e.g., the Bureau of Intercultural Education). President Truman's Committee on Civil Rights (President's Committee on Civil Rights 1947) was a national-level instance of the same approach.

Typically lacking legal enforcement powers and even statutes making discrimination illegal, these organizations relied primarily on persuasion and voluntary cooperation to advance their objectives. Information on the prevalence of discrimination and its adverse consequences was often their primary resource (Biondi 2003; Gordon 2015; Jackson Jr 1998, 2001; Jackson 1990; Richards 1997).

Three organizations – the Society for the Psychological Study of Social Issues (SPSSI), the NAACP, and the Commission of Community Interrelations (CCI) of the American Jewish Congress – were particularly prominent in connecting researchers and activists to generate that information (Cherry and Borshuk 1998; Jackson Jr 2001; Richards 1997). Advocating such linkages, a 1947 monograph for the Social Science Research Council by Cornell sociologist Robin Williams stated that the "necessary inclusion of fact finding among techniques of action suggests that research must be seen as an integral part of inter-group relations. Scientific study is a form of social action" (Williams 1947, p. 25). That same year, in book sponsored by the American Jewish Congress, psychologist Goodwin Watson cited the mantra of prominent social psychologist Kurt Lewin "No action without research; no research without action." Watson proposed university "action research service bureaus," particularly to evaluate the effectiveness of different approaches to reducing racial and religious discrimination (Watson 1947, p. 151).

At that time, Fisk University was already a center of scholarship embodying this approach. Envisioning fact finding as the basis for educational and legislative efforts, renowned sociologist Charles Johnson worked with Fisk's Department of Race Relations, which had been established in 1942 by the American Missionary Society as an action arm of Fisk's Social Sciences Department. Over many years, Johnson and his colleagues developed community self-surveys in which in-person and telephone interviews as well as other techniques were used to document race relations in multiple localities across the United States (Gilpin and Gasman 2003). Communities themselves collected the data, which then were analyzed by Fisk and published in the

form of statistical reports. For example, Sanders (2001) describes a 2-year community self-survey in Burlington Iowa conducted from 1949 through 1951.

A second center of scholarship in this tradition was the Commission on Community Interrelations (CCI) of the American Jewish Congress. Established by Kurt Lewin at the Massachusetts Institute of Technology, CCI engaged in various types of action research from the late 1940s through the early 1950s. The CCI approach moved even further than Fisk's in emphasizing community control over researcher leadership (Cherry and Borshuk 1998).

In particular, CCI researcher Claire Selltiz, working with housing activist Margot Haas Wormser, promoted researcher-community partnerships in which "citizens of a community are responsible for and participate in every phase of the investigation" (Wormser and Selltiz 1951). Their approach to community self-surveys reflected a concept enunciated by Kurt Lewin that local residents know best what would work for their community and how best to produce required changes.

Additionally, social psychologists at that time were beginning to argue that personal contact among individuals from different ethnic groups working together toward a common goal would itself importantly transform individual attitudes and behavior (Allport 1954). In particular, through collaborative efforts of Blacks and Whites studying their own communities, Whites would gain an understanding of the perspective of Black persons from whom they were normally segregated (Sanders 2001).

Auditing was one of several fact-gathering techniques promoted by Wormser and Selltiz as a logical outgrowth of this belief in the personally transformative experiences of people of different backgrounds working together. Structured audits in community self-surveys during the 1940s and 1950s can in part be thought of as formalizing the comparisons emerging naturally when members of different ethnic or religious groups examine prejudice and discrimination while sitting side by side.

In 1951, Wormser and Selltiz, produced a manual titled *How to Conduct a Self-Survey of Civil Rights* discussing participatory self-studies of housing, education, and public facilities and services (Wormser and Selltiz 1951). The manual's stated goal was to empower community groups to gather credible information in their own localities with only limited assistance from outside consultants. The manual included "test cases" – paired-comparison audits – as one method of data gathering.

Wormser and Selltiz played that consulting role in a pilot project in a small, highly segregated New Jersey town anonymized as "Northtown." They recruited local community organizations, both minority-based and not, to form a sponsoring committee. CCI staff provided technical advice on survey methods, trained volunteer interviewers, and participated in data interpretation. However, the local committee determined the effort's scope and style.

The Northtown sponsoring committee agreed to many of CCI's proposed approaches including random sampling, parallel interviews with both minority and non-minority individuals, and publication of findings with an action program based on those findings. However, it declined to implement other elements in the manual, including "test cases." The committee could not agree whether fact-finding should include observational procedures or only surveys and interviews; whether data should be used only in education and persuasion or put to more legal and confron-

tational uses; and the extent to which local employers should be publicly embarrassed. Moreover, during the McCarthyite anti-communist hysteria of the 1950s (Schrecker 1986), community activism was generally suspect, and rumors circulated that the Northtown study was the work of Communists. Within an already-controversial undertaking, "test cases" did not command consensus support.

Selltiz had more success in including "test cases" in a project of the Committee on Civil Rights of East Manhattan (CCREM), a group organized to address potential mistreatment of diplomats of color assigned to the new United Nations headquarters (Selltiz 1955; Biondi 2003). Along with a number of colleagues formerly at MIT, Selltiz was by then affiliated with the Research Center for Human Relations at New York University.

CCREM's first project was an audit study of Manhattan restaurants, for which Selltiz developed a study design, trained testers, and analyzed data. Pilot field work was conducted as thesis research by two Columbia University social work students (Landa and Littman 1950), and after methodological refinements based on that pilot, a full-scale study was conducted on 62 Manhattan restaurants in June of 1950. The study found no instances in which testers of color were refused service. However, differential treatment was documented in seating them in undesirable locations and providing poorer service (Schuman et al. 1983).

These findings were then moved into action, primarily in ways Selltiz (1955) described as "educational" and "persuasive" rather than "militant." CCREM representatives met with associations of restaurant owners and unions of restaurant employees seeking their pledge of equal treatment for all patrons. Letters requesting the same pledge were sent to the owners and managers of restaurants throughout the neighborhoods from which the audited restaurants had been sampled. Although no individual restaurant was publicly identified as having discriminated, private meetings were held with managers of some of those establishments. The Committee also issued a press release citing its findings and the restaurant industry's pledge. Members of the Committee were interviewed on radio, and 10,000 copies of a pamphlet, "Have You Heard What's Cooking?" were distributed. A follow-up audit study in 1952 found that discrimination had decreased significantly. However, it was not clear if the work of CCREM was the cause, as just before the new study, New York State legislation instituted more effective enforcement against discrimination in restaurants and other public accommodations (Selltiz, p. 25).

The relationships that communities formed with activist scholars at Fisk and CCI were not unique (Lambert and Cohen 1949; Torre and Fine 2011; Greenberg 1997). However, many other partnerships were short-lived and not well documented. Nevertheless, recognizing such activities, community self-studies and "test cases" were allocated a full chapter in the first edition of *Research Methods in Social Relations,* a textbook widely adopted in social science classes throughout the 1950s. That chapter described "test cases" as "the most direct method of getting information about possible discriminatory practices" and producing "evidence so clear-cut that it cannot be doubted." However, the chapter also noted that committees sponsoring community self-studies "represent a cross section of the community and are likely to be rather cautious" about approving "staged tests" that may

be seen as "dishonest" (Jahoda et al. 1951, pp. 621–622; see also Cherry 1995, and Torre and Fine 2011).

Towards the end of the 1950s, scholars' participation in action research waned. The anti-communist chill of cold war politics continued to reach onto college campuses to render academics' involvement in progressive political causes suspect (Schrecker 1986). Concurrently, some of the most prominent academic advocates of scholar-community partnerships were no longer active; for example, Kurt Lewin died in 1947, and Marie Jahoda moved to England.

Most importantly, academically prestigious, "cutting edge" attention in multiple social science disciplines was turning in other directions. In psychology, social psychologists increasingly defined their field as an experimental science marked by separation of research and application, experimenter and subject, and laboratories and real communities (Cherry 2009; Collier et al. 1991). In economics, the "institutionalist" tradition of researchers' involvement in practical issues such as development of the Social Security Act became less prestigious than analyses of economic behavior through mathematical models (Hodgson 2003). In sociology, academic interest in concrete interactions of individuals in real world situations was largely displaced by more abstract modeling of social structures and their functions (Fine 1995). These shifts were reflected in the evolution of *Research Methods in Social Relations.* By the 1959 edition of this textbook (Selltiz et al. 1959), the chapter on community self-surveys, including "test cases," was no longer included.

2.3 Auditing in the Era of Civil Rights Laws

During the 1960s, the Civil Rights Movement resulted in major federal legislation, prominently including the Equal Pay Act of 1963, Civil Rights Act of 1964, the Voting Right Act of 1965, and the Fair Housing Act of 1968. In concert with major Supreme Court decisions including *Brown v. Board of Education*, these developments collectively decimated *de jure* segregation across the southern states (Branch 1988).

With their counterpart statutes in many states and localities, these federal laws also prohibited *de facto* discrimination embodied in social custom rather than law and prevailing nation-wide rather than primarily in the South. These discriminatory practices were commonly embedded in the routine behavior of non-minority individuals and institutions, and the need to document that behavior sparked renewed interest in auditing. Housing was the first policy area in which auditing became central.

Fair housing committees of concerned citizens had existed in many localities across the nation since the 1940s or earlier, and many of them had applied auditing in investigating individual complaints against property owners, rental agents, real estate developers, mortgage lenders, and others (Yinger 1995, p. 28). In 1977, the U.S. Department of Housing and Urban Development (HUD) dramatically scaled up and systematized these local efforts by sponsoring a Housing Market Practices

Survey estimating the prevalence of these problems nation-wide, primarily with respect to African Americans. That project's 3,264 tests in 40 metropolitan areas documented widespread discrimination in both rental and owner-occupied housing (Wienk et al. 1979).

During the 1980s, auditing of housing discrimination matured into a sustainable activist scholar practice. In 1982, the U.S. Supreme Court's decision in *Havens Realty Corp. v. Coleman* unanimously upheld the standing of testers and the fair housing organizations employing them to bring litigation under the 1968 Fair Housing Act (Boggs et al. 1993, p. 346). A National Fair Housing Alliance was formed in 1988 and soon acquired 90 non-profit member organizations. At least 72 studies were conducted in individual cities, and in 1889, HUD sponsored a second nation-wide study, this time documenting the experiences of African Americans and Hispanics in 25 metropolitan areas (Yinger 1986, 1998). Findings from that HUD-sponsored study were credited with a major role in shaping the 1988 Amendments to the Fair Housing Act (Yinger 1998, p. 28).

Concurrently, auditing continued to be applied sporadically to discrimination in other aspects of daily life. Within housing, the initial focus on landlords' and owners' willingness to rent or sell individual properties broadened to examine "redlining" of neighborhoods in mortgage lending and homeowner insurance (Galster 1993). In 1988, social scientists from Howard University joined with a coalition of churches in Washington, DC to audit discrimination by taxicabs against African American riders and riders going to predominantly-African American neighborhoods. This project resulted in successful litigation against three cab companies and the important precedent of holding cab companies liable for discriminatory acts by individual drivers (Boggs et al. 1993, p. 348). A scholar at the Yale Law School used auditing to document race and gender discrimination in retail car sales (Ayres 1991). Psychologists used auditing procedures to measure the extent to which random samples of shoppers, motorists, and subway riders would assist strangers of different races (Cosby et al. 1980).

Such developments caught the attention of James Gibson, then head of the Equal Opportunity Program at the Rockefeller Foundation. Over the 1980s, Gibson had become increasingly concerned about erosion of public concern about racial discrimination throughout American society. Federal enforcement of anti-discrimination laws had substantially weakened since the 1981 advent of the conservative Reagan Administration (Clark 1989). Several Supreme Court decisions – notably *Regents of the University of California v. Bakke* in 1978 and *City of Richmond v. Croson* in 1989 – signaled increasing judicial skepticism of race-based affirmative action in education, employment, and government contracting. Perhaps most troubling, public opinion polls were reporting that increasing proportions of the non-minority U.S. population considered discrimination merely a problem of the past (Bendick 1999, p. 54). Gibson reasoned that auditing's success in housing might be expanded to provide fresh momentum to the flagging anti-discrimination cause.

Gibson translated this aspiration into substantial, multi-year grants to two non-profit organizations, The Urban Institute and the Washington Lawyers' Committee for Civil Rights and Urban Affairs. With respect to employment, these seminal grants

subsequently generated: a design for applying auditing to hiring (Bendick 1989); a study of employers' treatment of Hispanics under the federal Immigration Reform and Control Act of 1986 (Cross 1990); studies of hiring discrimination against African Americans (Turner et al. 1991; Bendick et al. 1994), Hispanics (Bendick et al. 1991), and older workers (Bendick et al. 1999); two successful testing-based lawsuits, one based on race and the other on gender (Boggs et al. 1993, pp. 362–363); a workshop training 72 academics and advocates on employment auditing (FEC 1993); and Congressional and state legislative testimony on the continued prevalence of discrimination and the continuing need for affirmative action (e.g., Bendick 1995). Promoting applications to other fields, the grants also supported two books (Fix and Struyk 1993; Fix and Turner 1999) offering creative ideas for auditing in retail sales, business lending, government contracting, and health care.

2.4 Audits Combining Rigor and Relevance

As other chapters in this volume reflect, over the two decades since these events, auditing research has expanded steadily, and its methodological sophistication has increased markedly (e.g., Edelman et al. 2017). These developments undoubtedly contribute to the method's credibility and influence today. But concurrently, some characteristics historically associated with auditing's unique contributions to societal and scientific advancement have become de-emphasized. In particular, four characteristics prominent in auditing prior to 2000 are relatively neglected today.

Human Testers One of these increasingly rare characteristics is audits employing live human testers rather than "correspondence studies" or "document studies" in which "testers" are presented only through written or electronic documents such as job resumes or mortgage applications. Because human testers are time-consuming and expensive to recruit and field, live tester studies tend to have samples on the order of 100 or fewer completed tests. In contrast, document audits, especially those in which application documents are computer generated, can afford samples of thousands. Larger samples increase the probability of observing statistically significant results and allow variations within the study design to examine multiple hypotheses and complex interactions. In addition, documents can be more rigorously controlled than the individual personality and appearance of live testers permit. Document studies thereby sidestep the inevitable skeptical questions about whether the tester within each pair who was treated less favorably was actually less qualified in some subtle, undocumented way (e.g., Heckman 1998).

These characteristics tend to make document studies easier to publish in scholarly outlets and more prestigious by conventional academic standards. However, their narrowness may limit the real-world applicability of their findings. Social psychologists have written extensively about the trade-off of experimental control and relevance since the late 1960s, when laboratory experimentation clearly became the preferred methodology in their field (Aronson and Carlsmith 1968;

Elms 1975; Cherry 2009). Parallel discussions can be found in other social science disciplines as well.

Have rigorously-controlled, document-based audit studies rendered live human audits obsolete? Are document-based audits the preferred approach when many selection processes today – such as job applications, college admissions, and loan applications – are commonly conducted at least partly on-line? The history of auditing suggests otherwise.

The most obvious reason for conducting in-person audits is that document studies typically cover only the initial stage of a selection process – for example, an employer's decision concerning which job applicants to invite to face-to-face interviews. Studies in which live testers have pursued the selection process all the way to the end often document discriminatory behavior appearing only at late stages – for example, where employers feel constrained by legal or social pressure to interview a racially-diverse slate of job candidates but then offer positions only to non-minority applicants (e.g., Bendick et al. 1994). In such circumstances, document audits of only the initial stages systematically under-estimate the overall prevalence of discrimination.

Correct measurement of outcomes is not the only benefit of deploying human testers. Contemporary concepts such as "implicit bias" and "micro-inequities" make clear that much discrimination today is unconscious and subtle (Jones et al. 2014). Especially when audio or video recordings provide word-for-word transcripts, in-person audits can illuminate the details of screening processes where such problems often lurk – for example, by documenting the influence of stereotypes on interviewers' judgments about job seekers' qualifications or the influence of in-group bias on interviewers' informal provision of assistance and encouragement to job seekers (Bendick and Nunes 2012).

The Lived Experience of Discrimination A second characteristic of historic auditing that is relatively rare today is efforts to communicate how discrimination feels to those experiencing it. Mainstream social and behavioral science research has tended to focus on the attitudes and behavior of the perpetrators of prejudice and discrimination rather than their targets. To be sure, such research is valuable in developing procedures for reducing harmful behavior. However, studying the other side of interactions between discriminators and their targets is also important for understanding the ways in which the targets – and the broader society – are harmed (e.g., Steeler 1997; Swim and Stangor 1998; Bendick and Nunes 2012).

The increasing visibility and power of community-based social movements – in forms such as marches and demonstrations advocating rights for women, visible minorities, person with disabilities, or based on sexual orientation or gender identity – often give a loud but unsystematic voice to persons experiencing discrimination in their daily lives. The community self-studies described earlier in this chapter began the process of systematically studying these perspectives by conducting extensive community interviews prior to developing their formal surveys (Wormser and Selltiz 1951). Live tester auditing studies tend to make these experiences even more central by collecting and disseminating detailed narratives of testers' actual experiences.

These narratives are particularly powerful in influencing the attitudes and behavior of individuals who have not personally experienced discrimination. As was discussed earlier in this chapter, when "test case" auditing started as part of community self-surveys, personal exposure to discrimination was intended to build understanding and empathy among minority and non-minority testers, members of local committees, and local residents. When auditing is part of lawsuits enforcing anti-discrimination law, the personal testimony of testers is often crucial in convincing judges and juries. And when audit findings have been presented to public policymakers – for example, in testimony to state or federal legislators – vivid anecdotes of testers' personal experiences tend to catch legislators' and media attention; as skilled public speakers know, human interest stories tend to be influential in ways that statistics alone are not (Bendick and Nunes 2012). Because auditing at its best provides *both stories and statistics*, it can uniquely retain the accuracy of the latter while mobilizing the persuasive power of the former.

The dominance of document-based auditing today tends to deprive audit studies of some of their most potentially influential findings. In addition, many audit studies are conducted with only academic peers as their target audience and incorporate few efforts to disseminate the findings more broadly. Career incentives in academia typically provide little credit for participating in community meetings, drafting pamphlets for distribution to consumers, or engaging with local news media. The earlier generations of activist scholars made such activities integral to their auditing research.

Community Partners A third characteristic of earlier auditing that is relatively rare today is partnerships between scholars and community groups.

In the early days of discrimination auditing, scholar-community partnerships were not formed merely for pragmatic reasons such community groups' need for trained researchers to analyze data or researchers' need to recruit community members as testers. Instead, both parties saw auditing as an important process of personal and organizational growth for the community groups, a strengthener of activist alliances through newly-shared perceptions and the team-building experience of working together. Promoting these processes was as integral an objective of the activity as were published reports.

Concurrently, working with a non-academic partner inevitably influences researchers as well. Interacting with individuals personally affected by discrimination often provides researchers with new insights into how institutions operate and new hypotheses to be studied. In addition, as the history in this chapter illustrates, community partners often shape studies in directions that researchers themselves would not necessarily have thought to initiate.

Decisions about paired versus unpaired audits provide an example of differences in priorities between researchers and community partners. The design of many audit studies today involves sending applications for the protected group tester and the control tester to different recipients, such as different employers or different mortgage lenders. The resulting response to the two groups of applications are then compared statistically to calculate overall rates of differences in treatment and to analyze the correlates of those differences. Researchers typically find this procedure

attractively efficient. However, when applications are presented to different recipients, it is not possible to identify individual decision-makers, such as employers or mortgage lenders, who have discriminated. The studies therefore document an abstract evil attributable only to the overall population from which the audit sample was drawn. Essentially, unpaired audits describe a *villainy without villains*. Community groups of adversely-affected individuals often want more specificity than that to facilitate concrete actions toward amelioration and often would not support such studies.

Objectives Beyond the Academic If properly conducted, unpaired auditing studies of the sort just described would be acceptable for scholarly publication. But that fact highlights an over-arching difference between many of today's auditing studies and auditing in the earlier activist scholar tradition. When Claire Selltiz and CCRM joined together to audit how restaurants treated customers of color, their goal was not to publish a study. Their goal was to reduce discriminatory behavior, with a study serving as an intermediate step. As was recounted earlier in this chapter, once the audit was completed, its findings were mobilized in multiple ways to promote behavioral change by restaurant owners, staff, and customers.

Activist scholars might or might not personally participate in such follow-up activities, but they must ensure that their studies are structured to support them. Invariably, this requirement means that audit studies take more time and resources. It often imposes study designs that might not be ideal from a pure research point of view, including use of live testers, having pairs of testers apply to the same company, and collecting narratives on testers' experiences in more detail than is statistically analyzable. In addition, researchers need to be prepared for the ego-crushing fact that the community groups see the audit as only a small piece of their long-term strategy.

2.5 Auditing Scholar-Activism Today

This section briefly sketches three examples of contemporary auditing scholar activism embodying most or all of the four characteristics just discussed.

Public Policy The most obvious example involves bringing audit evidence to bear on significant public policy issues.

In social policy today, few topics are more hotly debated than the complex relationships between race and the criminal justice system (Alexander 2012). One issue at the forefront of these discussions is so-called "ban the box" laws which limit employers' consideration of job applicants' criminal records in making hiring decisions. These laws have been adopted in 24 states and more than 150 localities, and have been considered in numerous additional jurisdictions (Rodriguez and Avery 2016).

When this issue is debated, the discussions almost inevitably cite the live-tester audit research of Harvard professor Devah Pager which examined the effect of criminal records on job applicants' chance of being hired and the interaction between

those effects and applicants' race (Pager 2007). A google search in November 2016 using the combined terms "ban the box" and "Devah Pager" produced more than 150 citations outside of her academic field of sociology, including by researchers in applied fields such as criminology, by news media, by advocacy organizations, and in public documents including hearings of the U.S. Congress and state legislatures and official EEOC guidelines for employers.

Pager herself has participated in these debates by, for example, serving on a government advisory panel of the National Academy of Sciences, writing opinion pieces in the news media, speaking at non-academic conferences of civil rights and criminal rights advocates, and testifying before public agencies such as the EEOC and the New York City Council (Pager 2016). However, the more fundamental way that her work reflects the activist scholar tradition is that the design of her audit studies provided information, both statistical and anecdotal, directly relevant to the "ban the box" issue. Had she not provided this information, more than a decade of policy debates would have been far less empirically grounded.

Legal Enforcement A second example involves using audit evidence in legal enforcement.

Make the Road New York (MTRNY) is a membership-based organization representing people of color in the New York City boroughs of Brooklyn, Queens, and Staten Island. After several of its transgender members complained about being turned down for jobs at fast food restaurants and a survey of their transgender members found that 59% of them reported similar experiences in a range of industries, the organization joined with economist Marc Bendick, Jr. to document the problem more systematically. They selected and trained two tester teams, one pairing a transgender and a cisgender woman and the other pairing a transgender and a cisgender man. Their resumes showed education and experience making them equally qualified for these positions. During the spring and summer of 2009, these teams applied for entry-level sales positions at 24 clothing retail stores in Manhattan. They found that, while transgender job applicants were often treated as politely as their testing partners, in some cases they were not, and the net rate of discrimination against the transgender testers was 42%.

The MTRNY report documenting this study (Bendick and Madar 2009) named the companies where discrimination had been encountered and included testers' "personal testimony" such as the following:

> When I went to apply at J Crew, I spoke to the manager, who said she was busy. I then spoke to a sales associate who gave me an application, but was vague about whether they were hiring. I filled out the application and submitted it to the manager then and there. She said she would give it to the hiring manager, and when I asked if they were currently hiring, she didn't say yes or no and said they would call me in for an interview. Twenty minutes later, my cismale partner went in, and…he ended up getting hired….I called twice over the next two weeks and they said they were still looking over applications and would call me. They never did.

> I was interviewed at a few of the stores that we tested. At some point during the interview, I would tell the employer that I was transgender and that my preferred pronoun was "he." In one interview, at DSW, I asked the manager whether I would feel comfortable working

> in the store as a transgender person and they said "that's up to you." I was also continually referred to as "she" despite my stated preference for the pronoun "he." Facing these kinds of experiences over and over again was humiliating. This process took an emotional toll on me…. Although this was a controlled research study, this experience mirrors my real life.

MTRNY then contacted the Civil Rights Bureau of the New York State Attorney General seeking relief under the anti-discrimination laws of New York City and New York State. It presented detailed documentation from audits of two employers, J. Crew and American Eagle Outfitters, each of which had been tested by both teams. The Attorney General found the evidence insufficient to proceed against J. Crew but opened an investigation of American Eagle. In May 2010, American Eagle agreed to a legal settlement that included adding gender identity and gender expression as a protected category in the company's anti-discrimination policy, training employees on transgender issues, training employees on how to file discrimination complaints, and revising the company dress code to no longer forbid men to wear women's clothing and men to wear women's clothing. In announcing this victory, MTRNY predicted that the settlement by such a prominent firm would encourage retailers nation-wide to rethink their policies and practices on the same issues (Taylor 2010).

Community Organizing A third example involves audit studies conducted as part of community organizing.

The Restaurant Opportunity Center United (ROC-U) is a non-profit worker center seeking to improve wages and working conditions for low-wage restaurant employees, including many people of color and recent immigrants. Starting in New York City with a core membership of restaurant workers who lost their jobs in the 2001 terrorist attack on the World Trade Center, the organization has developed into an influential presence throughout New York City's restaurant industry and then a multi-city group of similar organizations in a dozen states (Jayaraman 2013).

In achieving that growth, ROC-U has relied in part on a sequence of research activities first developed in New York and subsequently repeated in other cities. In New York, the sequence began with a research study in 2004 based on a structured survey of 530 workers, 45 semi-structured interviews with workers, and 35 semi-structured interviews with restaurant operators (ROC-NY 2005). Along with widespread issues of low wages, wage theft, and on-the-job harassment, this study highlighted occupational segregation in which immigrants and people of color were employed almost exclusively in low-paid positions in restaurant kitchens while white workers with similar skills, qualifications, and experience were overrepresented in server and manager jobs in the same restaurants' dining rooms. This pattern was especially stark in the city's upscale, "fine dining" establishments, where the earning opportunities were greatest.

To pursue that finding, ROC-NY joined with economist Marc Bendick, Jr. to conduct paired live-tester audits on the hiring of servers by fine dining restaurants in Manhattan (ROC-NY 2009; Bendick et al. 2010). During 2006 and 2007, the study recruited 37 volunteer testers primarily from among ROC members, partially

with the goal of actively engaging these individuals in the organization to strengthen their sense of affiliation. Testers were paired into teams of one white person and one person of color of the same gender, trained in effective interviewing techniques, provided with resumes showing equivalent qualifications, and assigned to apply for server positions at Manhattan's top restaurants, either responding to server openings advertised on-line or making "cold calls" at restaurants randomly selected from published lists of the city's most celebrated dining places. The study documented substantial differences in treatment adverse to persons of color at 31% of the restaurants audited and calculated that testers of color were only 55% as likely as their testing partners to receive job offers.

The study also illustrated employers' discriminatory behavior through narratives such as the following (Bendick et al. 2010, pp. 810–811):

> Answering a Craigslist advertisement, a white woman with no accent applied at an upscale Italian restaurant. She was promptly sent to an assistant manager, who, during an 18 minute interview, called her resume impressive, said that she presented herself well and that she'd "fit right in," and offered her specific work shifts. He emphasized that she would have opportunities to advance into management and that the restaurant would pay part of her health insurance. Meanwhile, a Chinese American woman with no accent, who had arrived half an hour before the white woman, was sent away with an interview appointment for the following day. During that interview, which lasted nine minutes, the same manager who had interviewed the white woman denied ever hearing of the restaurants on her resume and questioned whether she had worked in elegant establishments. He concluded that he would call her after consulting with other managers, but he never did.

The findings of this study were released at a well-attended "industry summit" hosted at his own restaurant by a celebrity chef who was a long-time ROC supporter. That release received considerable news media coverage, especially in the restaurant industry trade press, usually featuring vivid anecdotes that had been presented at the summit by testers themselves. The findings have been frequently cited by ROC throughout its organizing and lobbying activities in New York and elsewhere, whether appealing to potential worker members, restaurant operators, the news media, or public officials (ROC-U 2014). They are also reflected in ROC's list of "high road" restaurants they recommend to issue-conscious restaurant consumers (ROC-U 2016).

2.6 Barriers to Be Overcome

As multiple chapters in this book illustrate, the growing conceptual and methodological sophistication of much contemporary auditing research is impressive. However, the history reviewed in this chapter suggests that sophistication, experimental control, and academic credibility are often enhanced in ways that sacrifice other aspects of auditing that also contribute to the method's unique power. The history of scholar activism reviewed in this chapter suggests that such sacrifice is costly to both society and science. It is also often not necessary.

Of course, this history also documents that discrimination auditing in the activist scholar tradition is not without its challenges. Researchers often need to learn patience, a more user-friendly way of communicating, and sometimes humility to work effectively with non-academic partners whose perspectives and priorities often differ from their own. Incentives in the academic world tend to militate against overtly "applied" scholarship; involvement with community groups and in policy issues is sometimes viewed by academic colleagues, such as tenure committees, as a distraction at best and as indicating academic unworthiness at worst. Where activist audit studies achieve scholarly publication, that often occurs in interdisciplinary or second-tier journals that carry less academic prestige.

Furthermore, although some university Institutional Review Boards (IRBs) have enthusiastically endorsed audit studies, others have blocked them for political or other non-scientific reasons. One example is provided by Psychology Professor Jane Connor at the State University of New York at Binghamton (Connor 2000). In 1998, Connor taught a course on the Psychology of Racism in which her students watched a *Prime Time Live* television segment which followed two actors – John (White) and Glen (Black) – through a day of settling into a new city. Dressed similarly and coached to speak and behave in similar ways, the actors went apartment hunting, job seeking, and shopping while the video documented their strongly contrasting treatment.

Visible-minority students in Connor's class generally described the actors' experiences as similar to their own. In contrast, many non-minority students questioned whether, because the film has been made a decade earlier and in a different region of the United State and was made for television rather than as a scientific study, similar results would be found in their city. Accordingly, Connor organized a follow-up independent study course in which her students designed an in-person audit study of retailers in the Binghamton area. However, over the course of 2 years and multiple revisions, the proposal was never approved by the university's IRB. Connors' experiences offer a cautionary tale on the ethical and political controversies that auditing can trigger.

The twin goals of social progress and the advancement of human knowledge will be best served if the entire social science research community – funders, institutions, communities, and researchers themselves – invest creativity and effort in overcoming such obstacles and restoring the diversity of auditing activities more typical of auditing's earlier years. Without that breadth, auditing researchers may earn the title of scholars but not activist scholars, and all parties – activists, scholars, and society at large – will be the poorer for it.

References

Alexander, M. (2012). *The new Jim Crow, Mass incarceration in the age of colorblindness*. New York: New Press.

Allport, G. (1954). *The nature of prejudice*. Reading: Addison Wesley.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lidzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, 2nd ed., pp. 1–79). Reading: Addison-Wesley.

Ayres, I. (1991). Fair driving: Gender and race discrimination in retail car negotiations. *Harvard Law Review, 104*(4), 817–872.

Bendick, M., Jr. (1989). *Testing race discrimination in hiring: A research design*. Washington, DC: Bendick and Egan Economic Consultants, Inc.

Bendick, M., Jr. (1995, May 4). *Research evidence on discrimination and affirmative action in employment*. Testimony, Committee on the Judiciary, California State Assembly.

Bendick, M., Jr. (1999). Adding testing to the nation's portfolio of information on employment discrimination. In M. Fix & M. Turner (Eds.), *A national report card on discrimination: The role of testing* (pp. 47–68). Washington, DC: The Urban Institute.

Bendick, M., Jr., & Madar, C. (2009). *Transgender need not apply: Gender identity job discrimination in New York City's retail sector*. New York: Make the Road New York.

Bendick, M., Jr., & Nunes, A. (2012). Developing the research basis for controlling bias in hiring. *Journal of Social Issues, 68*, 238–263.

Bendick, M., Jr., Jackson, C., Reinoso, V., & Hodges, L. (1991). Discrimination against Latino job applicants: A controlled experiment. *Human Resource Management, 30*, 469–484.

Bendick, M., Jr., Jackson, C., & Reinoso, V. (1994). Measuring employment discrimination through controlled experiments. *Review of Black Political Economy, 23*, 25–48.

Bendick, M., Jr., Brown, L., & Wall, K. (1999). No foot in the door: An experimental study of employment discrimination against older workers. *Journal of Aging & Social Policy, 10*(4), 5–23.

Bendick, M., Jr., Rodriguez, R., & Jayaraman, S. (2010). Employment discrimination in upscale restaurants: Evidence from paired comparison testing. *The Social Science Journal, 47*, 802–818.

Biondi, M. (2003). *To stand and fight: The struggle for civil rights in postwar New York City*. Cambridge: Harvard University Press.

Boggs, R., Sellers, J., Bendick, M., & Jr, M. (1993). Use of testing in civil rights enforcement. In M. Fix & R. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 345–376). Washington, DC: Urban Institute Press.

Branch, T. (1988). *Parting the water, America in the King years 1954–1963*. New York: Simon & Schuster.

Cherry, F. (1995). *The "stubborn particulars" of social psychology: Essays on the research process*. London: Routledge.

Cherry, F. (2004). Kenneth B. Clark and social psychology's other history. In G. Philogene (Ed.), *Racial identity in context: The legacy of Kenneth B. Clark* (pp. 13–33). Washington, DC: American Psychological Association.

Cherry, F. (2008). Thomas F. Pettigrew: Building on the scholar/activist tradition in social psychology. In U. Wagner, L. T. Tropp, G. Finchilescu, & C. G. Tredoux (Eds.), *Improving intergroup relations: Building on the legacy of Thomas F. Pettigrew* (pp. 11–23). Oxford: Blackwell.

Cherry, F. (2009). Social psychology and social change. In D. Fox, I. Prilleltensky, & S. Austin (Eds.), *Critical psychology: An introduction* (2nd ed., pp. 93–109). Thousand Oaks: Sage.

Cherry, F., & Borshuk, C. (1998). Social action research and the Commission on Community Interrelations. *Journal of Social Issues, 54*, 119–142.

Clark, L. (1989). Ensuring equal employment opportunity through law. In D. L. Bawden (Ed.), *Rethinking employment policy* (pp. 155–167). Washington, DC: Urban Institute Press.

Collier, G., Minton, H. L., & Reynolds, G. (1991). *Currents of thought in American social psychology*. New York: Oxford University Press.

Connor, J. M. (2000). *Studying racial bias: Too hot to handle?* National Center for Case Study Teaching in Science. Case Collection http://sciencecases.lib.buffalo.edu/cs/collection/detail.asp?case_id=458&id=458

Cosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of black and white discrimination and prejudice. *A literature Review, Psych Bulletin, 87*, 546–563.

Crabtree, C. (2018). An introduction to conducting email audit studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Cross, H. (1990). *Employer hiring practices: Disparate treatment of Hispanic and Anglo job seekers*. Washington, DC: The Urban Institute.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*, 1–22.

Elms, A. C. (1975). The crisis in confidence in social psychology. *The American Psychologist, 30*, 967–976.

FEC. (1993). *Employment testing conference*. Washington, DC: Fair Employment Council of Greater Washington, Inc..

Fine, G. A. (Ed.). (1995). *A second Chicago School? The development of postwar American sociology*. Chicago: University of Chicago Press.

Fix, M., & Struyk, R. (Eds.). (1993). *Clear and convincing evidence, measurement of discrimination in America*. Washington, DC: The Urban Institute.

Fix, M., & Turner, M. A. (Eds.). (1999). *A national report card on discrimination in America: The role of testing*. Washington, DC: The Urban Institute.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Galster, G. (1993). Use of testing in investigating discrimination in mortgage lending and insurance. In M. Fix & R. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 287–334). Washington, DC: Urban Institute Press.

Giles, H. H., & Van Til, W. (1946). School and community projects. *The Annals of the American Academy of Political and Social Science, 244*, 34–41.

Gilpin, P. J., & Gasman, M. (2003). *Charles S. Johnson: Leadership beyond the veil in the age of Jim Crow*. Albany: State University of New York Press.

Gordon, L. N. (2015). The individual and "the general situation," The tension barometer and the race problem at the University of Chicago, 1947–1954. *Journal of the History of the Behavioral Sciences, 46*, 27–51.

Greenberg, C. (1997). Negotiating coalition: Black and Jewish civil rights agencies in the twentieth century. In M. Adams & J. Bracey (Eds.), *Strangers and neighbors: Relations between Blacks and Jews in the United States* (pp. 476–494). Amherst: University of Massachusetts Press.

Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives, 12*, 101–116.

Hodgson, G. M. (2003). John R. Commons and the Foundations of Institutional Economics. *Journal of Economic Issues, 37*, 547–576.

Jackson, W. A. (1990). *Gunnar Myrdal and America's conscience: Social engineering and racial liberalism, 1938–1987*. Chapel Hill: University of North Carolina.

Jackson, J. P., Jr. (1998). Creating a consensus: Psychologist, the Supreme Court, and school desegregation, 1952–1955. *Journal of Social Issues, 54*, 143–177.

Jackson, J. P., Jr. (2001). *Social scientists for social justice: Making the case against segregation*. New York: New York University Press.

Jahoda, M., Deutsch, M., & Cook, S. (1951). *Research methods in social relations with especial reference to prejudice. Part Two: Selected techniques*. New York: The Dryden Press.

Jayaraman, S. (2013). *Behind the kitchen door*. Ithaca: Cornell University Press.

Jones, J. M., Dovidio, J. F., & Vietze, D. L. (2014). *The psychology of diversity*. Malden: Wiley Blackwell.

Lahey, J., & Beasley, R. (2018). Technical aspects of correspondence studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Lambert, B. W., & Cohen, N. E. (1949). A comparison of different types of self-surveys. *Journal of Social Issues, 5*, 46–55.

Landa, P., & Littman, G. (1950). *A pilot study to test discriminatory practices against ethnic minority groups in public eating accommodations, An audit to determine the degree of discrimination practiced against Negroes in luncheonettes*. Unpublished thesis, New York School of Social Work, Columbia University.

Pager, D. (2007). *Marked: Race, crime, and finding work in an era of mass incarceration*. Chicago: University of Chicago Press.
Pager, D. (2016). *Curriculum vita*. Downloaded November 23, 2016 from www.scholar.harvard.edu
President's Committee on Civil Rights. (1947). *To secure these rights*. Washington, DC: Government Printing Office.
Richards, G. (1997). *"Race," racism, and psychology: Towards a reflexive history*. London: Routledge.
ROC-NY. (2005). *Behind kitchen doors: Pervasive inequality in New York City's thriving restaurant industry*. New York: Restaurant Opportunity Center of New York and the New York City Restaurant Industry Coalition.
ROC-NY. (2009). *The great service divide: Occupational segregation and inequality in the New York City restaurant industry*. New York: Restaurant Opportunity Center of New York and the New York City Restaurant Industry Coalition.
ROC-U. (2014). *Occupational segregation & inequality in the US restaurant industry*. New York: Restaurant Opportunity Center–United.
ROC-U. (2016). *Restaurant opportunities center united diners guide to ethical eating*. Downloaded October 17, 2016 from www.rocunited.org/dinersguide
Rodriguez, M. N., & Avery, B. (2016). *Ban the box, U.S. cities, counties, and states adopt fair hiring policies*. Downloaded November 23, 2016 from www.nelp.org
Sanders, K. M. (2001). The Burlington self-survey in human relations: Interracial efforts for constructive community change, 1949–1951. *The Annals of Iowa, 60*, 244–269.
Schrecker, E. (1986). *No ivory tower, McCarthyism and the universities*. New York: Oxford University Press.
Schuman, H., Singer, E., Donovan, R., & Selltiz, C. (1983). Discriminatory behavior in New York restaurants: 1950 and 1981. *Social Indicators Research, 13*, 69–83.
Selltiz, C. (1955). The use of survey methods in a citizens' campaign against discrimination. *Human Organization, 14*, 19–25.
Selltiz, C., Jahoda, M., Deutsch, M., & Cook, S. W. (1959). *Research methods in social relations [Rev]. Holt*. New York: Rinehart & Winston.
Steeler, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *The American Psychologist, 52*, 613–629.
Swim, J. K., & Stangor, C. (1998). *Prejudice: The target's perspective*. San Diego: Academic.
Taylor, S. (2010). *"Transgendering" personal appearance policies*. Downloaded November 22, 2016 from www.maketheroad.org
Torre, M. E., & Fine, M. (2011). A wrinkle in time: Tracing a legacy of public science through community self-surveys and participatory action research. *Journal of Social Issues, 67*, 106–121.
Torre, M. E., Fine, M., Stoudt, B. G., & Fox, M. (2012). Critical participatory action research as public science. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol 2. Research designs* (pp. 171–184). Washington, DC: American Psychological Association.
Turner, M., Fix, M., & Struyk, R. (1991). *Opportunities denied, opportunities diminished: Racial discrimination in hiring*. Washington, DC: The Urban Institute.
Watson, G. (1947). *Action for unity*. New York: Harper & Bros.
Wienk, R. E., Reid, C. E., Simonson, J. C., & Eggers, F. J. (1979). *Measuring racial discrimination in American housing markets: The housing market practices survey*. Washington, DC: Department of Housing and Urban Development, Office of Policy Development and Research.
Williams, R. M., Jr. (1947). *The reduction of intergroup tensions*. New York: Social Science Research Council.
Wormser, M. H., & Selltiz, C. (1951). *How to conduct a community self-survey of civil rights*. New York: Association Press.
Yinger, J. (1986). Measuring discrimination with fair housing audits: Caught in the act. *The American Economic Review, 76*, 881–893.
Yinger, J. (1995). *Closed doors, opportunities lost: The continuing costs of housing discrimination*. New York: Russell Sage Foundation.
Yinger, J. (1998). Evidence of discrimination in consumer markets. *The Journal of Economic Perspectives, 12*, 23–40.

Chapter 3
Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005

Stijn Baert

Abstract This chapter aims to provide an exhaustive list of all (i.e. 90) correspondence studies on hiring discrimination that were conducted between 2005 and 2016 (and could be found through a systematic search). For all these studies, the direction of the estimated treatment effects is tabulated. In addition, a discussion of the findings by discrimination ground is provided.

Keywords Hiring discrimination · Measurement · Correspondence experiments · Review · Ethnicity · Gender · Religion · Disability · Age · Military service · Wealth · Marital status · Sexual orientation · Political orientation · Union affiliation · Physical appearance

3.1 Triple Goal

The lack of labour market integration of vulnerable groups, such as refugees and other individuals with a migration background, the elderly, and people with a mental or physical health impairment, has received much attention in both policy and academic circles in the past decade (OECD 2008a, 2010). For policymakers, it is important to understand what factors cause this lack of integration in order to design the appropriate integration policies. Academic scholars have suggested discrimination in hiring as one important factor contributing to the poor labour market integration of these individuals (Altonji and Blank 1999; OECD 2008b). However, it is very challenging to measure discrimination in hiring, which makes it difficult to distinguish the effect of discrimination on employment from the effect of other factors, such as differences in human capital and other skills.

Historically, scholars have measured hiring discrimination through statistical analysis of non-experimental (survey or administrative) data. A commonly used

S. Baert (✉)
Ghent University, Ghent, Belgium
e-mail: stijn.baert@ugent.be

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_3

approach has been to try to control for as many observed individual factors as possible, such as education, experience, and occupation, and then interpret any unexplained part in employment between groups as pointing in the direction of discrimination (Blinder 1973; Oaxaca 1973). In general, these studies are likely to suffer from an important endogeneity bias, because job applicants who appear similar to researchers (except for their discrimination ground), based on non-experimental data, might in fact appear to be different to employers. For example, administrative data seldom contain information about language skills of individuals with a migration background, but this is likely to be observed by the employer, perhaps at a job interview. As long as not all relevant variables, taken into account by employers in making their hiring decisions, are controlled by the researcher, no conclusive proof of discrimination can be provided.

In response to this methodological problem, and inspired by the seminal work of Bertrand and Mullainathan (2004), scholars in labour economics, sociology of labour, and personnel psychology during the past decade have turned to so-called correspondence experiments to measure hiring discrimination (Gaddis 2018). In these experiments, fictitious job applications, differing only in a randomly assigned discrimination ground, are sent in response to real job openings. By monitoring the subsequent call-back from employers, unequal treatment based on this single characteristic is identified and can be given a causal interpretation.

Not surprisingly, given the seminal status of the correspondence experimentation framework[1] and the numerous academic studies that have adopted this framework, during the past years, scholars have written reviews and meta-analyses concerning this literature. We are aware of four such meta-studies: Bertrand and Duflo (2016), Neumark (in press), Rich (2014), and Zschirnt and Ruedin (2016). While all are inspiring high-quality syntheses, with excellent policy links and clever directions for further research, they share two limitations. First, these studies focus on an in-depth review of the field experimental evidence on labour market discrimination based on some grounds, while neglecting other grounds based on which unequal treatment is also forbidden. Second, none of these studies attempt to provide the reader with an exhaustive list of all experiments (conducted during a particular time frame). They all seem to focus on the better known (i.e. from their own country or highly cited) experiments while neglecting complementary work.

This chapter has a different ambition. It starts with identifying all discrimination grounds based on which unequal treatment is prohibited in at least one state of the United States and then provides the reader with a register of all correspondence experiments conducted (later than Bertrand and Mullainathan 2004) to measure these forms of discrimination. Given that the information provided for each study (i.e. particular treatment, country, and sign of the effect) is kept very limited—no effect size information is provided—this chapter has to be seen as a working instrument rather than as a classical review.

The register we will present serves three goals. First, it serves as a reference table to which later chapters of this book will refer. Second, and more broadly, it can be used

[1] Some deficiencies of the method were discussed in Chap. 2.

by scholars in search of a catalogue of all correspondence experiments on hiring discrimination based on a (cluster of) particular ground(s). Third, it implicitly indicates potentially fruitful directions for future correspondence experiments, as it unambiguously shows where the lacunae in this literature are, i.e. the discrimination grounds and regions to which researchers have paid little attention.

3.2 Scope

The register discussed in the next section is the result of a systematic search for correspondence experiments conducted after Bertrand and Mullainathan (2004) with the aim of measuring forms of unequal treatment in hiring which are prohibited by law in at least one state of the United States, i.e. the country in which the most correspondence experiments have been conducted. So, correspondence experiments included to assess the causal effect of, e.g., other cv characteristics such as juvenile delinquency, student employment and (former) unemployment spells were not included (Baert and Verhofstadt 2015; Baert et al. 2016d; Kroft et al. 2013; Eriksson and Rooth 2014).

Under US federal law, unequal treatment is forbidden based on nine (clusters of) discrimination grounds: (A) race and national origin, (B) gender and pregnancy, (C) religion, (D) disability, (E) (older) age, (F) military service or affiliation, (G) wealth, (H) genetic information, and (I) citizenship status.[2] With respect to (B), discrimination based on motherhood is also prohibited in Alaska[3] and California.[4] Finally, discrimination based on (J) marital status,[5] (K) sexual orientation and gender identity,[6] (L) political affiliation,[7] (M) union affiliation,[8] and (N) physical appearance[9] is forbidden in at least one state.

With this list of discrimination grounds at hand, a key word search (for the word groups 'correspondence test', 'correspondence experiment', 'correspondence study', 'fictitious resume', 'fictitious cv', 'fictitious application', and 'field experiment' in combination with 'discrimination') was conducted on three sources: Web of Science, Google Scholar, and the IZA Discussion Paper Series. This exercise was followed by the screening of all references in the relevant articles found and the screening of the studies citing these relevant articles.

[2] Source: https://www.eeoc.gov/

[3] Source: http://touchngo.com/lglcntr/akstats/Statutes/Title18/Chapter80/Section220.htm

[4] Source: http://www.dfeh.ca.gov/

[5] Source: http://touchngo.com/lglcntr/akstats/Statutes/Title18/Chapter80/Section220.htm

[6] Source: www.ilga.gov/legislation/ilcs/ilcs5.asp?ActID=2266

[7] Source: http://www.dfeh.ca.gov/

[8] Source: http://www.lexisnexis.com/hottopics/michie/

[9] Source: https://www.law.hawaii.edu/files/downloads/LAW%20589%20Appearance%20Discrimination_0.doc

3.3 The Register

Table 3.1 provides the reader with an overview of all studies (after Bertrand and Mullainathan 2004 of which we are aware that build on correspondence experiments aimed at measuring discrimination based on one of the grounds mentioned in the previous section. The unit of observation is the individual correspondence experiment. For each such experiment, there is a cell in column (3) of Table 3.1. Some cells contain more than one study, meaning that the studies exploited the same experimental data. Some studies focussed on more than one discrimination ground, and are therefore mentioned in more than one cell: Agerström et al. (2012), Albert et al. (2011), Arceo-Gomez and Campos-Vazquez (2014), Banerjee et al. (2009), Berson (2012), Capéau et al. (2012), Patacchini et al. (2015), Pierné (2013), and Stone and Wright (2013).

In total, we are aware of 90 correspondence experiments conducted between 2005 and 2016 with the aim of measuring discrimination based on prohibited grounds in at least one state of the United States. For 37 of these experiments, the focus (at least partly) was on measuring ethnic discrimination. Other commonly investigated discrimination grounds were gender (14 field experiments), age (11 experiments), and sexual orientation (12 experiments). In addition, at least five experiments focussed on religion, disability, and physical appearance as determinants of employers' hiring decisions. Only three experiments had a wealth-related focus and only two were related to military experience. Only one experiment has been conducted on hiring discrimination based on political affiliation and union membership. We are not aware of any experiments measuring unequal treatment based on genetic information, nor have any experiments—somewhat surprisingly given the massive migration flows to Europe in recent years—investigated citizenship status as a discrimination ground.

3.3.1 Treatment and Treatment Effects

As can be seen in column (1) of Table 3.1, for many discrimination grounds studied, a variety of particular treatments strategies have been used. For instance, ethnic origin is mostly revealed by means of the names of the candidates. The various minority groups studied are always groups that are substantially represented in the country where the data gathering took place. Alternative designs have disclosed ethnic origin by means of adding a resume picture or revealing one's nationality.

Column (4) shows the average treatment effect for each experiment (averaged across all vacancies and neglecting analyses by subsamples as presented in many studies). Overall, an overwhelming majority of the studies report negative treatment effects (i.e. discrimination of the group hypothesised to be discriminated against). More concretely, 80 (i.e. 78.4%) treatment effects are significantly negative, 17 (i.e.

Table 3.1 Register of correspondence experiments conducted between 2005 and 2016 with the aim of measuring discrimination based on prohibited grounds in US law

(1) Treatment	(2) Country of analysis	(3) Study	(4) Effect
A. Discrimination ground: race and national origin			
A.1. African (versus native) name	France	Cediey and Foroni (2008)	–
		Edo et al. (2013)	–
	US	Nunley et al. (2015)	–
		Gaddis (2015)	–
		Jacquemet and Yannelis (2012)	–
		Agan and Starr (2016)	–
A.2. African or Hispanic (versus native) name	Sweden	Bursell (2014)	–
	US	Darolia et al. (2016)	0
		Decker et al. (2015)	0
A.3. African, Asian, or German (versus native) name	Ireland	McGinnity and Lunn (2011)	–
A.4. African, Caribbean, Indian, or Pakistani (versus native) name	UK	Wood et al. (2009)	–
A.5. Albanian (versus native) name	Greece	Drydakis and Vlassis (2010) and Drydakis (2012a)	–
A.6. Antillean, Moroccan, Surinamese, or Turkish (versus native) name	Netherlands	Andriessen et al. (2012)	–
A.7. Arabian (versus native) name	Netherlands	Derous et al. (2012)	–
		Blommaert et al. (2014)	–
	Sweden	Agerström et al. (2012)	–
	US	Widner and Chicoine (2011)	–
A.8. Asian or Roma (versus native) name	Czech Republic	Bartoš et al. (2014)	–
A.9. Chinese, Greek, Indian, or Pakistani (versus native) name	US	Oreopoulos (2011)	–
A.10. Chinese, Indigenous, Italian, or Middle-Eastern (versus native) name	Australia	Booth et al. (2012)	–
A.11. Chinese, Nigerian, Serbian, or Turkish (versus native) name and appearance	Austria	Weichselbaumer (in press)	–
A.12. Congolese, Moroccan, Italian, or Turkish (versus native) name	Belgium	Capéau et al. (2012)	–
A.13. Ghanaian, Moroccan, Turkish, or Slovakian (versus native) name	Belgium	Baert et al. (2017)	–
A.14. Indigenous (versus native) name	Peru	Galarza and Yamada (2014)	–

(continued)

Table 3.1 (continued)

(1) Treatment	(2) Country of analysis	(3) Study	(4) Effect
A.15. Malaysian (versus Chinese) name	Malaysia	Lee and Khalid (2016)	–
A.16. Middle-Eastern (versus native) name	Sweden	Carlsson (2010), Carlsson and Eriksson (in press), Carlsson and Rooth (2007) and Carlsson and Rooth (2012)	–
		Attström (2007)	–
A.17. Mixed-race or Indigenous (versus white) skin	Mexico	Arceo-Gomez and Campos-Vazquez (2014)	–
A.18. Mongolian, Tibetan, or Uighur (versus native) name	China	Maurer-Fazio (2012)	–
A.19. Moroccan (versus native) name	France	Pierné (2013)	–
		Berson (2012)	–
		Duguet et al. (2010)	–
A.20. Pakistani (versus native) name	Norway	Midtbøen (2013) and Midtbøen (2016)	–
A.21. Turkish (versus native) name	Belgium	Baert et al. (2015)	–
		Baert and Vujić (2016)	–
	Germany	Kaas and Manger (2012)	–
A.22. Ukraine or Vietnamese (versus native) name	Poland	Wysienska-Di Carlo and Karpinski (2014)	–
B. Discrimination ground: gender and motherhood			
B.1. Being a mother (versus a childless woman)	US	Correll et al. (2007)	–
B.2. Being pregnant (versus revealing no pregnancy)	Belgium	Capéau et al. (2012)	–
B.3. Female (versus male) gender	Australia	Booth and Leigh (2010)	+
	Belgium	Capéau et al. (2012)	0
		Baert (2015) and Baert et al. (2016a)	0
	China	Zhou et al. (2013)	+
	France	Petit (2007)	–
		Berson (2012)	+
	Spain	Albert et al. (2011)	0
	Sweden	Agerström et al. (2012)	0
		Carlsson (2011)	0
	UK	Jackson (2009)	+
		Riach and Rich (2006b)	–
B.4. Transgender sexual identity	US	Make the Road New York (2010)	–
C. Discrimination ground: religion			
C.1. Muslim (versus majority religion)	France	Adida et al. (2010)	–
		Pierné (2013)	–
	India	Banerjee et al. (2009)	0

(continued)

Table 3.1 (continued)

(1) Treatment	(2) Country of analysis	(3) Study	(4) Effect
C.2. Pentecostal, Evangelical, or Jehovah's Witness (versus majority religion)	Greece	Drydakis (2010b)	–
C.3. Religious group membership	US	Wright et al. (2013)	–
C.4. Wearing headscarves	Germany	Weichselbaumer (2016)	–
D. Discrimination ground: disability			
D.1. Blindness, deafness, or autism	Belgium	Baert (2016)	–
D.2. Former depression	Belgium	Baert et al. (2016b)	–
D.3. Former mental illness (versus physical injury)	US	Hipes et al. (2016)	–
D.4. HIV	Greece	Drydakis (2010a)	–
D.5. Obesity	Sweden	Agerström and Rooth (2011) and Rooth (2009)	–
D.6. Spinal cord injury or Asperger's Syndrome	US	Ameri et al. (2015)	–
D.7. Unspecified physical disability	Belgium	Capéau et al. (2012)	–
D.8. Wheelchair user	UK	Stone and Wright (2013)	–
E. Discrimination ground: age			
E.1. Age 21 or age 27 (versus age 39 or age 47)	UK	Riach and Rich (2010)	–
E.2. Age 24 or age 25 (versus age 50 or age 51)	UK	Tinsley (2012)	–
E.3. Age 24 or age 28 (versus age 38)	Spain	Albert et al. (2011)	–
E.4. Age 27 (versus age 57)	France	Riach and Rich (2006a)	–
	Spain	Riach and Rich (2007)	–
E.5. Age 29, age 30, or age 31 (versus age 64, age 65, or age 66)	US	Neumark et al. (2015) and Neumark et al. (2016)	–
E.6. Age 35 or age 45 (versus age 50, age 55, or age 62)	US	Lahey (2008)	–
E.7. Age 35, age 47, or age 53 (versus age 23, age 35, or age 47)	Belgium	Capéau et al. (2012)	–
E.8. Age 46 (versus age 31)	Sweden	Ahmed et al. (2012)	–
E.9. Age 50 or age 44 (versus age 44 or age 38)	Belgium	Baert et al. (2016c)	–
E.10. Age 50 or older (versus younger)	US	Farber et al. (2016)	–
F. Discrimination ground: military service or affiliation			
F.1. Military work experience	Belgium	Baert and Balcaen (2013)	0
F.2. Military service	US	Kleykamp (2009)	+

(continued)

Table 3.1 (continued)

(1) Treatment	(2) Country of analysis	(3) Study	(4) Effect
G. Discrimination ground: wealth			
G.1. Residence in neighbourhood with poor (versus bland) reputation	UK	Tunstall et al. (2014)	0
G.2. Non-upper-caste (versus upper-caste)	India	Banerjee et al. (2009)	0
		Siddique (2011)	–
H. Discrimination ground: genetic information			
No related correspondence experiments found.			
I. Discrimination ground: citizenship status			
No related correspondence experiments found.			
J. Discrimination ground: marital status			
J.1. Married (versus unmarried)	Mexico	Arceo-Gomez and Campos-Vazquez (2014)	0
K. Discrimination ground: sexual orientation			
K.1. LGBT organisation member	Cyprus	Drydakis (2014)	–
	Germany	Weichselbaumer (2015)	–
	Greece	Drydakis (2009)	–
		Drydakis (2011)	–
		Drydakis (2012b)	–
	Italy	Patacchini et al. (2015)	0
	Sweden	Ahmed et al. (2013)	–
		Bailey et al. (2013)	0
	UK	Drydakis (2015)	–
	US	Tilcsik (2011)	–
		Mishel (2016)	–
K.2. Same-sex marriage partner	Belgium	Baert (2014)	0
L. Discrimination ground: political orientation			
L.1. Orientation of mentioned youth political organisation	Belgium	Baert et al. (2014)	0
M. Discrimination ground: union affiliation			
M.1. Youth union membership	Belgium	Baert and Omey (2015)	–
N. Discrimination ground: physical appearance			
N.1. Lower attractiveness of resume picture	Argentina	Lopez Bóo et al. (2013)	–
	Belgium	Baert (in press)	–
	China	Maurer-Fazio and Lei (2015)	–
	Israel	Ruffle and Shtudiner (2015)	–
	Italy	Patacchini et al. (2015)	0
N.2. Facial disfigurement (in resume picture)	UK	Stone and Wright (2013)	–

+ (0) ((−)) indicates an overall significantly positive (neutral) ((negative)) effect of the treatment in column (1) on call-back outcomes. Used abbreviations: *LGBT* Lesbian, Gay, Bisexual, and Transgender; *UK* United Kingdom; *US* United States. This register is kept updated at the author's homepage [http://users.UGent.be/~sbaert]

16.7%) are insignificantly different from 0, and 5 (i.e. 4.6%) are significantly positive.[10]

Most of the cases document discrimination against ethnic minorities. There are two important exceptions with respect to this empirical pattern. First, in two recent studies with experiments conducted in the United States, no ethnic discrimination in hiring was found (Darolia et al. 2016; Decker et al. 2015). Second, in Malaysia the (expected) unfavourable treatment of the ethnic *majority* was found (Lee and Khalid 2016).[11] In addition, research in Belgium (Baert and Vujić 2016; Baert et al. 2015, 2017) revealed situations in which ethnic discrimination disappeared there, i.e. when ethnic minorities mentioned volunteer work for mainstream organisations, when they applied for occupations in which labour market tightness was high, and when they had many years of work experience. For an in-depth review of a selection of the studies in Panel A of Table 3.1, we refer to Bertrand and Duflo (2016), Neumark (in press), Rich (2014), and Zschirnt and Ruedin (2016).

With respect to evidence on gender discrimination, i.e. the experiments comparing call-back for male and female candidates, the evidence is very mixed. This is related to the particular occupations tested. Indeed, many authors mentioned that gender discrimination was heterogeneous by occupational characteristics (Baert et al. 2015; Petit 2007; Carlsson 2011). On the other hand, a significant penalty for being pregnant or being a mother was found in a study from Belgium and one from the United States, respectively (Capéau et al. 2012; Correll et al. 2007). Disclosing one's transgender identity was found to be detrimental to labour market success in the United States (Make the Road New York 2010).

With respect to discrimination based on religion, a majority of the studies focussed on the signal of being a Muslim (directly mentioned or indicated by means of a resume picture in which headscarves were worn), compared with being a Christian (in countries where Christianity was the majority religion). Affiliation with Islam always yielded lower call-back rates (Adida et al. 2010; Banerjee et al. 2009; Pierné 2013; Weichselbaumer 2016). Somewhat surprisingly, no correspondence experiments have been conducted yet with respect to other leading religions (e.g., Hinduism, Buddhism, and Judaism) as well as to various folk religions.

Remarkably, all experiments on discrimination against the disabled have focussed on different dimensions of disability. Thus, we are in favour of replication studies for this dimension of discrimination. Nevertheless, each form of disability revealed in the hiring process seems to result in adverse hiring outcomes. The same is true with respect to age discrimination: across all studies listed in Table 3.1, older age is always punished.

[10] These numbers do not sum up to 90, as some studies were included multiple times in Table 3.1 (as mentioned in the first paragraph of Sect. 3.3).

[11] In general, comparing the results across the rows of Table 3.1 is very tricky, as the experiments differed substantially with respect to at least the following characteristics of their design: (i) region of the experiment; (ii) experimental population (e.g., with respect to age and education level); and (iii) sectors, occupations, and vacancies tested.

A minority sexual orientation, revealed by means of mentioning membership in a rainbow organisation or the name of one's (same-sex) marital partner in the resume, has a non-positive effect on employment opportunities. Including an attractive facial picture (compared to a less attractive one) with one's resume has a beneficial effect. Finally, Table 3.1 lists little evidence for non-negative effects of military service and higher wealth (Baert and Balcaen 2013; Kleykamp 2009), a negative effect of trade union membership (Baert and Omey 2015), and zero effects for marital status (Arceo-Gomez and Campos-Vazquez 2014) and political affiliation (Baert et al. 2014).

3.3.2 Country of Analysis

Column (2) of Table 3.1 shows that the summarised literature on labour market discrimination is unbalanced with respect to the country of analysis. Grouped at the continental level, 59 of the 90 correspondence experiments were conducted in Europe, compared to 20 in North America, only 7 in the largest continent of Asia, 2 in South America, 2 in Australia, and none in Africa.

At the country level, most experiments (19) were conducted in the United States. The European countries of Belgium (13 experiments), France (8 experiments), Greece (6 experiments), Sweden (9 experiments), and the UK (8 experiments) are clearly overrepresented. On the other hand, these European countries are, together with the United States, the only ones in which within-country comparisons can be made of the discrimination measured for different grounds. In 6 of the 10 largest countries by population (Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, and Russia), no correspondence experiments have been conducted yet.

3.4 Conclusion

This chapter provided the reader with a catalogue of all correspondence experiments on hiring discrimination conducted after Bertrand and Mullainathan (2004) that could be found through a systematic search. It shows that these experiments have focussed on a few specific grounds for discrimination (race, gender, religion, disability, age, sexual orientation, and physical appearance). An overwhelming majority of these studies reported unfavourable treatment of the group hypothesised to be discriminated against. On the other hand, other topical forms of potential hiring discrimination (e.g., based on genetic information, citizenship status, or political orientation) have hardly been assessed. Moreover, in 6 of the 10 largest countries by population, no correspondence experiments have been conducted yet.

The register presented in Table 3.1—enriched with hyperlinks to the electronic versions of the included studies—is kept updated at the author's homepage [http://users.UGent.be/~sbaert].

References

Adida, C. L., Laitin, D. D., & Valfort, M. A. (2010). Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 22384–22390.

Agan, A., & Starr, S. B. (2016). *Ban the box, criminal records, and statistical discrimination: A field experiment* (pp. 16–012). University of Michigan Law School, Law and Economics Research Paper Series.

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96*, 790–805.

Agerström, J., Björklund, F., Carlsson, R., & Rooth, D. O. (2012). Warm and competent Hassan = Cold and incompetent eric: A harsh equation of real-life hiring discrimination. *Basic and Applied Social Psychology, 34*, 359–366.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2012). Does age matter for employability? A field experiment on ageism in the Swedish labour market. *Applied Economics Letters, 19*, 403–406.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2013). Are gay man and lesbians discriminated against in the hiring process? *Southern Economic Journal, 79*, 565–858.

Albert, A., Escot, L., & Fernández-Cornejo, J. A. (2011). A field experiment to study sex and age discrimination in the Madrid labour market. *International Journal of Human Resource Management, 22*, 351–375.

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. *Handbook of Labor Economics, 3*, 3143–3259.

Ameri, M., Schur, L., & Meera, A. (2015). *The disability employment puzzle: A field experiment on employer hiring behavior* (NBER Working Paper Series 21560).

Andriessen, I., Nievers, E., Dagevos, J., & Faulk, L. (2012). Ethnic discrimination in the Dutch Labor Market: Its relationship with job characteristics and multiple group membership. *Work and Occupations, 39*, 237–239.

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2014). Race and marriage in the labor market: A discrimination correspondence study in a developing country. *American Economic Review, 104*, 376–380.

Attström, K. (2007). *Discrimination against native Swedes of immigrant origin in access to employment*. Geneva: International Labour Office.

Baert, S. (2014). Career lesbians. Getting hired for not having kids? *Industrial Relations Journal, 45*, 543–561.

Baert, S. (2015). Field experimental evidence on gender discrimination in hiring: Biased as Heckman and Siegelman predicted? *Economics: The Open-Access, Open-Assessment E-Journal, 9*, 25.

Baert, S. (2016). Wage subsidies and hiring chances for the disabled: Some causal evidence. *European Journal of Health Economics, 17*, 71–86.

Baert, S. (in press). Facebook profile picture appearance affects recruiters' first hiring decisions. *New Media & Society, 17*, 1377–1396.

Baert, S., & Balcaen, P. (2013). The impact of military work experience on later hiring chances in the civilian labour market. Evidence from a field experiment. *Economics: The Open-Access, Open-Assessment E-Journal, 7*, 37.

Baert, S., & Omey, E. (2015). Hiring discrimination against pro-union applicants: The role of union density and firm size. *Economist, 163*, 263–280.

Baert, S., & Verhofstadt, E. (2015). Labour market discrimination against former juvenile delinquents: Evidence from a field experiment. *Applied Economics, 47*, 1061–1072.

Baert, S., & Vujić, S. (2016). Immigrant volunteering: A way out of labour market discrimination? *Economics Letters, 146*, 95–98.

Baert, S., Jong, A. Pin, R., De Freyne, L., & Parmentier, S. (2014). *Political ideology and labour market discrimination*. Conference presentation at the spring meeting of Young Economists 2014.

Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2015). Is there less discrimination in occupations where recruitment is difficult? *ILR Review, 68*, 467–500.

Baert, S., De Pauw, A. S., & Deschacht, N. (2016a). Do employer preferences contribute to sticky floors? *ILR Review, 69*, 714–736.

Baert, S., De Visschere, S., Schoors, K., Vandenberghe, D., & Omey, E. (2016b). First depressed, then discriminated against? *Social Science & Medicine, 170*, 247–254.

Baert, S., Norga, J., Thuy, Y., & Van Hecke, M. (2016c). Getting grey hairs in the labour market. A realistic experiment on age discrimination. *Journal of Economic Psychology, 57*, 86–101.

Baert, S., Rotsaert, O., Verhaest, D., & Omey, E. (2016d). Student employment and later labour market success: No evidence for higher employment chances. *Kyklos, 69*, 401–425.

Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination? *Economics Letters, 155*, 35–38.

Bailey, J., Wallace, M., & Wright, B. (2013). Are gay men and lesbians discriminated against when applying for jobs? A four-city, internet-based field experiment. *Journal of Homosexuality, 60*, 873–894.

Banerjee, A., Bertrand, M., Datta, S., & Mullainathan, S. (2009). Labor market discrimination in Delhi: Evidence from a field experiment. *Journal of Comparative Economics, 37*, 14–27.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2014). *Attention discrimination: Theory and field experiments with monitoring information acquisition* (IZA Discussion Paper Series 8058).

Berson, B. (2012). *Does competition induce hiring equity? Documents de travail du Centre d'Economie de la Sorbonne 12019*.

Bertrand, M., & Duflo, E. (2016). Review on field experiments on discrimination. In A. Banerjee & E. Duflo (Eds.), *Handbook of field experiments*. Cambridge: Abdul Latif Jameel Poverty Action Lab.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991–1013.

Blinder, A. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources, 8*, 436–455.

Blommaert, L., Coenders, M., & van Tubergen, F. (2014). Discrimination of Arabic named applicants in the Netherlands: An internet-based field experiment examining different phases in online recruitment procedures. *Social Forces, 92*, 957–982.

Booth, A. L., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters, 107*, 236–238.

Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics, 74*, 547–573.

Bursell, M. (2014). The multiple burdens of foreign-named men—evidence from a field experiment on gendered ethnic hiring discrimination in Sweden. *European Sociological Review, 30*, 399–409.

Capéau, B., Eeman, L., Groenez, S., & Lamberts, M. (2012). *Two concepts of discrimination: Inequality of opportunity versus unequal treatment of equals* (Ecore Discussion Paper Series 2012–58).

Carlsson, M. (2010). Experimental evidence of discrimination in the hiring of first- and second-generation immigrants. *Labour, 24*, 263–278.

Carlsson, M. (2011). Does hiring discrimination cause gender segregation in the Swedish labor market? *Feminist Economics, 17*, 71–102.

Carlsson, M., & Eriksson, S. (in press). Do attitudes expressed in surveys predict ethnic discrimination? *Ethnic and Racial Studies, 40*(10), 1739–1757.

Carlsson, M., & Rooth, D. O. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics, 14*, 716–729.

Carlsson, M., & Rooth, D. O. (2012). Revealing taste-based discrimination in hiring: A correspondence testing experiment with geographic variation. *Applied Economics Letters, 19*, 1861–1864.

Cediey, E., & Foroni, F. (2008). *Discrimination in access to employment on grounds of foreign origin in France*. Geneva: International Labour Office.

Correll, S. J., Benard, B., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology, 112*, 1297–1338.

Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2016). Race and gender effects on employer interest in job applicants: New evidence from a resume field experiment. *Applied Economics Letters, 23*, 853–856.

Decker, S. H., Ortiz, N., Cassia, S., & Hedberg, E. (2015). Criminal stigma, race, and ethnicity: The consequences of imprisonment for employment. *Journal of Criminal Justice, 43*, 108–121.

Derous, E., Ryan, A. M., & Nguyen, H. H. (2012). Multiple categorization in resume screening: Examining effects on hiring discrimination against Arab applicants in field and lab settings. *Journal of Organizational Behavior, 33*, 544–570.

Drydakis, N. (2009). Sexual orientation discrimination in the labour market. *Labour Economics, 16*, 364–372.

Drydakis, N. (2010a). Labour discrimination as a symptom of HIV: Experimental evaluation: The greek case. *Journal of Industrial Relations, 52*, 201–217.

Drydakis, N. (2010b). Religious affiliation and labour bias. *Journal for the Scientific Study of Religion, 49*, 472–488.

Drydakis, N. (2011). Women's sexual orientation and labor market outcomes in Greece. *Feminist Economics, 11*, 89–117.

Drydakis, N. (2012a). Estimating ethnic discrimination in the labour market using experimental data. *Southeast European and Black Sea Studies, 12*, 335–355.

Drydakis, N. (2012b). Sexual orientation and labour relations: New evidence from Athens, Greece. *Applied Economics, 44*, 2653–2665.

Drydakis, N. (2014). Sexual orientation discrimination in the Cypriot labour market. Distastes or uncertainty? *International Journal of Manpower, 35*, 720–744.

Drydakis, N. (2015). Measuring sexual orientation discrimination in the UK's labour market; a field experiment. *Human Relations, 68*, 1769–1796.

Drydakis, N., & Vlassis, M. (2010). Ethnic discrimination in the Greek labour market: Occupational access, insurance coverage and wage offers. *Manchester School, 78*, 201–218.

Duguet, E., Leandri, N., L'Horty, Y., & Petit, P. (2010). Are young french jobseekers of ethnic immigrant origin discriminated against? A controlled experiment in the Paris area. *Annals of Economics and Statistics / Annales d'Économie et de Statistique, 99*(100), 187–215.

Edo, A., Jacquemet, N., & Yannelis, C. (2013). *Language skills and homophilous hiring discrimination: Evidence from gender- and racially-differentiated applications* (CES Working Paper Series, pp. 13–58).

Eriksson, S., & Rooth, D. O. (2014). Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. *American Economic Review, 104*, 1014–1039.

Farber, H. S., Silverman, D., & von Wachter, T. (2016). Factors determining callbacks to job applications by the unemployed: An audit study. *American Economic Review, 106*, 314–318.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*, 1451–1479.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Galarza, F. B., & Yamada, G. (2014). Labor market discrimination in Lima, Peru: Evidence from a field experiment. *World Development, 58*, 83–94.

Hipes, C., Lucas, J., Phelan, J. C., & White, R. C. (2016). The stigma of mental illness in the labor market. *Social Science Research, 56*, 16–25.

Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *British Journal of Sociology, 60*, 669–692.
Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics, 19*, 824–832.
Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review, 13*, 1–20.
Kleykamp, M. A. (2009). Great place to start? The effect of prior military service on hiring. *Armed Forces & Society, 35*, 266–285.
Kroft, K., Lange, F., & Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *Quarterly Journal of Economics, 128*, 1123–1167.
Lahey, J. N. (2008). Age, women, and hiring: An experimental study. *Journal of Human Resources, 43*, 30–56.
Lee, H. A., & Khalid, M. A. (2016). Discrimination of high degrees: Race and graduate hiring in Malaysia. *Journal of the Asia Pacific Economy, 21*, 53–76.
Lopez Bóo, F., Rossi, M., & Urzúa, S. (2013). The labor market return to an attractive face: Evidence from a field experiment. *Economics Letters, 118*, 170–172.
Make the Road New York. (2010). *Transgender need not apply: Gender identity job discrimination in New York City's retail sector*. New York: Make the Road New York.
Maurer-Fazio, M. (2012). Ethnic discrimination in China's internet job board labor market. *IZA Journal of Migration, 1*, 1–24.
Maurer-Fazio, M., & Lei, L. (2015). As rare as a panda. How facial attractiveness, gender, and occupation affect interview callbacks at Chinese firms. *International Journal of Manpower, 36*, 68–85.
McGinnity, F., & Lunn, P. D. (2011). Measuring discrimination facing ethnic minority job applicants: An Irish experiment. *Work, Employment and Society, 25*, 693–708.
Midtbøen, A. H. (2013). The invisible second generation? Statistical discrimination and immigrant stereotypes in employment processes in Norway. *Journal of Ethnic and Migration Studies, 40*, 1657–1675.
Midtbøen, A. H. (2016). Discrimination of the second generation: Evidence from a field experiment in Norway. *Journal of International Migration and Integration, 17*, 253–272.
Mishel, E. (2016). Discrimination against Queer women in the U.S. Workforce: A Résumé audit study. *Socius, 2*, 2378023115621316.
Neumark, D. (in press). Experimental research on labor market discrimination. *Journal of Economic Literature*.
Neumark, D., Burn, I., & Button, P. (2015). *Is it harder for older workers to find jobs? New and improved evidence from a field experiment* (NBER Working Paper Series 21669).
Neumark, D., Burn, I., & Button, P. (2016). Experimental age discrimination evidence and the Heckman critique. *American Economic Review, 106*, 303–308.
Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *B.E. Journal of Economic Analysis & Policy, 15*, 1093–1125.
Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review, 14*, 693–709.
OECD. (2008a). *Jobs for immigrants. Labour market integration in France, Belgium, the Netherlands and Portugal*. Paris: OECD Publishing.
OECD. (2008b). *The price of prejudice: Labour market discrimination on the grounds of gender and ethnicity*. Paris: OECD Publishing.
OECD. (2010). *Sickness, disability and work. Breaking the barriers—A synthesis of findings across OECD countries*. Paris: OECD Publishing.
Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy, 3*, 148–171.
Patacchini, E., Ragusa, G., & Zenou, Y. (2015). Unexplored dimensions of discrimination in Europe: Homosexuality and physical appearance. *Journal of Population Economics, 28*, 1045–1073.

Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector. *Labour Economics, 14*, 371–391.
Pierné, G. (2013). Hiring discrimination based on national origin and religious closeness: Results from a field experiment in the Paris area. *IZA Journal of Labor Economics, 2*, 4.
Riach, P. A., & Rich, J. (2006a). *An experimental investigation of age discrimination in the French labour market* (IZA Discussion Paper Series 2522).
Riach, P. A., & Rich, J. (2006b). An experimental investigation of sexual discrimination in hiring in the English labor market. *B.E. Journal of Economic Analysis & Policy, 5*, 1–22.
Riach, P. A., & Rich, J. (2007). *An experimental investigation of age discrimination in the Spanish labour market* (IZA Discussion Paper Series 2654).
Riach, P. A., & Rich, J. (2010). An experimental investigation of age discrimination in the English labor market. *Annals of Economics and Statistics, 99*(100), 169–185.
Rich, J. (2014). *What do fild experiments of discrimination in markets tell us? A meta analysis of studies conducted since 2000* (IZA Discussion Paper Series 8584).
Rooth, D. O. (2009). Obesity, attractiveness, and differential treatment in hiring: A field experiment. *Journal of Human Resources, 44*, 710–735.
Ruffle, B., & Shtudiner, Z. (2015). Are good-looking people more employable? *Management Science, 61*, 1760–1776.
Siddique, Z. (2011). Caste-based discrimination: Evidence and policy. *Labour Economics, 18*, S146–S159.
Stone, A., & Wright, T. (2013). When your face doesn't fit: Employment discrimination against people with facial disfigurements. *Journal of Applied Social Psychology, 43*, 515–526.
Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology, 117*, 586–626.
Tinsley, M. (2012). *Too much to lose: Understanding and supporting Britain's older workers*. London: Policy Exchange.
Tunstall, R., Green, A., Lupton, R., Watmough, S., & Bates, K. (2014). Does poor neighbourhood reputation create a neighbourhood effect on employment? The results of a field experiment in the UK. *Urban Studies, 51*, 763–780.
Weichselbaumer, D. (2015). Testing for discrimination against lesbians of different marital status: A field experiment. *Industrial Relations, 54*, 131–161.
Weichselbaumer, D. (2016). *Discrimination against female migrants wearing headscarves* (IZA Discussion Paper Series 10217).
Weichselbaumer, D. (in press). Discrimination against migrant job applicants in Austria: An experimental study. *German Economic Review.*
Widner, D., & Chicoine, S. (2011). It's all in the name: Employment discrimination against Arab Americans. *Sociological Forum, 26*, 806–823.
Wood, M., Hales, J., Purdon, S., Sejersen T., & Hayllar O. (2009). *A test for racial discrimination in recruitment practice in British cities* (DWP Research Reports 607)
Wright, B. R. E., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. *Research in Social Stratification and Mobility, 34*, 111–126.
Wysienska-Di Carlo, K., & Karpinski, Z. (2014). *Discrimination facing immigrant job applicants in Poland—Results of a field experiment*. In Conference presentation at the XVIII ISA World Congress of Sociology.
Zhou, X, Zhang, J, & Song, X. (2013). *Gender discrimination in hiring: Evidence from 19,130 resumes in China*. https://doi.org/10.2139/ssrn.2195840. Accessed 11 Nov 2016.
Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*, 1115–1134.

Part II
The Method of Audit Studies: Design, Implementation, and Analysis

Chapter 4
Technical Aspects of Correspondence Studies

Joanna Lahey and Ryan Beasley

Abstract This chapter discusses technical concerns and choices that arise when crafting a correspondence or audit study using external validity as a motivating framework. The chapter discusses resume creation, including power analysis, choice of inputs, pros and cons of matching pairs, solutions to the limited template problem, and ensuring that instruments indicate what the experimenters want them to indicate. Further topics about implementation include when and for how long to field a study, deciding on a participant pool, and whether or not to use replacement from the participant pool. More technical topics include matching outcomes to inputs, data storage, and analysis issues such as when to use clustering, when not to use fixed effects, and how to measure heterogeneous and interactive effects. The chapter ends with a technical checklist that experimenters can utilize prior to fielding a correspondence study.

Keywords Audit studies · Correspondence review · Field experiment · Experimental design · Experimental analysis · Power analysis · Matched pairs · Internal and external validity · Heterogeneous and interactive effects

---

This chapter has been prepared for the volume *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis. The authors thank all of the researchers who have provided feedback on the Resume Randomizer program, and Joanna Lahey also thanks the many editors who, through referee requests, have forced her to keep up-to-date on the state of correspondence studies. Thanks also to Patrick Button, S. Michael Gaddis, R. Alan Seals, Jill E. Yavorsky, and an anonymous reviewer for helpful feedback.

J. Lahey (✉)
Texas A&M University and NBER, College Station, TX, USA
e-mail: jlahey@tamu.edu

R. Beasley
SimQuest Solutions, Inc., Boston, MA, USA

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_4

4.1 External Validity and the Audit Study

External validity[1] concerns drive many technical choices in correspondence studies. While it is tempting to believe that a single study can answer "Is there *X* discrimination?" or "Do for profit colleges and universities provide value?", an audit study can only test a limited market for a specific subset of applicants during a specific time period. It is therefore vital to design the experiment carefully, so that it is clear how the study's results will further knowledge. In general, we will use examples from employment audit studies to illustrate ideas in this chapter, but correspondence review is a powerful tool that can be used more broadly to study differential treatment across many settings.

Ultimately, the external validity of an experiment is constrained by each decision made in the design. For example, studies that only apply to ads within big cities may not be applicable to smaller towns or rural areas. Similarly, resumes in which every person over the age of 50 also has a multi-year employment gap may provide results that are driven by the age, by the gap, or by their combination. Results may translate completely, partially, or not at all to other populations and settings depending on the similarities between the experiment and the population or setting of interest. Questions to ask in the initial design phase include: Who will you use as participants? When and for how long will you field the study? Where will you get correspondence inputs? Taking the design as a whole, for what group will the results of the experiment be externally valid?

The most important external validity question to ask is whether the indicator that separates the treatment group(s) from the control group tests what it is supposed to test and does not inadvertently test something different. (This type of threat to validity is termed "ecological validity" in some social sciences. See Brewer, M. (2000) for more information.) Examples of indicators include names for race discrimination (e.g. Bertrand and Mullainathan 2004; Gaddis 2017a, b; Oreopoulos 2011), date of school graduation for age (e.g. Lahey 2008; Neumark et al. 2016), or name of school when testing the effect of colleges (e.g. Gaddis 2015; Deming et al. 2016; Darolia et al. 2015). It is important that the indicator indicates what it is intended to indicate and is not just measuring that a resume or other piece of correspondence is unusual. For example, indicating age by date of high school graduation is something that most real job seekers do, but listing age on a resume is frowned upon in the United States. The most troubling examples occur when the unintended "unusual" negative signal only signals negatively for treatment. For example, putting union membership on a nursing resume is not just testing the effect of union status, and similarly, listing number of children does not just indicate that the applicant is a mother, but that the applicant does not know not to put things on a resume that do not belong there. (This type of threat to validity is termed "construct validity" in many social sciences. See Trochim and Donnelly (2006) or Shadish

[1] External validity here is defined as results from the experiment being generalizable to other populations and settings (as in Stock and Watson 2011).

et al. (2002) for detailed discussions of construct validity.) It is our belief that if this indicator is "unusual" rather than something that normally appears in resumes that the study should not be performed and resources should be devoted elsewhere because the study results will not be generalizable. As a caveat, avoiding testing "unusual" items does not mean that it would be inappropriate for someone from a discriminated group to apply for a position. Men do apply for clerical jobs, women do apply for truck driving positions, and minorities do apply for high powered jobs; the general equilibrium employment ratio does not necessarily indicate that applicants are not interested in a job.

A final external validity concern is the open question of how call-backs translate into job offers. It is important to note that this translation will be different for different types of jobs. Although there are scattered answers to this question from various industry surveys and studies (e.g. Barron et al. 1985; Howden 2016; Maurer 2016) and studies of job seekers (e.g. Moynihan et al. 2003), we do not know what the average translation from call-back to job offer is or how this number varies by industry, occupation, unemployment rates, educational level of the applicant and so on. During the design phase, it is important to investigate how actual job-seekers enter the selection process and to be careful about making broader claims on how interviews translate into hires.

While decisions driven by external validity motivations should guide study design, this chapter will also discuss technical considerations including power analysis, matching outcomes to inputs, data storage, how to deal with changes while fielding the experiment, and post-collection data analysis concerns. The chapter ends with a technical checklist to aid researchers.

4.2 Determining the Pool

4.2.1 Participants

In a correspondence audit, the participants will generally be companies, landlords, purchasers, and so on, that is, members of the group whose biases are being tested, not the hypothetical applicants. Results will only be externally valid to the larger population from which the participants tested in the experiment are drawn. Results may be different if participants are drawn from, for example, urban vs. rural areas, from the Southern US region vs. the Northeast, or from Belgium compared to Mexico. Studies that cover a broad geographical area may be affected by heterogeneous effects across different cities or states or countries, and will need to have a large enough sample size to be able to detect, and preferably test, those differences. On the other hand, because effects may be different across regions, results for one area are only externally valid for that area, thus broader geographic coverage may give effects that are externally valid on average even if they do not provide a good representation for what any individual faces in a smaller market.

The choice of how to find participants is important, especially in these times of rapid technological change. For example, in the past, classified ads in the newspaper were a primary way that jobs were posted, which meant that early studies could use Sunday want-ads in order to run an experiment that was externally valid for a large population of job seekers. Companies still take out classified ads in trade magazines/journals in order to reach a specific audience, but online resources have risen in prominence. Craigslist in particular is a popular website for researchers and has increased its market penetration across the United States. Online job sites such as Indeed.com or snagajob.com have also become more prevalent and potentially more useful than their earlier incarnations ten or twenty years ago.[2] Not all sites have the same job ad penetration across geographic markets or fields and a researcher should investigate these differences before committing to a specific source of advertising. What may be a good source of jobs for computer science positions may not be as good a source for nursing positions. Other researchers may avoid general want-ad postings and pick specific companies to target with unsolicited applications. This method could include, for example, targeting Fortune 500 companies (e.g. Bendick et al. 1999) or all the hospitals and nursing homes in a specific area. In some regions researchers can use job banks, such as Belgium's job bank (e.g. Baert and Dieter 2014). Some professions rely on walk-ins or networking for the majority of their job openings and as such are more difficult to test in a correspondence framework (Holzer 1996). Again, these decisions should be guided by both feasibility and external validity of the sample to the question you are interested in answering. Carbonaro and Schwarz (2018) in Chap. 7 of this volume will go into more detail on these concerns.

When choosing specific participants, it is important to have a systematic rule in place that provides the most externally valid sample possible. For example, a simple rule could be to apply to all ads posted on your online site during the course of the study, checking for new ads once a day. Drawing a sample may require more complicated rules that should be decided on in advance or during a pilot study.

4.2.2 Length of Time

Another important choice is when and for how long to field a survey. Using employment audits as an example, it is important to think about how business cycles might affect hiring. Results during the holiday hiring season, when many lower-level companies and job seekers are looking for holiday work, may be different than results during a hiring lull. A study that is externally valid for college students looking for summer work will not be externally valid for applicants searching for a full-time career. Similarly, many companies will advertise for positions after the first of the year, and industries tied to fair weather will advertise in the Spring (see JOLTS 2016 for data on hiring seasonality).

[2] Note that, as always, you should check with your IRB about what job sites are allowable based on their Terms of Service (TOS). Some IRB allow TOS violations that could happen in the normal course of use, whereas others do not allow such usage.

An additional factor that may determine how long you field a study is more practical—the necessary number of observations to find statistical significance for a given power, which we discuss below in the section on power analysis. Similarly, the expected response rate can mechanically affect a study's ability to obtain variation even with a large sample size. Response rates can depend on the type of participants, the number of participants, whether the participants are actively hiring, annual cycles, long-term recessions or expansions/recoveries, how many resumes are sent to each participant, and participant strategies of satisficing vs optimizing. Expanding on that last point, for positions that expect a lot of turnover the participant may use a satisficing rule and hire the first number of applicants that meet certain criteria, so there may be more call-backs overall and the timing of when resumes are sent may be important for detecting differential treatment. Other job openings, particularly those with more limited positions and longer tenure, may use an optimizing rule in order to get the best applicant possible, so there may be fewer call-backs and the quality of the resume will be important for finding differential treatment.

4.2.3 How Many Pieces of Correspondence to Send

The choice of how many resumes to send to a single participant at one time has tradeoffs. An obvious benefit of sending multiple resumes to a single firm is that it is an easy way to increase the number of total resumes sent. As with matched-pairs designs, this choice makes it easier to see how a single participant treats different types of resumes and can help to make a compelling argument for differential treatment that media reporters can easily understand. However, the choice to send multiple resumes to a firm comes with several potential drawbacks. One problem common to any within-subject design compared to between-subjects design is that inclusion of different treatments and controls can cause the participant to more directly compare these treatments to each other than he or she would if only viewing one treatment or control, thus decreasing levels of detected discrimination. These types of effects are seen in experiments generally (e.g. Charness et al. 2012; Tversky and Kahneman 1981) and there is some evidence of spillover effects of resumes within audit studies themselves (Phillips 2016). A related problem is that with more hypothetical resumes, the participant may change his or her priors about the underlying quality distribution and number of potential employees within the applicant pool. Thus any results from these studies will be externally valid to a different sample than reality. With a large enough number of resumes sent for a small number of interviews, there may also be mechanical effects—weak levels of discrimination will be magnified if, with a smaller number of applicants, equivalent resumes from both groups would receive an interview. Finally, there may be ethical concerns if the number of resumes is large enough to affect the hiring manager's practices; he or she may have trouble hiring if, for example, the opening has received a larger number of highly qualified applications than usual because of a large number of

hypothetical applications. There is not one right answer for how many resumes to send to an open position. The benefits and disadvantages will vary by job type. In general, the disadvantages will be lower with openings that receive a larger number of applications than those which receive a smaller number. For example, sending four resumes of varying quality to a low-level job during a recession for an opening that receives hundreds of resumes will probably still produce externally valid results and not harm the company, but sending four high quality resumes for a job that has a pool of maybe twenty qualified applicants (e.g. Horton forthcoming) can provide biased results and harm the company.

Another decision to make regarding the participant pool is whether to sample participants with or without replacement. For example, if an employer advertises a second time during the sampling period, will it receive multiple sets of resumes from the study? External validity concerns would suggest considering if an actual seeker would apply for the same job or company again. This answer may depend on the time between reposting, and if it is for the same job that has already rejected the applicant or for a different job in the same company. Sampling with replacement has some downsides, however. If the resumes sent are from a quality pool that is sufficiently different from real job applications to the firm, then the study itself may be changing the employers' beliefs about the applicant pool which may have spillovers to the results. Another design concern may make this decision mechanically—if the sets of resumes are similar but not identical across items, for example, they use the same names and contact information but the other resume items vary, then a second set of applications to the same firm will be testing the effect of seeing resumes for what seems to be the same applicant but with at least one set of qualifications forged. Again, this concern ties back to the original guideline to not test "unusual."

4.3 Crafting Correspondence

4.3.1 Choosing Correspondence Inputs

After selecting the participant pool, the next question to address is how to build correspondence inputs. In general, correspondence should be both realistic and externally valid to the pool tested. A common tactic in employment audit studies is to take inputs from real resumes gathered from online resume banks. These inputs are then either mixed and matched or modified slightly and used for a different employment pool so as to not negatively interfere with the job search of the applicants whose inputs were used. Care should be taken with this strategy; while it may be more externally valid than entirely fabricating inputs, it is still only externally valid for applications of the same quality or composition group from which the inputs came. In particular, the quality pools for resume banks may differ greatly. For example, resume audits from the early 2000s often used Americasjobbank.com (now careeronestop.org), which was a government-run job bank program (i.e. Bertrand and Mullainathan 2004; Lahey 2008). The resumes in this bank were often low quality, e.g., full of typographical errors. Resumes that remained in the bank for

longer periods of time tended to be of especially poor quality. More modern resume banks, for example, Indeed.com, seem to have higher quality examples on average. There is no guarantee that the composition of resumes on a resume bank site is equivalent to the composition of resumes that a posted advertisement will receive.

Quality of correspondence is additionally important for theoretical reasons. For example, with theories of variance-based statistical discrimination there is an interaction between quality of the resume and the treatment variable, with the dominant group preferred at higher quality levels, but the group for which there is less information preferred at lower quality levels. If the quality distribution of correspondence is small, the experiment may only be able to pick up a portion of this activity and may potentially give misleading results about the market as a whole. If the question being asked focuses on a specific quality segment of the market, the correspondence quality will be less of a problem because the pool is externally valid to the question being asked. An additional concern with quality levels is a mechanical one—if the quality of correspondence is too low, it may be difficult to get any positive responses from participants; treatment and control correspondence will have been treated the same, but that does not prove the lack of discrimination in the labor market and the results will not provide useful information on the impact of individual resume characteristics and their interactions.

As discussed in the first section, the choice of indicator that separates the treatment group(s) from the control groups is a key decision in the study design. Particularly, researchers should avoid correspondence that stands out for reasons unrelated to the study. Otherwise the external validity is reduced because the results show how participants treat unusual correspondence rather than showing how they treat the variable of interest.

It is important to be aware of correspondence trends. For example, styles change with regard to resumes and are not consistent across countries. Using recommendations for how to create a resume from 10 or 20 years ago may show that the applicant has not kept up with the times; in this case the results would only be externally valid for the group of applicants who submit old-fashioned resumes. Objective statements have fallen in and out of favor, various sections on the resume are given more or less weight, what type and how much previous experience to include varies, and so on. What is true at the time this volume is being written may be outdated in ten years. Prior to starting a study, determine what is "normal" for the study's specific area of interest. In the employment context, this can be done via viewing actual resumes submitted for a recent job opening, talking to HR representatives or hiring managers for positions similar to the type you are testing, and reading recent popular advice for job seekers.

4.3.2 Creating Correspondence

Once inputs have been gathered and the indicator has been chosen, those elements are combined to create the correspondence. Early studies based on matched-paired audits would often have a small number of correspondence templates, perhaps as many as eight, that they manually assigned names of different races or genders. This type of

study is only externally valid for the types of people similar to those that the template represents, making it impossible to get a full view or even a large view of the labor market. In addition, without variation within the templates, it is difficult if not impossible to get a full picture of who within the broad group is being discriminated against, how they are being discriminated against, and why they are being discriminated against.

Our previous paper (Lahey and Beasley 2009) addressed these concerns and argued that three common problems with audit experiments were surmountable through automated random generation of correspondence. First, with limited numbers of templates, all items except the variable of interest are correlated within each pair of templates, so the results can only predict the outcomes and interaction effects for specific bundles of characteristics rather than individual characteristics. For example, in an age discrimination study, if employers only discriminate against older workers without computer experience and all templates have computer experience, then not only would such a study not find evidence of differential treatment by age, but it would not be able to determine that lack of computer experience was a reason for age discrimination. Similarly, limited templates can group specific work experiences with specific education experiences, for example with "high quality" and "low quality" templates, so that it is not clear if effects are coming from the work experience or the education experience or a combination of the two. We term this problem "template bias". Sending out a large number of dissimilar resumes can isolate the predictive effects of individual characteristics and their interactions with group status. Second, experimenter bias is exacerbated when humans are responsible for manually generating correspondence or matching templates to jobs, because the human may subconsciously deviate from random assignment. Third, early in-person matched-pairs audits were limited in scale and scope by expense, which necessitated small sample audit analysis.

With automated, random generation of correspondence, the number of templates is no longer limited because each correspondence can have some probability to contain any given characteristic, robust pseudo-random number generators replace human action and thus avoid experimenter bias, and (given sufficient input material) generating large numbers of unique correspondence is quick and inexpensive. With enough responses, standard econometric techniques (OLS or Probit/Logit) can be used to test the impact of individual correspondence characteristics and their interactions with group differences on the outcome of interest. Additionally, with many templates or completely unique randomized correspondence, the researcher can allow the market to determine what the quality of a resume is rather than imposing one's own beliefs about what employers are looking for, something we discuss in the analysis section. At the same time, each additional variable may decrease the power of the study. In general, we are in favor of large audit studies that are powered for main pre-specified hypotheses but that also allow for tests of secondary hypotheses that the study may not have enough power to test.

A simple approach to generating correspondence is via "mail merge", a thirty-year-old method in which a form letter has blanks that get filled from a list of text inputs, e.g., names and addresses (Friedman et al. 2013). The resulting correspondence outputs are generally nearly identical because the majority of the text is

unchanged. While straightforward to use and supported by most word processors (current versions of Microsoft Word have a Mailings tab with a "Start Mail Merge" option), mail merge does no more than fill form letters by copying text from a list of inputs. The experimenter must take care in creating the list of text inputs to avoid experimenter bias, then create a dataset to link correspondence characteristics to outputs, then prepare different form letters (i.e., templates) if extensively different correspondence is desired. If different blanks in a form should relate (e.g., employment history is a function of bachelor's degree) then the experimenter must create the list of text inputs to contain that relationship. So while mail merge can fill a form letter with input text, the experimenter must manually generate the form letter and the inputs, and is saved only the effort of copy/pasting the latter into the former. Thereby mail merge solves the small sample problem because it assists in generating more correspondence quickly and easily, but it does not help with either limited templates or experimenter bias.

To help in the implementation of audit studies that surmount all three problems, we have developed a free open-source computer program named Resume Randomizer.[3] The program can create correspondence with a large number of experimenter-defined characteristics, and comes in two parts. The first part is an HTML-based user interface used to create templates. These templates can randomize inputs across the correspondence, including specifying the probability that an input will be included or the number of times an input will be included. For example, each correspondence may start with the same salutation, then have a random slot that selects between many unique first sentences for an objective statement, then have another random slot that has a twenty-five percent chance of outputting nothing and otherwise randomly chooses four different job history statements, and so on. Each input (i.e., characteristic) is simply text provided by the researcher; Resume Randomizer was initially envisioned as a Mad Libs-like game for quasi-randomly making unique resumes from lists of names, jobs, dates, etc. Thus, in creating a template, the researcher is making a file that contains: (1) a Mad Libs-like text containing blanks (or slots), (2) lists of text that can fill each blank, and (3) control or flow logic used to delineate repeating sections, nested random sections, and various other options affecting resumes created from the template. Thus, the template file created using the user-interface is a plain-text file that can be written manually, but we do not recommend that because the user-interface is more intuitive.[4]

[3] Available at http://www.nber.org/data/ (under "Other"), at https://github.com/beaslera/resumerandomizer, or from the authors by request.

[4] The simplest non-trivial template file might be:

```
24 gui version number
*constant* 1 1
*random* 1-1 2 *matchDifferent*
*leaf* 1-1-1
John
*end_leaf* 1-1-1
*leaf* 1-1-2
Jane
```

The second part of the program is an executable that uses that template file to generate multiple correspondence to be sent to the same participant. This part of the program allows for "matching" between correspondence so that either all the correspondence generated have the same characteristic for a given item, or so that none of them share characteristics for that item. The generated correspondence are plaintext, but various approaches can be used to add formatting, including generating the correspondence in TeX or HTML, or once the characteristics are chosen via Resume Randomizer then using mail merge to put those characteristics into Word documents (Oreopoulos 2011). Along with each correspondence, the second part of the program saves a "variable file" that, when combined with the input texts and template, contains all the information necessary to re-create the correspondence. This variable file can be imported into a statistical program, e.g., Stata, to analyze the impacts of characteristics.

With this program, researchers can avoid the three common problems with audit experiments, and generate correspondence sufficient for using standard econometric techniques to test the impact of individual correspondence characteristics and their interactions. First, template bias, in which all items except the variable of interest are correlated within each pair of templates, can be avoided by making each characteristic have some probability of being placed onto each resume or letter, so the impact of each characteristic (or group of characteristics) can be tested separately. Continuing the example above, if all resumes have some chance of containing computer experience independently of age, that will allow for separately testing the effects of age and computer experience. Second, the problem of experimenter bias can be mitigated because the software composes the correspondence randomly, so an un-biased template will lead to un-biased correspondence, in aggregate. Third, as with mail merge, this program substantially reduces the expense of generating additional correspondence, though the researcher must still provide sufficient input texts.

Since the initial release, we have revised the Resume Randomizer program for clarity, additional features, and ease of use. Random sections can now be configured to specify the exact percentage chance of choosing each potential result. Sections of the template can now be chosen based on the selection made in a previous random section, e.g., fraternity vs. sorority membership at the end of a resume can naturally depend upon a random gender choice at the start of the resume. Text can be saved into variables defined on-the-fly in the template, and then recalled from those variables later in the template, e.g., randomly choose the name at the top of the letter and save the corresponding initials for use later in the letter. To ease analysis, the

---

end_leaf 1-1-2
end_random 1-1 2
end_constant 1 1

which defines a single slot into which will go either "John" or "Jane" in the correspondence, and which appears in the user-interface as two text boxes that each contain one of the names plus drop-down boxes for various options. Example template files are distributed with the program, and the HTML user-interface has buttons that can load sixteen examples of templates, e.g., https://raw.githubusercontent.com/beaslera/resumerandomizer/master/example_cover_letter_template.rtf

executable now automatically generates a codebook that maps the variables saved in the variable file to the text that gets placed in the correspondence. To simplify assembly of the input text snippets, templates can now import text files that solely contain such text items. We will continue to incorporate useful features as we get feedback from users.

4.3.3 Matched Correspondence

An important choice is whether or not to use matched pairs in the audits. This study design essentially sends two resumes to the same firm that are identical except for the group characteristic of interest. Matched pairs were originally used for in-person audits because they dramatically increase power for small sample sizes. For studies that are necessarily small, matched pairs may still be the best design choice. However, there are drawbacks that come with matching pairs in audit studies. Using matched pairs is a within-subjects study design rather than a between-subjects design, which means that the same participant sees both the treatment(s) and the control (Charness et al. 2012). Even if participants do not realize that they are participating in an experiment, they are more likely to make a direct comparison between the treatment(s) and control which may change the effects of discrimination, most likely decreasing them by reducing implicit bias (e.g. Olian et al. 1988). A more ethical concern is that sending a participant matched sets of correspondence may be more likely to distort the participant's view of the labor market if they think that a specific type of hypothetical applicant is more heavily represented in the labor market pool than is actually true. Unmatched sets send a less focused signal and may be less likely to harm a participant's overall view of the market.

It is possible that the matched-pairs design may be better able to test for differences in situations in which some element of what is being tested can affect the general equilibrium applicant pool. For example, a hypothetical resume audit could find that firms that advertise as being Affirmative Action/Equal Employment Opportunity (AA/EEO) are less likely to interview hypothetical black workers than firms that do not advertise as being AA/EEO. These AA/EEO firms may still be less discriminatory if general equilibrium effects of having AA/EEO advertising mean that more black applicants are applying to the firm (Kang et al. 2016).[5] From the standpoint of a single minority job seeker the reason for not getting called for an interview is less relevant, but from the standpoint of the labor market we would not be able to make the claim that firms with AA/EEOC are more discriminatory. The black/white comparison within firms that advertise AA/EEOC is important, and matched pairs may be the best way of getting enough power to test for these effects. Chapter 6 by Mike Vuolo and colleagues (2018) will discuss concerns about matched pair audits in more detail.

[5] See Pager and Pedulla (2015) for more information on how perceived discrimination affects job application behavior.

4.4 Sample Size

An important part of the experimental design phase is figuring out the minimum sample size needed to find significant results for a reasonable effect size given a set power. Determining necessary sample size via power analysis requires information on effect size, desired significance level and desired power. Ideally the effect size will come from a pilot study. However, it is possible to get suggested effect sizes for field experiments from previously completed laboratory work or from related field studies. Psychologists have long been interested in many of the questions that other social scientists are just now testing in the field. In the absence of any prior related work, experimenters can use the default effect sizes of small, medium, or large based on beliefs about the size of the effect or based on the practical impact of an effect that is small, medium, or large. That is, if it is believed that a small effect size would be unimportant for the population in question, then it may be sufficient to gather a sample that could only capture a medium size effect. In general, one can choose standard levels for significance (0.05) and power (0.8), although these heuristics may be overly simplistic (Cohen 1977, 1992).

Power analysis has become easier in recent years given the availability of the program G*Power.[6] Current versions of G*Power can even determine sample size for matched pair studies. While G*Power is remarkable in many respects, as of this writing, it still lacks in two areas important to researchers planning audit studies. First, G*Power does not take into account clustering. If the study design includes sending multiple pieces of correspondence to the same participant, G*Power does not account for how power is affected by the loss in variation due to multiple samples per participant. To take into account the additional sample size needed because of the clustered design, sample size calculations from multi-level modeling for two levels can be used.

$$Sample\ Size_{final} = Sample\ Size_{G*Power}{}^{*}\left(1+\left(number\ of\ items\ per\ cluster-1\right)^{*}\ ICC\right)$$

The desired sample size, $Sample\ Size_{final}$, is calculated by multiplying the sample size (given by G*Power) that does not take into account clustering by a factor that takes into account both the number of items per cluster (ex. the number of resumes being sent to a firm) and the average inter-correlation between clusters (ICC). With a pilot study, the ICC can be determined using the *xtmixed* or *mixed* commands in Stata to determine standard deviations and applying the following formula:

$$ICC = \frac{\sigma^2_{cons}}{\sigma^2_{cons} + \sigma^2_{residual}}$$

[6] Stata's currently supported sample size calculator is *power*, but as of this writing has limited options compared to G*Power and thus is only recommended for simple designs, although its *nratio* option is useful for unbalanced designs.

G*Power is available for free from http://www.gpower.hhu.de/en.html and is available for both Mac and Windows.

where σ^2_{cons} is the standard deviation of the constant and $\sigma^2_{residual}$ is the standard deviation of the residual. In the absence of a pilot study, default ICC range from 0.10 to 0.30 (Gulliford et al. 1999; Maas and Hox 2005). The Stata .ado file, *clustersampsi,* may also be used to find appropriate sample sizes for clustered designs.[7]

A second drawback of G*Power is that how to test power for interactive effects is unclear—the "Linear Regression Model" options do not provide information on power to test the significance of an interacted coefficient, but test the effect of the interaction on the regression's R^2. Instead, G*Power's ANOVA framework can provide sample size analysis for interactive effects.

4.5 Datamining Concerns: Pre-registration and Mid-Experiment Analysis

Pre-registering experimental plans has become more de rigueur in recent years. Grant proposals, which are often necessary to pay for experimental work, function in a similar way to pre-registration because they force researchers to outline their hypotheses and analysis plans a priori. Olken (2015) does an excellent job explaining the pros and cons of pre-analysis plans and discussing the elements they should include. Such plans remove problems of data-mining and remove the need for most robustness checks, but also limit exploration and are difficult to implement for tests of more complicated theories. Our general belief is that there are benefits to plan pre-registration but that one should not be dissuaded from doing exploratory secondary analysis in conjunction with or after completing the primary analysis. Correspondence review studies are large undertakings and are often our first glimpse at the hiring sides of various markets. One correspondence review cannot provide the definitive answer to any economic question and there is a place for exploratory work that informs future pre-planned studies.

How often to analyze the data while the study is being run is a related concern that has trade-offs with data-mining. In the ideal world, researchers would design the study, do a small pilot study to make sure everything was in working order and to get information for sample size calculations, and then they would run the experiment without looking at the results until it had completed. In the real world, however, mid-stream checks are important to make sure that the experiment is still running smoothly and is free from human error or unforeseen external shocks. For example, researchers may want to check to see that resume inputs from the sent resumes are balanced in the way that researchers expect them to be and that response rates are not dramatically lower than expected because of mechanical issues. Any dramatic changes in results over time may also warrant exploration to make sure that they are based on changes in the hiring environment and not, for example, a major email provider deciding to send all the experiment's emails directly to their

[7] Thanks to R. Alan Seals for this suggestion. He also notes that there is room for a methodology paper on best practices for finding sample sizes in audit studies.

spam filter or a new research assistant accidentally sending resumes to a holding folder rather than out to participants. While it may be tempting to use mid-stream checks to make major changes in the experiment based on results, doing so comes at the expense of data-mining concerns.

4.6 Technical Data Concerns

4.6.1 Sending Correspondence

How resumes are submitted has changed over the past few decades. In early studies it was standard to mail applications or to submit them by hand. Studies from 15 years ago generally faxed resumes to prospective employers. Today, emailed and online applications are much more common. One new program to facilitate mass emailing of correspondence is an automation program by Chehras (2017). Her code will match correspondence to openings based on location and date, generate an email, attach the correspondence, and send the email including delays as desired. Crabtree's (2018) Chap. 5 in this volume also discusses email audit studies.

4.6.2 Matching Responses to Correspondence

Once the experiment has been planned, the participants chosen, the correspondence generated, and the correspondence sent to the participants, the experimenter will still need to match the participants' responses to the characteristics of the correspondence. In a laboratory experiment, this matching can be automated because the experimenter can directly collect the responses from the participant. However, when doing a field experiment, the response can be at some remove from the stimulus. Virtual voice-boxes, P.O. boxes, and email addresses are common ways of collecting responses and should be chosen with external validity concerns in mind.[8] With generous resources or with a limited number of templates, each stimulus would have its own unique phone number and email address and thus the responses would be directly connected to the correspondence. With more limited resources, it is possible to bin responses based on the main variable combinations of interest, for example, a

[8] Note that researchers using their own domain, such as those from hostgator, can quickly create hundreds of email addresses all with the same passwords and settings, facilitating exact matches when responses come via email. Voicemail matches are more difficult. Neumark et al. (2016) populated voicemail bins such that each voicemail only had one version of each first name and last name used, which helped with matching. "So if a bin got a call, and they said, 'Hi Jennifer, we'd like to interview you,' then we knew the exact applicant since there was only one Jennifer in that bin," (personal communication, Patrick Button, October 20, 2016). R. Alan Seals (personal communication, November 13, 2016) recommends using Google Voice to transcribe phone messages from employers for easy text analysis.

researcher looking at the effects of race for different 10-year age intervals by gender could have a separate phone number or email address for each age interval*race*gender combination. A drawback of binning rather than doing exact matching is that because correspondence is not directly matched to its response it is difficult to explore the effect of any variables that were not used to create the bins. Without making separate bins by characteristics, it is necessary to match the resumes to the responses using clues from the responses. However, this is costly in terms of person-hours and is not always possible when, for example, firms call back from a number unrelated to the one in the advertisement and do not provide any other identification. Even with binning, it may be difficult to determine when the same company is calling back multiple times in response to the same application.[9]

4.6.3 Data Storage

If possible, keep a copy of everything pertaining to the experiment. In these days of inexpensive storage, it is better to have unused data than to need something and realize it was not preserved and is no longer available. As an example of data size, three thousand resumes, including all the data plus images of the resumes, can take under three hundred megabytes. Each resume's pertinent features must be saved for use in the analysis, commonly stored as variables and a codebook mapping those variables to the resume text. Saving a copy of exactly what is sent to the participant is also a good idea to be able to answer any questions that may arise about what the participants actually received. If the IRB permits, saving prompts such as original job advertisements may allow for later in-depth analysis that uses text from these prompts.[10]

Additionally, save the template or process used to generate the submission material. For the Resume Randomizer program, these files consist of scripting commands that detail which inputs should be chosen with specific probabilities and matching constraints. By saving this information, if there are any questions about how the resumes were supposed to be generated, those can be quickly answered. As an example, after the study is run there could be a question about what probabilities were intended during resume creation for the years of high school graduation. While the variables and codebook can detail what resumes were actually generated, the template is necessary to know the process that generated them. Furthermore, the template can be used as a starting point for future experiments.

The final recommendation regarding data storage is to store an off-site backup of everything in case of hard drive failure, fire, or natural disaster. For those who do not

[9] In order to reduce the burden on companies, it is common for experimenters to respond to firms that the employee has taken another job after being contacted for an interview during this step.

[10] R. Alan Seals (personal communication, November 13, 2016) notes that if you save prompts electronically as webpages, it is important that workers all use the same web browser to facilitate text scraping.

have secure online back-ups available from their place of work, Amazon currently sells unlimited storage via Amazon Drive for sixty dollars per year, and a variety of other companies offer similar storage services (e.g., Google Drive, DropBox, iCloud, OneDrive). Sharing data with other researchers after publication at a site such as ICPSR will also protect from data loss. In doing so, be mindful of appropriate data-protection/anonymization protocols and any restrictions imposed by IRB or any governing body for the data.

4.6.4 If You Need to Change the Resumes Mid-Experiment

Sometimes correspondence will need to be changed mid-experiment. For example, summary statistics or initial analysis can indicate that a mistake was made in the template(s). Inaccurate calculations of numbers/ages/years, using an outdated version of the template, or completely omitting a section of the resume are all examples of unintentional actions/inactions that might substantially reduce external validity. Alternatively, even after a careful pilot study, unexpected events or findings after the experiment has started can encourage researchers to make modifications to the study. This chapter encourages (and facilitates) mindful preparation, but unforeseen and unavoidable occurrences happen and can lead to the decision to make a correspondence revision mid-experiment despite the reduction in power that comes from dividing the samples.

Mid-experiment revision leads to data storage and data connection challenges. The first challenge is keeping track of the data from resume inputs. For simple designs that use a limited number of templates matched by hand or via mail merge, it is sufficient to mark the resumes before and after the change. Researchers using our program (as of this writing) to create correspondence will end up with two separate datasets, one from before the change and one from after the change. Depending on the change that has been made, the variable names or values may no longer map to each other. Researchers should then post process these two datasets separately before combining in order to match the correct variables together. The second challenge is that responses to the new correspondence need to be identifiable from responses to the old correspondence. If using bins for response collection, that separation may require new email addresses or phone numbers. For more complicated matching procedures it may be sufficient just to keep track of the date at which the change was made. Finally, it is important to keep a clear record of any changes made and when they were made. On hard drives, it is helpful to keep the new data (template, codebook, variable files, etc.) in a separate folder from the pre-revision data to avoid any confusion or lost data. Obviously, a researcher should also clear the changes in correspondence with their IRB if required to do so.

4.7 Analysis Concerns

The choice of dependent variable will vary by study. In resume audits, the choice between call-back (when the company sends any non-negative response) versus interview (when the company specifically requests an interview) is a common one. There does not seem to be a consensus on which numbers to present, and in our opinion researchers should present both for comparability across studies. Researchers using other types of correspondence audits should use what is most common in their specific literature unless there is a strong theoretical reason not to.

When multiple stimuli are sent to the same participant (ex. multiple resumes are sent to the same want-ad), it is important to account for between observation (intra-class) correlation. In that case, one should cluster on participant in a regression framework (Lahey and Beasley 2009). For many cases, simply clustering on participant will be sufficient, however some studies may require more complicated methods of correcting standard errors. Clustering can be nested, but if non-nested clusters exist (e.g., different participants sampled over time), traditional cluster inference can only handle one of the dimensions (Cameron and Miller 2015). Alternatively, random effects modeling is commonly used in the metrics of panel data and can be used if the group coefficients are assumed to be uncorrelated with observed group covariates. Both random effects modeling and the more general multilevel modeling (MLM, also called "mixed" models), can handle multiple levels of correlation (e.g., state and participant). A detailed discussion of these different ways of dealing with clustered data is beyond the scope of this chapter, but a good place for interested readers to start is the UCLA Institute for Digital Research and Education webpage on analyzing correlated data (http://www.ats.ucla.edu/stat/stata/library/cpsu.htm).

The related question of when to use participant fixed effects is non-trivial. When sending multiple resumes to a firm, it is tempting to use firm or job opening fixed effects to control for firm characteristics, all items on matched resumes that are matched, and even the point in the business cycle at which the resumes were sent. However, using firm fixed effects when the dependent variable is binary and the researcher is using logit or probit analysis leads to a mis-estimation of the level of differential treatment because it drops all instances where the stimuli were treated the same by the firm, leading to the standard Heckman critique (Heckman 1998). That is, the measure of differential treatment is solely determined by instances in which the firm treats candidates differently, but ignores all instances when firms treat candidates the same.

A second Heckman critique about audit discrimination studies is that the magnitude of market discrimination that these studies find has no real-world meaning because the treatment and control are equivalent by design except for the characteristic of interest (Heckman 1998). Thus discrimination magnitudes can only be compared across audit studies but have no real world relevance other than their sign and significance. Neumark (2012) provides a clever method of translating the results from a discrimination audit study into meaningful numbers by essentially anchoring

the audit results onto population characteristics while Lanning (2013) proposes a method to translate audit-pair findings into wage differentials.

Existence and magnitude of discrimination are not the only outcome of interest even in a discrimination correspondence study. A primary benefit of the larger sizes and better technology with modern correspondence studies is that they are no longer limited to addressing the question, "Is there differential treatment?" and now can start to answer questions of, "Why is there differential treatment?" and "Which subgroups are most affected?" Pedulla (2018) in Chap. 9 of this volume goes more into detail about these important theoretical questions. A simple interaction with main effects can be used to test both of these types of questions. One caveat is that interactive effects require larger sample sizes to find significance at a reasonable power, and researchers should be cognizant of these requirements.

One specific avenue of interest may be testing differential effects by the "quality" of the correspondence. Rather than having the researcher decide what items constitute high quality vs. low quality, it is best to let the market decide what items they prefer. A simple way to get predicted quality is to regress the outcome measure on all items that vary absent the ones that you care about, for example, regress callback outcomes on all resume items except name (which indicates race/gender), or on all resume items except high school graduation date (which indicates age). Then the predicted Y would be the quality measure absent the variable of interest.

4.8 Beyond Standard Audit Studies

Although we have motivated much of this chapter with resume audits, the correspondence review technology does not need to be limited to employment audits. This technology can be expanded to many types of laboratory or natural field experiments (Harrison and List 2004). For example, there is no reason hypothetical correspondence cannot be used with subject pools like Amazon's Mechanical Turk (see Porter et al. 2017 for more discussion of Mechanical Turk) or used in conjunction with a natural experiment as in Agan and Starr (2016). The technology can be combined in a laboratory setting with surveys, eye-tracking (ex. Lahey and Oxley 2016), IAT tests (ex. Rooth 2010), other types of laboratory experiments, and so on, to get a richer understanding of what motivates people's choices. Much of this technology has historically been used to explore discrimination in markets, but it does not need to be limited to employment, mortgage markets, or purchasing (Bertrand and Duflo 2016; Neumark 2016). Potential future avenues could include experiments looking at soliciting donations, responses to consumer complaints or political concerns, or the effects of advertising. The use of these methods is only limited by ethical concerns and the researcher's imagination.

4.9 Technical Checklist

- Determine an externally valid (unobtrusive) signal for the treatment(s)
- Talk with practitioners and explore current practices in your market
- Decide on a participant pool
- Choose how to gather representative inputs
- Plan response collection method (e.g., email addresses)
- Review design choices with respect to the expected external validity
- Get IRB approval for pilot and run a pilot study (optional)
- Estimate necessary sample size from pilot, previous research, or default estimates
- Decide on length of time to field experiment
- Decide on data storage including off-site back-ups and regular back-up schedule while experiment is running
- Register experiment (optional)
- Get IRB approval
- Generate correspondence
- Submit correspondence
- Collect participant responses
- Match responses to correspondence
- Respond to employers (optional)
- Mid-experiment analysis, revision, and IRB changes (optional)
- Do primary data analysis as specified in registration, grant proposal, or other initial plan
- Do exploratory secondary data analysis

References

Agan, A. Y., & Starr, S. B. (2016). *Ban the box, criminal records, and statistical discrimination: A field experiment* (U of Michigan Law & Econ Research Paper No. 16–012). Available at SSRN: https://ssrn.com/abstract=2795795

Baert, S., & Dieter, V. (2014). *Unemployment or overeducation: Which is a worse signal to employers?* (No. 8312). Institute for the Study of Labor (IZA).

Barron, J. M., Bishop, J., & Dunkelberg, W. C. (1985). Employer search: The interviewing and hiring of new employees. *The Review of Economics and Statistics, 67*(1), 43–52.

Bendick Jr, M., Brown, L. E., & Wall, K. (1999). No foot in the door: An experimental study of employment discrimination against older workers. *Journal of Aging & Social Policy, 10*(4), 5–23.

Bertrand, M., & Duflo, E. (2016). *Field experiments on discrimination via National Bureau of Economic Research*. http://www.nber.org/papers/w22014. Accessed 13 Oct 2016.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013.

Brewer, M. (2000). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372.

Carbonaro, W., & Schwarz, J. (2018). Opportunities and challenges in designing and conducting a labor market resume study. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8.

Chehras, N. (2017). Automating correspondence study applications with python and SQL: Guide and code. *Mimeo.*

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Crabtree, C. (2018). An introduction to conducting email audit studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Darolia, R., Koedel, C., Martorell, P., et al. (2015). Do employers prefer workers who attend for-profit colleges? Evidence from a field experiment. *Journal of Policy Analysis and Management, 34*(4), 891–903.

Deming, D. J., Yuchtman, N., Abulafi, A., et al. (2016). The value of postsecondary credentials in the labor market: An experimental study. *The American Economic Review, 106*(3), 778–806.

Friedman, S., Reynolds, A., & Scovill, S. (2013). *An estimate of housing discrimination against same-sex couples*. Available via the US Department of Housing and Urban Development: http://big.assets.huffingtonpost.com/hud.pdf

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.

Gaddis, S. M. (2017a). How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*, 469–489.

Gaddis, S. M. (2017b). Racial/ethnic perceptions from Hispanic names: Selecting names to test for discrimination. *Socius, 3*, 1–11.

Gulliford, M. C., Obioha, U. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology, 149*(9), 876–883.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055.

Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives, 12*(2), 101–116.

Holzer, H. J. (1996). *What employers want: Job prospects for less-educated workers*. New York: Russell Sage Foundation.

Horton, J. J. (forthcoming). The effects of algorithmic labor market recommendations: Evidence from a field experiment. *Journal of Labor Economics.*

Howden, D. (2016). *Interviews per hire: Recruiting KPIs*. Available via Workable. https://resources.workable.com/blog/interviews-per-hire-recruiting-metrics/. Accessed 13 Oct 2016.

Jobs Opening and Labor Turnover Survey (JOLTS). (2016). *Bureau of labor statistics*. http://www.bls.gov/jlt/home.htm. Accessed 13 Oct 2016.

Kang, S. K., Decelles, K. A., Tilcsik, A., et al. (2016). Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly, 61*(3), 469–502.

Lahey, J. N. (2008). Age, women, and hiring an experimental study. *Journal of Human Resources, 43*(1), 30–56.

Lahey, J. N., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior & Organization, 70*(3), 508–514.

Lahey, J. N., & Oxley, D. (2016). *Discrimination at the intersection of age, race, and gender: Evidence from a lab-in-the-field experiment*. Working Paper.

Lanning, J. A. (2013). Opportunities denied, wages diminished: Using search theory to translate audit-pair study findings into wage differentials. *BE Journal of Economic Analysis and Policy, 13*(2), 921–958.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

Maurer, R. (2016). *More employers moving to fewer interviews*. Available via Society for Human Resources Management. Accessed 13 Oct 2016.

Moynihan, L. M., Roehling, M. V., LePine, M. A., et al. (2003). A longitudinal study of the relationships among job search self-efficacy, job interviews, and employment outcomes. *Journal of Business and Psychology, 18*(2), 207–233.

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *Journal of Human Resources, 47*(4), 1128–1157.

Neumark, D. (2016). *Experimental research on labor market discrimination*. NBER working paper series. http://www.nber.org/papers/w22022. Accessed 17 Oct 2016.

Neumark, D., Burn, I., & Button, P. (2016). Evidence from lab and field experiments on discrimination: Experimental age discrimination evidence and the Heckman critique. *The American Economic Review, 106*(5), 303–308.

Olian, J. D., Schwab, D. P., & Haberfeld, Y. (1988). The impact of applicant gender compared to qualifications on hiring recommendations: A meta-analysis of experimental studies. *Organizational Behavior and Human Decision Processes, 41*(2), 180–195.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives, 29*(3), 61–80.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy, 3*(4), 148–171.

Pager, D., & Pedulla, D. S. (2015). Race, self-selection, and the job search process. *American Journal of Sociology, 120*(4), 1005–1054.

Pedulla, D. S. (2018). Emerging frontiers in audit study research: mechanisms, variation, and representativeness. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Phillips, C. (2016). *Do comparisons of fictional applicants measure discrimination when search externalities are present? Evidence from existing experiments*. Working Paper.

Porter, N. D., Verdery, A. M., & Gaddis, S. M. (2017). *Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities*. Available at SSRN: https://ssrn.com/abstract=2844155

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523–534.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont: Wadsworth Cengage Learning.

Stock, J. H., & Watson, M. W. (2011). *Introduction to econometrics* (3rd ed.). Boston: Addison-Wesley.

Trochim, W. M. K., & Donnelly, J. P. (2006). *Research methods knowledge base*. Mason: Atomic Dog/Cengage Learning.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Vuolo, M., Uggen, C., & Lageson, S. (2018). To match or not to match? Statistical and substantive considerations in audit design and analysis. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Chapter 5
An Introduction to Conducting Email Audit Studies

Charles Crabtree

Abstract This chapter offers the first general introduction to conducting email audit studies. It provides an overview of the steps involved from experimental design to empirical analysis. It then offers detailed recommendations about email address collection, email delivery, and email analysis, which are usually the three most challenging points of an audit study. The focus here is on providing a set of primarily technical recommendations to researchers who might want to conduct an email audit study. The chapter concludes by suggesting several ways that email audit studies can be adapted to investigate a broader range of social phenomena.

Keywords Audit studies · Correspondence studies · Experiment

5.1 Introduction

What is an audit study? As other chapters in this volume note (Gaddis 2018), an audit study (or correspondence study) is one way of assessing hard-to-observe phenomena, such as discrimination (Heckman 1998).[1] The general structure of an audit study is very simple. To begin with, researchers create some set of identities. The initial identities share the same characteristics. Scholars then randomize one or more attributes of these identities, such as race or gender. Next they use the identities to accomplish some task, like applying for jobs, renting housing, or contacting legislators. These tasks can done via phone, mail, and email. Finally, scholars compare how individuals — such as prospective employers, landlords, or legislators — respond to the putative identities. Any difference in treatment across the randomized

[1] See Gaddis (2018) for a history of audit studies and an overview of the approach.

I thank Volha Chykina for her helpful comments. I particularly thank Holger L. Kern for teaching me about audit studies and providing me some of the code used to conduct email audit studies.

C. Crabtree (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: ccrabtr@umich.edu

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_5

attributes is interpreted as evidence of some latent bias. For example, if landlords respond to inquiries from Blacks less frequently than inquiries from Whites, then scholars would infer that landlords are biased again Blacks. Scholars have used audit studies to observe biases in nearly every facet of common life — in political interactions (Butler 2014; Broockman 2013; Butler and Broockman 2011; Grose 2014; Costa 2017), in housing transactions (Gaddis and Ghoshal 2015; Turner et al. 2002; Hogan and Berry 2011; Oh and Yinger 2015), in economic exchanges (Riach and Rich 2002), in employment decisions (Neumark et al. 1995; Bertrand and Mullainathan 2004), and in many other spheres (Pager and Shepherd 2008). Taken together, the results from these studies have considerably improved our collective understanding of discrimination.

The important point for this chapter is that an increasing number of these audit studies are being conducted over email.[2] There are several reasons for this. One reason is that email is an extremely common means of communication; approximately 2.6 billion people sent over 205 billion messages in 2011 (Radicati and Hoang 2011). Email can be used to accomplish virtually any communication related task — from exchanging documents, to sharing personal news, to organizing collective actions, to conducting business transactions, or even to requesting assistance from public officials. The dominance of email as a mode of communication is indicated by the fact that workers report spending up to 50 percent of their day reading, writing, and managing emails (Stocksdale 2013). This widespread use of email helps researchers because it provides them with opportunities to engage in many different types of interactions and thus potentially observe discrimination (or other phenomena) across many contexts. A second reason relates to external validity. As an increasing number of interactions occur over email, researchers would limit the generalizability of their findings if they only conducted audit studies through other media.

A third reason why the number of email audit studies is increasing is because they are relatively inexpensive to implement. There are costs to conducting audit studies through other means, such as the mail, that simply do not apply to email studies. For instance, in the case of mail, these costs might include stamps, post office boxes, or enumerators in different locations. In contrast, anyone with an Internet connection can send and receive emails for free. This means that researchers with limited resources — such as graduate students and junior faculty — might find email a particularly attractive means of conducting their correspondence studies.

Despite these advantages, email audit studies are perhaps underused. Certainly, many audit studies have been conducted over email since electronic messaging became widely available. This number could be even higher, though, as every published audit study suggests (implicitly or explicitly) a large number of possible extensions and adaptations.

[2] Some recent examples of this include Gaddis (2015); Gaddis and Ghoshal (2015); Sharman (2010); Radicati and Hoang (2011); Oh and Yinger (2015); Milkman et al. (2012, 2015), Lahey and Beasley (2009); Hogan and Berry (2011); Giulietti et al. (2015), Findley et al. (2015), Bushman and Bonacci (2004), Butler (2014), Ahmed et al. (2012, 2013), Baert (2016), and Baert et al. (2016a, b).

One reason why email audit studies might not be used more is that they are often difficult to implement. This is particularly true for scholars who are inexperienced with conducting audit studies in any medium. The issue here is that there are no general introductions to audit studies. Another reason why email audit studies might be underused is because scholars might think that they can only use them to examine a narrow range of social phenomena. While the vast majority of email audit studies have focused on unearthing evidence of discrimination, this general form of study can be easily adapted to examine a wider range of social phenomena.

In this chapter, I address both of these issues with the goal of increasing email audit study use.[3] The first section of the chapter attempts to reduce the complexity of email audit studies by providing a comprehensive guide to implementing one. This guide describes the steps involved in conducting an audit study. It also offers detailed recommendations about how researchers should collect, send, and code emails, since these are perhaps the most intimidating steps to inexperienced scholars. The primary focus of this section is on describing computerized, time-saving solutions to common issues. The R code used to address these issues is available online at charlescrabtree.com/email_audit and at auditstudies.com.[4] The second section of the chapter offers several suggestions about how scholars can adapt audit studies to investigate a broader range of social phenomena. It provides examples of non-traditional audit studies and discusses how those designs might be modified to answer other theoretical questions. This deconstruction of prior research might be helpful to scholars who are interested in audit studies but think that they cannot be used in their research.

5.2 Guide to Implementation

How can a researcher conduct an email audit study? This section addresses that question by providing an overview of the implementation process. Before discussing individual steps in detail, we first provide a general outline of the stages involved in a typical email audit study. These eight stages are listed in Fig. 5.1. They include

[3] I acknowledge that there are instances in which researchers cannot or should not implement an audit study over email. Perhaps the biggest reason for this is it might be impossible to collect email addresses for some populations. For instance, it would be very difficult to get email address information for a random sample of Americans. Similarly, one can imagine international contexts, such as many emerging market economies, where it might even be difficult to gather email addresses for public figures, such as government members. In addition to this concern, it is also probably true that some interventions are less plausible over email than through the regular mail or via phone. To the extent that researchers want to maximize the ecological validity of their interventions, they might want to conduct them via alternative means. Yet, despite these limitations, I still think that there are substantial opportunities for conducting additional email audit studies. These opportunities will continue to increase so long as email remains one of the most widely used means of communication.

[4] While I focus on using R to address some implementation issues, researchers should be able to accomplish similar tasks in Stata or using other programming languages, such as Python.

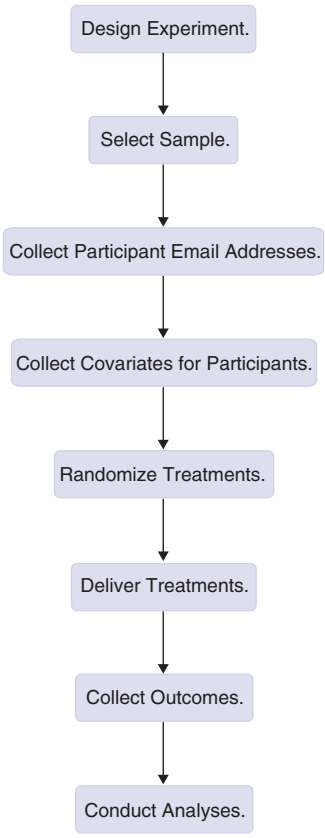

Fig. 5.1 The eight stages of a typical email audit study

(1) experimental design, (2) sample selection, (3) email address collection, (4) covariate collection, (5) treatment randomization, (6) treatment (i.e. email) delivery, (7) outcome collection, and (8) analysis.

One additional stage not discussed here is getting institutional approval, typically provided by an institutional review board (IRB), for conducting the intended study. Since the requirements of these boards vary considerably across institutions (Driscoll 2015), it is difficult to provide useful, general recommendations about how to manage their potential concerns. Interested readers should consult Hauck (2008), Fujii (2012), and Yanow and Schwartz-Shea (2008) for overviews of potential IRB-related problems and solutions. More generally, researchers should carefully consider the ethics of their experimental interventions. Desposato (2015) and Riach and Rich (2004) provide great introductions to these issues, and Gaddis (2018) offers several suggestions regarding best practices.

5.2.1 Experimental Design and Sample Selection

While experimental design and sample selection are extremely important, I do not discuss them here. Many excellent texts deal with issues related to design and sampling (e.g., (Gerber and Green 2012; Lohr 2009). I refer interested readers to them.[5]

One important issue related to sampling stands out — power calculations. These calculations are used to determine whether experiments are sufficiently powered to detect treatment effects. In other words, they help scholars determine if they have included a sufficient number of participants. The Evidence in Governance and Politics Group provides a useful guide (goo.gl/HXOK5Q) and a couple calculators (goo.gl/CJ8zox, goo.gl/0ucE9G) that researchers can use to think through their potential statistical power concerns.

Regardless of what researchers decide regarding experimental design and sample selection, they should consider pre-registering these choices, along with their theoretical expectations and analytic strategy (Olken 2015; Franco et al. 2014).[6] There are many possible reasons to write a pre-analysis plan.[7] If scholars pre-register their research designs, they might think more clearly about their theoretical expectations and the extent to which their proposed design might satisfactorily test them. Pre-registration should also lead to fewer questionable research practices, such as analyzing the data in whatever way leads to statistically significant results (i.e. 'p-hacking') or hypothesizing after results are known (i.e. 'HARKing'). This is because it forces researchers to commit to analyzing and discussing the results as discussed in the pre-analysis plan (Olken 2015). Finally, researchers might want to pre-register their designs because journals in some fields, such as political science and psychology, are increasingly encouraging this practice. Pre-analysis plans can be posted on sites like the AEA RCT Registry, the Evidence in Governance and Politics site, or on personal academic webpages.

5.2.2 Email Address Collection

Once a researcher has designed an experiment and selected a sample, they need to collect email addresses for each participant in their sample. This is typically one of the most difficult and time-consuming steps in conducting an email audit study. One of the things that make this so difficult is that researchers often want to recruit a large number of participants. This could be because they want to maximize statistical

[5] For ease of exposition, I assume that researchers are implementing a between-subjects design. The general process described in this chapter can be easily adapted to accommodate a within-subjects design. The only potential difficulty in doing this would be in modifying the email delivery script available in the online appendix. I have addressed this issue by modifying the code to deal with both types of design.

[6] Lin and Green (2015) provide excellent guidance on some of these decisions.

[7] Coffman and Niederle (2015) discusses some of the limitations of pre-analysis plans.

power or because they want to increase the external validity of their findings. Regardless of the reason, gathering contact information and other details for large samples can be intimidating. I briefly discuss here some of the ways that researchers can efficiently collect contact information for their sample. Thankfully, this task is now perhaps easier than ever before. In many cases, researchers can find participants' emails online, either individually or together as part of a mailing list. This is particularly true in the case of political figures. Sites like everypolitician.org and sunlightfoundation.com provide data for elected officials. Lists of unelected officials emails are often available from offices in Washington, D.C. or at state capitals.

Even when the information has not already been previously compiled by others, researchers still have many tools at their disposal that can reduce the time they would spend on data collection. One quick way to collect contact information is by scraping it from websites, such as job boards, or state agency employee listings. Building a web scraper used to be something that only a well-trained programmer could manage, but the diffusion of programming tutorials and the ready availability of example code at sites like github.com or stackexchange.com, have made it so that even individuals inexperienced with programming can adapt existing scrapers to their own purposes.

Some sites present problems to basic scrapers, though, such as login screens or paywalls. In these cases, researchers have two options. If they have research funds, they might consider paying a programming freelancer to create a custom scraper for them. Sites like elance.com and guru.com can help researchers find qualified help. Since building a scraper is a rather basic programming task, the job would not cost much. If researchers, however, cannot (or will not) pay for a freelance programmer to build a scraper, then they can explore what-you-see-is-what-you-get solutions, such as the excellent Web Scraper extension for Chrome.

After collecting emails, researchers should drop observations with obviously invalid email addresses. This includes emails that do not contain an '@' symbol, emails that contain spaces, and emails that are actually website addresses, among others. One reason to drop bad email addresses before implementing the experiment is to reduce the number of invalid email notifications received post-implementation. Scholars should not worry too much about catching every invalid address, though. Since treatment is randomized, they should be able to drop observations that contain bad contact information without biasing inferences.

5.2.3 Covariate Collection

Researchers might gather covariates on their participants either prior to or alongside email addresses. There are two general reasons to collect covariates related to their sample. One is to examine treatment effect heterogeneity. This is the "degree to which different treatments have differential causal effects on each unit" (Imai et al. 2013, 443). The idea here is to use pre-treatment covariates to determine the effect of treatments on different subpopulations. Another reason to collect covariates is to

include them in the randomization scheme, such as through block randomization (described below) (Suresh 2011). In many cases, scholars can use the same techniques to collect covariates as they do to collect email addresses.

5.2.4 Treatment Randomization

After collecting covariates, researchers should then decide how they intend to randomize treatment. There are many ways that you can do this. One approach would be to just use a random number generator. A more sophisticated approach would be to assign treatments within blocks. This is done by dividing subjects into homogeneous blocks and then assigning treatments within those blocks. The goal here is to increase efficiency by decreasing variability between units. When randomizing this way, I typically use the R package blockTools (Moore and Schnakenberg 2012). The choices that researchers face at this step are not unique to email audit studies, though, so I do not discuss them at length here. Gerber and Green (2012) offer a particularly good guide to the pros and cons of various randomization schemes.

5.2.5 Email Delivery

After scholars randomize treatment assignment, they need to assign those treatments to participants. Since this chapter focuses on email audit studies, I assume that treatments are being delivered via email. In order to assign treatment then, researchers need to email study participants.

Researchers can send emails manually. This would involve sending each email one-by-one through an email client or web application, such as gmail.com. There are two problems with this approach, though. The first is that it can be time-consuming to send many emails this way. It might also be impractical for researchers who intend to contact very large samples (Butler and Crabtree 2017). The second is that researchers might make mistakes when sending emails manually. They could, for example, assign the wrong treatment to a participant, or accidentally fail to send emails to some participants. This is a problem because mistakes such as these could lead to invalid inferences.

Researchers can also send emails automatically with the help of a programming script. There are several advantages to sending emails like this. The first is that it can dramatically reduce the time that researchers spend actually sending emails. Instead of addressing emails to individual participants, scholars would only need to execute a loop of code that would iteratively email each participant. The second is that it reduces the possibility of error. If prepared properly, the script should correctly assign treatments and email all participants. A third advantage is that a script can record the exact time that emails are sent. This is useful if scholars have theoretical expectations regarding how treatments influence not only whether individuals

respond but how long they take to respond as well. Taken together, these advantages suggest that scholars should send emails through scripts.

While researchers might understand *why* they should do this, it is often less clear about *how* they should do this. I provide a detailed outline of this process below. This is based on a set of best practices developed over more than a dozen email audit studies with various collaborators. The outline is broken down into two sections. The first describes the steps researchers should take prior to sending emails. The second describes the steps involved in sending the emails.

5.2.6 Pre-implementation

To begin with, researchers should create an email delivery account for every putative identity used in the experiment. In the past, I used free email accounts from services like gmail.com and yahoo.com. Many free email providers have changed their security policies, though, making them potentially untenable solutions for researchers who want to quickly send their emails through programming scripts. One potential workaround is to modify the script so that it pauses between email sending attempts.[8] Scholars who want to use these services should check their security policies before implementation.

Recently, I have used Google Apps to send email, though other domain hosting services like dreamhost.com would work. While this approach imposes a marginal monthly cost ($5–$10 a month), it allows scholars to get around the security restrictions now common with free accounts. The main downside of this approach is that it requires emails be sent from a domain name that the researcher registers. In several experiments, I have registered and used domains that include a combination of the first and last name for a putative identity. The potential problem with this, however, is that individuals who send emails from custom domains are presumably different from other individuals in important ways. For example, they probably possess higher tech skills and they might have more disposable income. Another option is to register a domain name for a dummy corp (e.g., dummy-corp.org) or email provider (e.g., thefastestemailever.org). In order to make the domain name seem more legitimate, I typically put up a basic webpage at that domain. The trick with this approach is that it can be difficult to register domain names that do not bring to mind specific association(s).

Another potential problem with using a custom domain name is that it might raise participant suspicions. This could increase the risk of experiment discovery. Just as worrying, it could also cause participants to behave in ways other than usual. Unfortunately, there is not a clear solution to this problem, and researchers simply have to evaluate the advantages and disadvantages of each email sending approach within the context of their experiment. Regardless of how they decide to send email, they will need to think carefully about the problems their method might pose to the interpretation and external validity (i.e. scope conditions) of their findings.[9]

[8] I provide an annotated example of this in the online appendix for this chapter.

[9] Pedulla (2018) discusses some of the other issues that potentially limit the generalizability of audit study findings.

After researchers have created the email accounts they will use in their experiment, they should create an additional email account. This will be the master account from which researchers can monitor initial responses and collect final outcome data. All email delivery accounts should be set to forward email to this account.

There are three primary reasons to create a master account. The first is that researchers might want to monitor emails as they arrive, so as to make sure that the experiment was successfully implemented. Researchers should avoid monitoring the original replies, though, as it is very easy to accidentally respond to a message. In some cases, a reply might raise participant concerns and lead to unnecessary problems. The second reason is that it is easier to collect outcome data from one account than many. The third is that bad things can happen with email accounts. Researchers can, for example, be locked out of accounts. It is therefore wise to keep multiple copies of the emails across accounts. Since the master email account will only be used to receive emails, I often create a gmail.com account. This is because Google provides an easy interface for exporting emails.

Once researchers have setup the email delivery and master email accounts, they can attend to other details. They need to write the code that links treatment assignments to strings of treatment text, such as the name of the sender. Scholars should also create the strings of text that comprise the non-random email components, such as email valedictions or salutations.[10] After that, scholars will need to write the code that combines the random and non-random strings of text into a complete email. The online appendix for this chapter includes R code for both steps.

Next scholars should create a script that will deliver their emails. The script should loop through each observation in the dataset. In each iteration, it should extract an observation's email address and treatment details, combine the treatments and other text elements into a complete email, and send the email. After sending the email, the script should save the time that it was sent. This information can be used to confirm that individual emails were sent. It can also be used to create a 'time to reply' outcome measure, as I discuss later. After that, the script should print the observation number for that iteration. This is for diagnosing potential problems later. The online appendix for this chapter includes R code for this loop. It is highly annotated and can be easily adapted to fit a variety of needs.

The final step before implementing the experiment is to test the script. I suggest that researchers do this by sending a limited run of emails (20 or 50) to all project collaborators. The idea here is to test all of the email settings saved in the script. An additional benefit of doing this is that everyone working on the project can look carefully through the sent emails. Particular attention should be paid to the email headers and subject lines, which can be easily ignored. If these emails look good, then the experiment is ready to implement.

[10] In some cases, researchers might want to randomize the valedictions or salutations. This could be a good idea if scholars are concerned about some actor observing similarities across delivered emails (Butler and Crabtree 2017).

5.3 Implementation

Researchers begin implementation by executing the script. In an ideal world, the script will execute successfully, only finishing when all emails are sent. Unfortunately, the script will most likely fail at some point, causing the loop to stop. This can happen, for instance, because an invalid email address remains in the dataset. Most scripts will be unable to parse invalid email addresses and will register an error when reading them. Since the script prints the observation number at the end of each iteration, researchers can manually inspect the dataset to see if the error was caused by an invalid email. If researchers cannot fix the email address, they then should skip that iteration of the loop.[11]

The script can also stop because of email server problems. Sometimes servers, even gmail.com servers, are unable to accept email commands. Sometimes servers will only take so many email commands within a short period of time. In either case, the script available in the online appendix will register a server error. The best way to deal with this problem is to wait a few minutes and restart the loop at the current iteration.

While the script is running, researchers should open the master email account and monitor it for responses. Unless the emails are sent at a really odd time, the participant pool is really small, or the requests will take a while to address, responses should pour in shortly after the script has been executed. There are several reasons to check the responses. The biggest reason is to ensure that the experiment was successfully implemented. Evidence for this can come from email replies, which often include the full text of the sent email. Another reason is to ensure that participants appear unaware that they are part of a study.

5.4 Outcome Collection

Having sent emails, scholars can begin collecting outcomes measures. While audit studies make use of many different outcomes, the primary outcome of interest in many *email* audit studies is a binary indicator that is coded 1 if participants replied and 0 otherwise (e.g., Butler 2014, Bertrand and Mullainathan 2004, and Grose (2014)). There are two ways that scholars can construct this indicator. The first and most common way of collecting this outcome is to read and manually code email responses. The benefit of this approach is that it can be very accurate compared to automatic coding. The problem, however, is that it can be extremely time-consuming to process a large number of emails. Given a sufficiently large sample, it might simply be impractical to do so.

[11] I have assumed here that all emails can be delivered in a single wave. This might not be possible depending on the email solution used and the size of the participant pool. One potential problem here is that some servers might limit the number of emails sent in any given 24-hour period. If researchers need to send emails across multiple waves, they will then need to subset their data into different waves prior to implementation and then execute the script for each wave.

The second way that scholars can collect this outcome is by using a script to automatically code replies. This approach has the benefit of speed, as a script can code thousands of emails in minutes. The disadvantage of this approach, however, is accuracy. In some cases, emails might not be accurately matched with observations. Most of the time this loss in accuracy is relatively trivial, influencing only a small number of observations.

Before using a script to code emails, scholars first need to download the data from the master email account. The exported data will likely be in .mbox format. At this point, scholars could either use the script available in the online appendix or one that they create. The heavily annotated R script performs a number of functions. First, it converts the .mbox file into *N* .eml files, where *N* represents the number of email replies. Second, it reads the emails. Third, it extracts the email addresses that are included in each reply. Fourth, it matches those email addresses to observations in the dataset, link email reply and participant information. Fifth, it creates the outcome measure for each observation.

While a binary *email reply* indicator might be a suitable outcome measure for many research questions, scholars might also be interested in other outcomes. For instance, researchers might want to code whether the replies they receive are positive or negative. This would be easy to do manually. Researchers, however, could also do this automatically. The key here would be to identify words and phrases that are unique to positive or negative replies. Once this is done, scholars could adapt the script discussed above to search for these terms within the email texts that have been linked to participants. If the email contains one or more of the words that uniquely identify positive replies, then an observation can be coded as receiving a positive response. An example of how to do this is included in the script.

They might also have theoretical expectations about how treatments influence *when* participants reply. In this case, they might want to record the time participants take to reply. The R code included in the online appendix can be easily adapted to extract this information from the email replies. Once researchers know when they received email replies, they can subtract the email sent time recorded in the delivery script from this value.

Scholars might also be interested in the length of replies. Reply length could, for instance, be used as a measure of email helpfulness. While scholars can count the words in each reply, it is much easier to do this automatically using either the included code or commercially available software, such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker 2015).

Finally, researchers might be interested in examining the sentiment of the replies. For example, they could be interested in how positive or negative the replies were. Scholars could create this measure manually, by reading and assessing each email. Or they could use one of several software solutions. For example, LIWC can generate measures of positive and negative emotion (Pennebaker 2015). The difference of these two quantities can be taken as a measure of positive sentiment (Crabtree et al. n.d.). Another way that researchers can code this measure is through natural language processing (Manning et al. 2014).

5.5 Analysis

Once scholars have collected their outcomes of interest, they can analyze the results. There are good guides for analyzing experimental results, such as Gerber and Green (2012). For any additional data analysis needs, I recommend Gelman and Hill (2006).

5.6 Extending Audit Studies

Having explained how scholars can conduct audit studies, I want to suggest several ways that researchers can use this study type to examine social phenomena other than discrimination. One potentially promising direction would be to use email audit experiments to test the theoretical determinants of compliance. The idea here is that researchers can create experimental interventions that treat email responses (or non-responses, depending on the case) as a sign of participant adherence to some law, norm, or convention. For example, Terechshenko et al. (n.d.) investigated how international norms and the prospect of public sanctions might influence state respect for human rights. To examine the influence of these factors, they conducted an email audit experiment with a sample of 984 foreign diplomatic missions in the United States, Canada, and the United Kingdom. They emailed each mission with a request for information about contacting domestic prisoners, a right that has long been acknowledged by the United Nations, and varied several attributes of the email. The important point is that receiving an email reply was interpreted as an act of compliance with an international norm. While a design like this could be extended to study compliance with other laws or norms, the main idea here is that email audit studies can be adapted to answer a wide range of substantive inquiries. Scholars have recently used email audit studies to examine the efficacy of economic regulations (Findley et al. 2015) among other phenomena. The only real constraint is the imagination of the researcher.

Another way of adapting email audit studies is to use them as the second part of a larger experimental design. For example, Butler and Crabtree (2017) conduct an experiment to reduce discrimination among public officials. In the first stage of their experiment, they sent a random sample of elected municipal officials an email that called attention to the growing literature on racial discrimination by political elites. In the second stage, they emailed nearly all elected municipal officials with requests for information, varying the racial identity of the putative constituent. They then examined whether the level of discrimination exhibited by officials in their treatment group was lower than the level of discrimination exhibited by officials in the control group.

This type of study suggests the potential of two-stage email audit studies. While Butler and Crabtree (2017) use this design to test the effect of an information treatment aimed at reducing bias, scholars can adapt this two-stage approach to examine the effect of other treatments on discrimination, compliance, and other types of sensitive behavior.

5.7 Discussion

In this chapter, I provided an overview of the steps involved from experimental design to empirical analysis. I then offered detailed recommendations about email address collection, email delivery, and email analysis, which are usually the three most challenging points of an audit study. The focus was on providing a set of primarily technical recommendations to researchers who might want to conduct an email audit study. I concluded by suggesting several ways that email audit studies can be adapted to investigate a broader range of social phenomena. While going from the first to final stage in any email audit study can take considerable time, I think that the results they generate are often worth this cost. I hope that this chapter has helped reduce some of the effort for novice email auditors and thus encouraged the use of this simple but powerful study type.

References

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2012). Does age matter for employability? A field experiment on ageism in the Swedish labour market. *Applied Economics Letters, 19*(4), 403–406.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2013). Are gay men and lesbians discriminated against in the hiring process? *Southern Economic Journal, 79*(3), 565–585.

Baert, S. (2016). Wage subsidies and hiring chances for the disabled: Some causal evidence. *The European Journal of Health Economics, 17*(1), 71–86.

Baert, S., Norga, J., Thuy, Y., & Van Hecke, M. (2016a). Getting grey hairs in the labour market. An alternative experiment on age discrimination. *Journal of Economic Psychology, 57*, 86–101.

Baert, S., De Visschere, S., Schoors, K., Vandenberghe, D., & Omey, E. (2016b). First depressed, then discriminated against? *Social Science & Medicine, 170*, 247–254.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013.

Broockman, D. E. (2013). Black politicians are more intrinsically motivated to advance Blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science, 57*(3), 521–536.

Bushman, B. J., & Bonacci, A. M. (2004). You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs. *Journal of Experimental Social Psychology, 40*(6), 753–759.

Butler, D. M. (2014). *Representing the advantaged: How politicians reinforce inequality*. New York: Cambridge University Press.

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science, 55*(3), 463–477.

Butler, D. M., & Crabtree, C. (2017). Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. *Journal of Experimental Political Science, 4*, 57–67.

Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *The Journal of Economic Perspectives, 29*(3), 81–97.

Costa, M. (2017). How responsive are political elites? A meta-analysis of experiments on public offcials. https://doi.org/10.1017/XPS.2017.14

Crabtree, C., Golder, M., Gschwend, T., & Indriðason, I. H. (n.d.). *Campaign sentiment in European party manifestos* (Technical Report Working Paper).
Desposato, S. (2015). *Ethics and experiments: Problems and solutions for social scientists and policy professionals*. London: Routledge.
Driscoll, J. (2015). Prison states & games of chicken. In *Ethics and experiments: problems and solutions for social scientists and policy professionals*. New York: Routledge.
Findley, M. G., Nielson, D. L., & Sharman, J. C. (2015). Causes of noncompliance with international law: A field experiment on anonymous incorporation. *American Journal of Political Science, 59*(1), 146–161.
Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505.
Fujii, L. A. (2012). Research ethics 101: Dilemmas and responsibilities. *PS: Political Science & Politics, 45*(04), 717–723.
Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.
Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. *The Annals of the American Academy of Political and Social Science, 660*(1), 282–299.
Gelman, A., & Hill, J. (2006). *Data analysis using regression and multi-level/hierarchical models*. New York: Cambridge University Press.
Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.
Giulietti, C., Tonin, M., & Vlassopoulos, M. (2015). Racial discrimination in local public services: a field experiment in the US (IZA DP No. 9290 Working paper).
Grose, C. R. (2014). Field experimental work on political institutions. *Annual Review of Political Science, 17*, 355–370.
Hauck, R. J. P. (2008). Protecting human research participants, IRBs, and political science Redux: Editor's introduction. *PS: Political Science & Politics, 41*(03), 475–476.
Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives, 12*(2), 101–116.
Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & Community, 10*(4), 351–372.
Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*(1), 443–470.
Lahey, J. N., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior & Organization, 70*(3), 508–514.
Lin, W., & Green, D. P. (2015). Standard operating procedures: A safety net for pre-analysis plans, *Berkeley. Retrieved from* www.stat.berkeley.edu/~winston/sop-safety-net.pdf *(2014). Promoting transparency in social science research. Science (New York, NY), 343*(6166), 30–31.
Lohr, S. (2009). *Sampling: Design and analysis*. Boston: Nelson Education.
Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)* (pp. 55–60).
Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination an audit study in academia. *Psychological Science, 23*(7), 710–717.
Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology, 100*(6), 1678.

Moore, R. T., & Schnakenberg, K. (2012). BlockTools: Blocking, assignment, and diagnosing interference in randomized experiments. *R package Version, 0.5–7*. http://rtm.wustl.edu/software.blockTools.htm.

Neumark, D., Bank, R. J., & Van Nort, K. D. (1995). *Sex discrimination in restaurant hiring: An audit study* (Technical Report National Bureau of Economic Research).

Oh, S. J., & Yinger, J. (2015). What have we learned from paired testing in housing markets? *City, 17*(3), 15.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives, 29*(3), 61–80.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181.

Pedulla, D. S. (2018). Emerging frontiers in audit study research: mechanisms, variation, and representativeness. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Pennebaker, J. W. (2015). *LIWC: How it works.* http://liwc.wpengine.com/how-it-works/

Radicati, S., & Hoang, Q. (2011). "*Email statistics report, 2011-2015*." Retrieved 25 May 2011.

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal, 112*(483), F480–F518.

Riach, P. A., & Rich, J. (2004). Deceptive field experiments of discrimination: Are they ethical? *Kyklos, 57*, 457–470.

Sharman, J. C. (2010). Shopping for anonymous shell companies: An audit study of anonymity and crime in the international financial system. *The Journal of Economic Perspectives, 24*(4), 127–140.

Stocksdale, M. (2013). E-mail: Not dead, evolving. http://bit.ly/2AdoOMH

Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences, 4*(1), 8.

Terechshenko, Z., Crabtree, C., Eck, K., & Fariss, C. J. (n.d.). *International norms, sanctioning, and prisoners' rights: A field experiment with foreign missions* (Technical report Working Paper).

Turner, M. A., Ross, S., Galster, G. C., & Yinger, J. (2002). *Discrimination in metropolitan housing markets: National results from phase 1 of the housing discrimination study (HDS)* (Technical report).

Yanow, D., & Schwartz-Shea, P. (2008). Reforming institutional review board policy: Issues in implementation and field research. *PS: Political Science & Politics, 41*(03), 483–494.

Chapter 6
To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis

Mike Vuolo, Christopher Uggen, and Sarah Lageson

Abstract In audits, as in all experiments, researchers are confronted with choices about whether to collect and analyze repeated measures on the unit of analysis. In typical social science practice, this decision often involves consideration of whether to send single or multiple auditors to test for discrimination at a site that represents the unit of analysis, such as employers, landlords, or schools. In this chapter, we provide tools for researchers considering the statistical and substantive implications of this decision. For the former, we show how sample size and statistical efficiency questions hinge in large part on the expected concordance of outcomes when testers are sent to the same unit or site. For the latter, we encourage researchers to think carefully about what is gained and lost via matched and non-matched designs, particularly regarding the finite nature of certain populations, resource constraints, and the likelihood of detection in the field. For both approaches, we make recommendations for the appropriate statistical analysis in light of the given design and direct readers to software and code that may be helpful in informing design choices.

Keywords Audits · Matching · Power · Sample size

In prior work on choosing a sample size for a paired audit (Vuolo et al. 2016), we recommended that researchers consider unmatched designs under certain research conditions. This would typically involve sending one (unmatched) randomly assigned treatment/control "tester" to a single experimental unit (e.g. employer, school, landlord), rather than sending multiple testers to the same unit. In this

M. Vuolo (✉)
Department of Sociology, The Ohio State University, Columbus, OH, USA
e-mail: vuolo.2@osu.edu

C. Uggen
Department of Sociology, University of Minnesota, Minneapolis, MN, USA

S. Lageson
School of Criminal Justice, Rutgers University, Newark, NJ, USA

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_6

chapter, we consider this recommendation in more detail, specifically from the perspective of sample size requirements and statistical efficiency. We weigh the advantages and disadvantages of matched and unmatched approaches statistically and substantively. First, we compare sample size requirements for matched and unmatched designs. Then, we examine empirical data from our own audit (see Uggen et al. 2014) and a series of simulations to contrast the statistical efficiency of modeling approaches for matched and unmatched audits, including predictors that vary both within and between experimental units. Next, we repeat this exercise for a hypothetical case with a much different distribution of outcomes than our empirical audit results. The lesson from these exercises is that the more efficient approach rests heavily on the expected degree of concordance of outcomes. Then, we discuss substantive considerations for matching that should be taken into account alongside efficiency. Finally, in light of these results, we offer further recommendations for the specific choice of whether to match or not.

6.1 Sample Size Requirements for Matched and Unmatched Approaches

Issues of statistical power are central in determining the proper sample size to detect effects, and thus are crucial to careful research design that maximizes the chance for meaningful results and well-expended resources. Although there are several values that determine power, one intuitive way to understand power is in terms of the magnitude difference. That is, at what magnitude difference in the population does a particular sample size have a reasonable chance (typically 80 or 90%) that a statistically significant effect ($p < 0.05$) will be detected in a given sample? In determining an appropriate sample size then, we select a magnitude difference and determine the minimum sample size that would result in a significantly significant effect ($p < .05$) in 80–90% of samples, should a difference of that magnitude or greater exist in the population. In this chapter, we examine this question for audit studies that employ either matched or unmatched designs. Table 6.1 shows audits from our disciplines of sociology and criminal justice. While this list is by no means exhaustive, it does show a rapid increase in the number of published audit studies since Pager's (2003) landmark research. Importantly here, while 9 of the 10 audits prior to 2015 employed matched designs, there is a recent rise in unmatched designs (4 of 11 since 2015). In light of this current landscape of audits, this is a particularly opportune moment to assess the differences in these two design approaches.

Table 6.2 provides notation for the cases of matched and unmatched experimental designs. Throughout this chapter, we use the example of a 2 × 2 experiment, where there is a single treatment and control condition both sent to an experimental unit where a response is measured either affirmatively or negatively. In our empirical matched example (Uggen et al. 2014), we sent two job applicants (testers) to employers (experimental unit) and measured whether they were called back by the

Table 6.1 Examples of audits in sociology and criminal justice by matching design

Author (Year)	Journal	Treatment(s)	Unit	Matched?
Vuolo et al. (2017)	*Res Sociol Work*	Race	Employer	No
Michel (2016)	*Socius*	Sexual identity	Employer	Yes
Rivera and Tilcsik (2016)	*Am Sociol Rev*	Class	Employer	No
Pedulla (2016)	*Am Sociol Rev*	Employment history, gender	Employer	Yes
Hipes et al. (2016)	*Soc Sci Res*	Mental illness	Employer	No
Kugelmass (2016)	*J Health Soc Behav*	Race, class, gender	Psychotherapists	Yes
Gaddis (2015)	*Soc Forces*	Race, academics	Employer	Yes
Gaddis and Ghoshal (2015)	*Ann Am Acad Polit SS*	Race/ethnicity	Roommates	Yes
Wright et al. (2015)	*J Sci Stud Relig*	Race/ethnicity	Churches	No
Decker et al. (2015)	*J Crim Just*	Race/ethnicity, prison	Employer	Yes
Evans and Porter (2015)	*J Exp Crim*	Criminal record, gender	Landlords	Yes
Uggen et al. (2014)	*Criminology*	Misdemeanor arrest, race	Employer	Yes
Wallace et al. (2014)	*Soc Currents*	Religion	Employer	Yes
Wright et al. (2013)	*Res Soc Stratif Mobil*	Religion	Employer	Yes
Widner and Chicoine (2011)	*Sociol Forum*	Arab ethnicity	Employer	Yes
Lauster and Easterbrook (2011)	*Soc Problems*	Sexual orientation, parenthood	Landlords	No
Hogan and Berry (2011)	*City Community*	Race	Landlords	Yes
Tilcsik (2011)	*Am J Sociol*	Sexual orientation	Employer	Yes
Pager et al. (2009)	*Am Sociol Rev*	Felony, race/ethnicity	Employer	Yes
Correll et al. (2007)	*Am J Sociol*	Parenthood, gender	Employer	Yes
Pager (2003)	*Am J Sociol*	Felony, race	Employer	Yes

Table 6.2 Audit notation

		Treatment		
		Affirmative response	Negative/no response	Total
Control	Affirmative response	n_{11}	n_{10}	n_{1+}
		p_{11}	p_{10}	p_{1+}
	Negative/no response	n_{01}	n_{00}	n_{0+}
		p_{01}	p_{00}	p_{0+}
	Total	n_{+1}	n_{+0}	n
		p_{+1}	p_{+0}	$p_{++} = 1$

employer (affirmative response) or not called back by the employer (negative/no response) when one presented an arrest record (treatment) and the other presented a clean record (control). In the table, p_{11} represents the proportion of experimental units where both testers received an affirmative response, while p_{00} is the proportion where both testers received a negative (or no) response. Together, $p_{11} + p_{00} = p_{CC}$, or the total concordance. By contrast, p_{10} represents the proportion of units where the control received an affirmative response and the treatment did not (that is, the cell where discrimination is observed), while p_{01} is the proportion where the treatment received an affirmative response when the control did not. Together, $p_{10} + p_{01} = p_{DD}$, or the total discordance. The difference in the discordance is the object of the test of statistical significance for matched designs and is known as McNemar's test (McNemar 1947). In prior work (Vuolo et al. 2016), we demonstrated that the sample size requirements for a matched audit of a given power level and alpha are based not only on the difference in the discordant proportions ($p_{10} - p_{01}$), but also the total amount of concordance (p_{CC}). We emphasized that the distribution of total concordance p_{CC} across the two constituent concordant cells was irrelevant for sample size calculations; only the total mattered.

In an unmatched design, only the marginal distribution, or the proportions in the "Total" column and row, is observed. But the difference in the marginals will always equal the difference in the two discordant cells, or $p_{10} - p_{01} = p_{1+} - p_{+1}$ (see Vuolo et al. 2016 for a proof). So regardless of whether matched or not, the observed percentage point difference is the same. The sample size requirements to detect that effect, however, may vary greatly. In fact, while the total concordance is the primary consideration for matched designs, the required sample size for an unmatched design must take the values of the two concordant cells into account because as they change, the marginal values p_{1+} and p_{+1} change, despite the discordant cells and $p_{10} - p_{01} = p_{1+} - p_{+1}$ remaining the same.

In our prior recommendation regarding matching versus non-matching (Vuolo et al. 2016:295), we stated that smaller sample sizes are required for matched tests when there is greater concordance, while lower sample sizes are required for independent tests when there is greater discordance. This result is straightforward: when the experimental unit (say, the employer in a study of job discrimination) has little effect on the results of the test, it is irrelevant whether one sends testers to the same unit. We based this conclusion on results from Donner and Li (1990), who showed that the size requirements for the unmatched Pearson's chi-square test are related to a matched test via a weight that measures intraclass correlation. Based solely on sample size, the key to deciding between a matched and unmatched approach depends on whether this weight exceeds or falls below 1.

As expressed in the formula (see Appendix B in Vuolo et al. 2016), this is equivalent to asking whether concordance is above or below .5. This result is intuitive, as .5 marks the threshold at which the effect of the experimental unit (e.g. an employer) becomes more or less influential. That is, it represents the point at which half of the tests had either the same or different outcome occur at the same experimental unit. When more than half of the tests have the same outcome, the experimental unit exerts an effect and matching is preferred, and vice versa. Mathematically, this

weight produces lower sample size values for the matched case when concordance is above .5 and the matched design is preferred. The opposite is true when the concordance falls below .5 because the weight would then produce lower sample size values for the unmatched design, which would then be preferable.

Our prior work focused primarily on matched designs. We here build upon that study by comparing such matched designs against unmatched designs. We begin by further examining sample size requirements, which results in some caveats to our prior recommendation. According to the weight described above, a 50–50 concordance-discordance split designates the point at which one design is favored over the other. While we find that the unmatched case is always preferable in terms of sample size when concordance is below .5, there are still instances where the unmatched case requires lower sample size even when the concordance is above .5. As noted above, the sample size required for McNemar's test for matched designs for a given p_{10} and p_{01} does not depend on the breakdown of the two concordant cells, as only the total p_{CC} matters. For the unmatched case, this breakdown does matter and produces different sample size requirements depending on the marginal distribution, with the well-known result that a higher sample size is required as the distribution approaches values of 50-50, or a random coin flip. Depending on how far the marginal distribution departs from this even split, the sample size requirements for the unmatched case are lower than for the matched case. In our prior work, we emphasized that only total concordance mattered for sample size calculations in matched designs. The distribution across the concordant cells, however, becomes an important factor when deciding between matched and unmatched designs.

We show this result graphically in Fig. 6.1, with each panel displaying sample size requirements for a power of .9 (equivalently a Type II error of $\beta = .1$), Type I error of $\alpha = .05$, and population difference of 5 percentage points between affirmative responses for the treatment and control. Substantively then, the figures display the sample size required to have a 90% chance of observing a sample that would result in a statistically significant effect at the 0.05 level for a population difference of a given amount (in this example, 5 percentage points). To compute sample sizes, we use the formula (Rosner 2011:384–86) and function for the statistical software program R (R Core Team 2015) presented in our prior work (Vuolo et al. 2016). For the unmatched case, we use the R function "power.prop.test" (for derivation and more on this R function, see Chen and Peace 2011:163–66). We note that our McNemar's function calculates the number of experimental units, and the unmatched sample size function calculates the number of units per group. That is, the total observations for both are actually twice the amount shown. We return to substantive implications of this below.

Panel A shows sample size requirements for a distribution where $p_{01} = .05$, $p_{10} = .10$, and $p_{CC} = .85$. In other words, the control receives an affirmative response at 10% of experimental units when the treatment did not, and the treatment receives an affirmative response at 5% of experimental units when the control did not. The remaining 85% of experimental units are concordant, which means both testers uniformly received an affirmative or negative response. For McNemar's test for

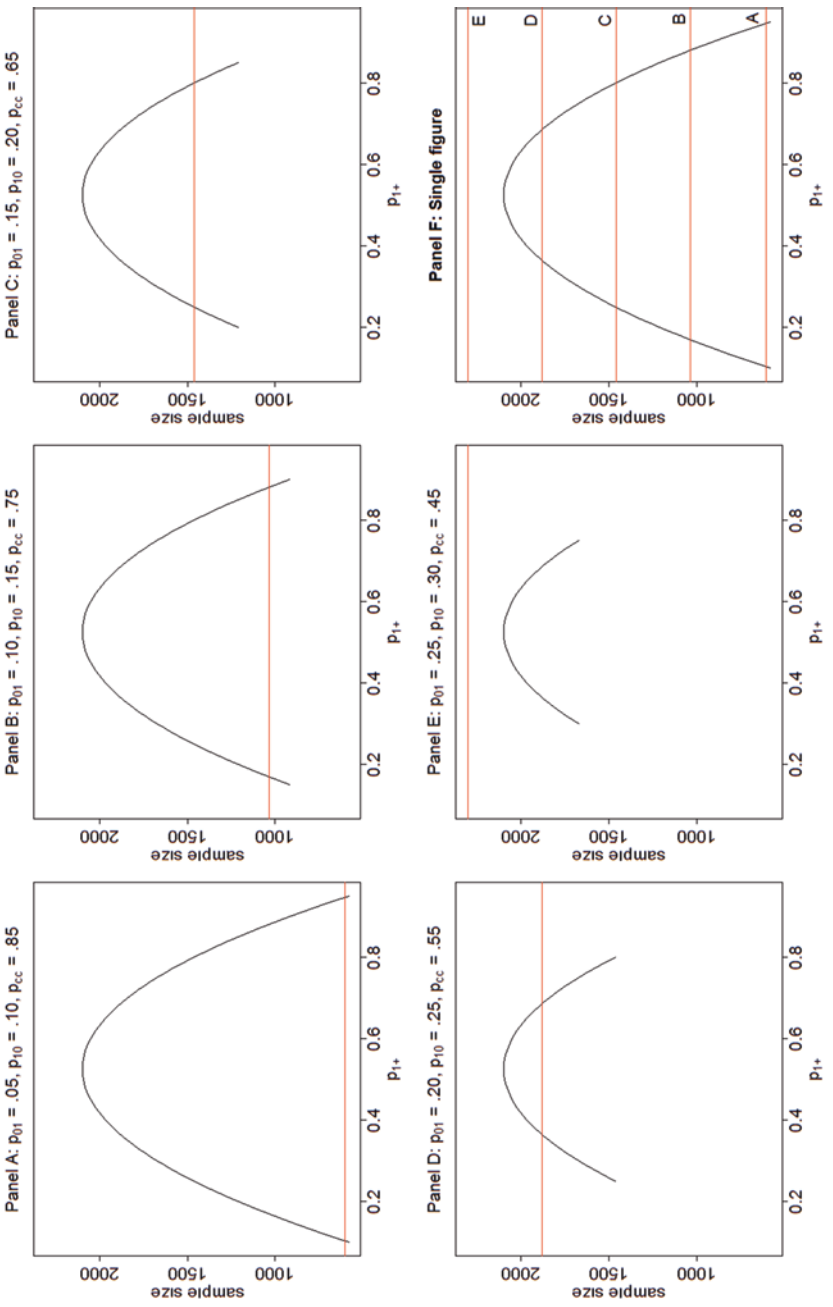

Fig. 6.1 Sample size requirements for unmatched test of proportions vs. matched McNemar's Test for a difference in the concordant cells (and thus equivalently for the marginals) of .05 percentage points
Note: All calculations are for power of .9 and α = .05. Horizontal lines represent sample size requirement for matched design. The curve represents the sample size requirement for the unmatched design

matched pairs, the sample size requirement is the same regardless of the values in the two concordant cells, represented by the horizontal line with a value of 603.

The values of the concordant cells do, however, matter for the unmatched test of proportions, which is shown by the curve. The x-axis is the marginal value p_{1+}, or the total proportion of affirmative responses for the control regardless of the outcome of the treatment test (again, this is all that would be observed in the unmatched case). Since the gap in the marginals is the same as the gap in the discordant cells, the x-axis could also represent p_{+1} (the total proportion of affirmative responses for the treatment regardless of the outcome of the control test) by simply subtracting .05. The implications for sample size are clear and corroborate the results of the weight discussed above: the matched case virtually always requires a lower sample size with a concordance of 85%.

Perhaps unexpectedly, as we move across the panels and the concordance decreases, however, there are scenarios in which the unmatched is preferred from a sample size perspective even when the concordance is below .5. In Panel B with 75% concordance, the required sample size for McNemar's test represented by the horizontal line is 1035, but at the ends of the marginal distribution, a small proportion of possible distributions fall below the horizontal line, thus favoring the unmatched design. As we move across concordance totals through the panels, more of the unmatched curve falls below the required sample size for the matched design. Once the total crosses .5, the curve is completely below the required sample size for the matched case. In Panel E with 55% concordance, the maximum of the unmatched curve, with a value of 2100, is completely below the required sample size of 2305 for the matched case. The curve actually remains constant across the panels, but there is less of the curve to display because the possible marginal distributions decrease with concordance. We illustrate these by showing each of the panels together in Panel F.

Figure 6.2 changes the percentage point difference for the discordant cells and the marginals to 0.15. A virtually identical pattern is observed, but of course with a y-axis that includes much lower sample sizes given the larger discordant difference and one fewer panel since the concordance starts at a lower value of .75. Again, at the highest concordance, the matched case always requires lower sample size (Panel A). After crossing .5 concordance, the unmatched yields lower sample size requirements (Panel D), but there are scenarios in between where the unmatched case still requires lower sample size (Panel B and C).

6.2 Statistical Efficiency in Matched and Unmatched Designs

A more statistically efficient design or experiment is one that requires fewer observations. Paired designs are often encouraged due to a widely-held belief, sometimes made explicit in introductory texts (e.g., Kramer 1991; Dalgaard 2008; Shih and Aisner 2016), that they are more statistically efficient than unmatched designs. This perception is likely based on the case of a paired *t*-test, where paired designs with a

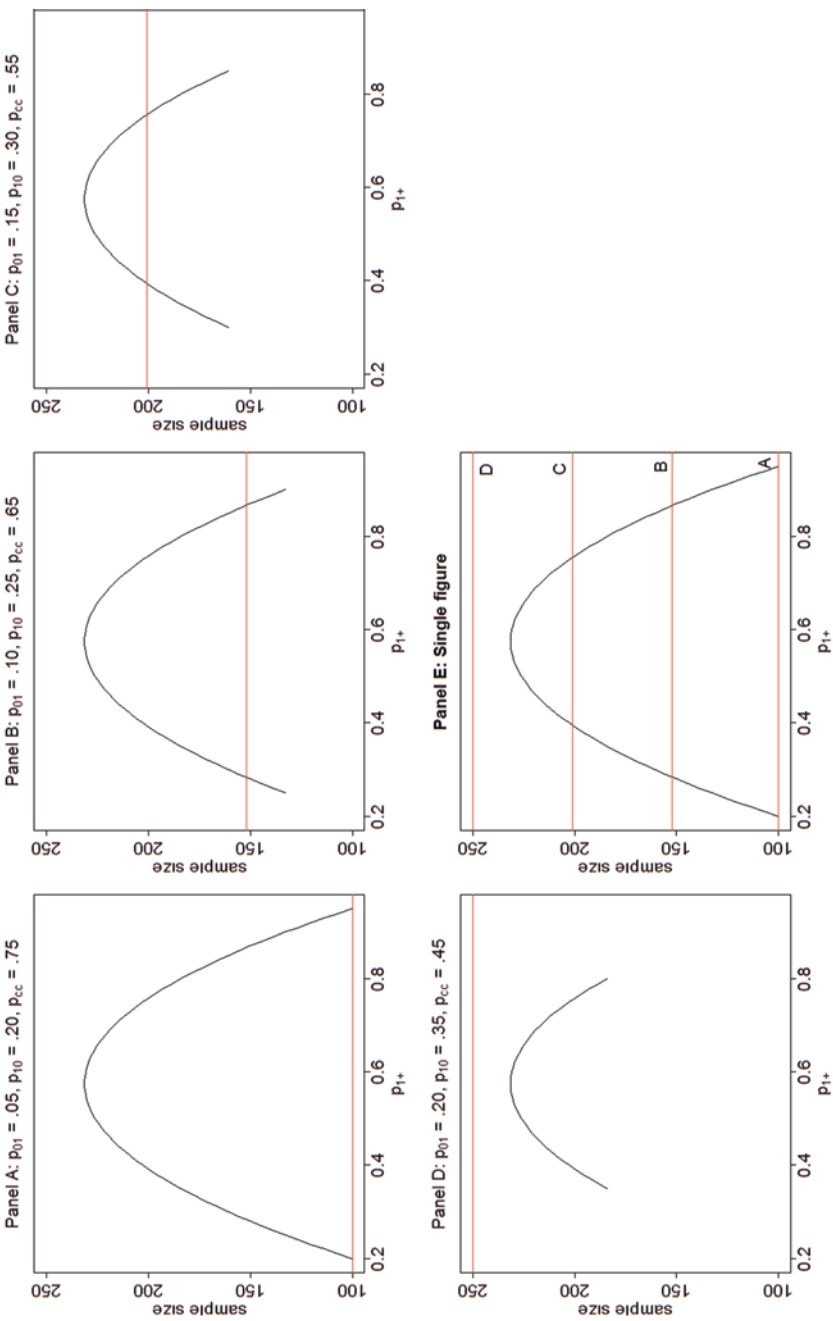

Fig. 6.2 Sample size requirements for unmatched test of proportions vs. matched McNemar's Test for a difference in the concordant cells (and thus equivalently for the marginals) of .15 percentage points
Note: All calculations are for power of .9 and $\alpha = 0.05$. Horizontal lines represent sample size requirement for matched design. The curve represents the sample size requirement for the unmatched design

continuous outcome are often stated as more statistically efficient, even though this is not always the case (Hedberg and Ayers 2015). With McNemar's test as the analogue to a paired *t*-test when the outcome is dichotomous, clearly from what we showed above, the same claim about the efficiency of this design is also misplaced. More specifically, the matched case is only more efficient when concordance is higher than .5, and still only with certain marginal distributions.

In practice, an estimator is more efficient if its standard error, or the standard deviation of its sampling distribution, is smaller. Thus, in this section, we consider the standard errors of the coefficients from standard modeling approaches that would follow from matched and unmatched designs. For unmatched designs, the model is simply a standard logistic regression. For the matched case, we display both a logistic regression with cluster-corrected standard errors and a multilevel logistic regression of observation nested within experimental unit, as both are often seen in audit studies. As would be expected from the sample size figures above, the standard errors are smaller in the matched case when concordance is high (>.5) and smaller in the unmatched case when concordance is low (<.5). Beyond demonstrating this result, however, we also consider two related components that are often important in an audit that are not reflected in the preceding section: (1) a randomized blocking variable at the experimental unit level (as opposed to an observed covariate at the experimental unit level) that thus only exhibits between-unit effects; and (2) the interaction of that variable with the treatment condition. Thus, while the calculations in the prior section solely consider the efficiency of the focal treatment (that which would vary within-units in the matched case), here we consider both blocking effects and the (cross-level in the matched case) interaction. As is shown below, we find the efficiency of the focal treatment and interaction to be at odds with that of the blocking effect.

For example, in our own audit (Uggen et al. 2014), half of the experiments were White tester pairs and half were African American tester pairs. In the interest of testing the arrest record, these pairs were always sent to the same employer, such that the effect of race was actually at the level of the experimental "block" (here, the employer level). This was also the design of Pager's (2003) Milwaukee audit (in which same-race pairs were sent to employers, with one of the testers presenting a felony conviction) and Correll et al.'s (2007) audit (in which same-gender pairs were sent to employers, with one tester signaling parenthood status). In our prior work on sample size for matched audits (Vuolo et al. 2016), we stated that the blocks should be considered separate experiments for the sake of sample size calculations (that is, for example, one experiment for Whites and one for African Americans). Since we are often interested in the effect of this blocking variable, as well as its interaction with the treatment, we now consider this in more detail.

We begin by considering the efficiency of a case with more concordance by using the empirical results from our own audit (Uggen et al. 2014). As described above, we sent same-race pairs to employers to test the effect of an arrest record on employer callbacks. The top of Table 6.3 shows the results of the audit for the whole sample. From the discordant cells, 13% of employers called back the control with the clean record but not the treatment with the arrest record, while 9% of employers

Table 6.3 Distribution of callbacks by criminal record for each paired employer audit (Uggen et al. 2014)

Total		Misdemeanor arrest		
		Callback	No callback	Total
No misdemeanor arrest	Callback	$n_{11} = 60$	$n_{10} = 39$	$n_{1+} = 99$
		$p_{11} = .200$	$p_{10} = .130$	$p_{1+} = .330$
	No callback	$n_{01} = 27$	$n_{00} = 174$	$n_{0+} = 201$
		$p_{01} = .090$	$p_{00} = .580$	$p_{0+} = .670$
	Total	$n_{+1} = 87$	$n_{+0} = 213$	$n = 300$
		$p_{+1} = .290$	$p_{+0} = .710$	$p_{++} = 1$
African American		Misdemeanor arrest		
		Callback	No callback	Total
No misdemeanor arrest	Callback	$n_{11} = 24$	$n_{10} = 18$	$n_{1+} = 42$
		$p_{11} = .157$	$p_{10} = .118$	$p_{1+} = .275$
	No callback	$n_{01} = 12$	$n_{00} = 99$	$n_{0+} = 111$
		$p_{01} = .078$	$p_{00} = .647$	$p_{0+} = .725$
	Total	$n_{+1} = 36$	$n_{+0} = 117$	$n = 153$
		$p_{+1} = .235$	$p_{+0} = .765$	$p_{++} = 1$
White		Misdemeanor arrest		
		Callback	No callback	Total
No misdemeanor arrest	Callback	$n_{11} = 36$	$n_{10} = 21$	$n_{1+} = 57$
		$p_{11} = .245$	$p_{10} = .143$	$p_{1+} = .388$
	No callback	$n_{01} = 15$	$n_{00} = 75$	$n_{0+} = 90$
		$p_{01} = .102$	$p_{00} = .510$	$p_{0+} = .612$
	Total	$n_{+1} = 51$	$n_{+0} = 96$	$n = 147$
		$p_{+1} = .347$	$p_{+0} = .653$	$p_{++} = 1$

actually called back the tester with the record but did not call back the tester without a criminal record. This amounts to a difference of $p_{10} - p_{01} = .04$. The concordance is high at $p_{CC} = p_{11} + p_{00} = .20 + .58 = .78$. Thus, 78% of employers provided the same response to both testers who applied for jobs at their workplace, either calling both or neither job applicants, regardless of criminal record. These specific values of p_{11} and p_{00} produce marginal values of $p_{1+} = .33$ and $p_{1+} = .29$, which represent the only values that would be observed in the unmatched case. The frequencies in the table represent employers, such that there are 300 total employers and 600 observations. The distribution of callbacks for the African American and White testers are also shown in the table.

In the following exercise, we assume that our original audit data collection constitutes the population of employers and draw samples from within to demonstrate efficiency. We know the population parameters of the statistical models for the matched case (as that was our original design), but not for the unmatched case. We therefore simulate an unmatched design by drawing just a single observation from each employer, while maintaining the balance of race and arrest record. The result is 150 observations per treatment/control group (300 total observations), with equal

Table 6.4 Logistic model simulations for case of high concordance

	Unmatched	Matched cluster correction		Matched hierarchical model	
	Simulated	Actual	Simulated	Actual	Simulated
Model 1					
Record	−0.186	−0.187	−0.187	−0.364	−0.368
	(0.215)	(0.127)	**(0.126)**	(0.249)	(0.254)
Model 2					
Record	−0.190	−0.190	−0.190	−0.364	−0.370
	(0.219)	(0.129)	**(0.129)**	(0.249)	(0.254)
White	0.532***	0.530*	0.536*	1.018*	1.028*
	(0.130)	(0.218)	(0.220)	(0.422)	(0.429)
Model 3					
Record	−0.210	−0.207	−0.209	−0.390	−0.398
	(0.326)	(0.189)	**(0.192)**	(0.363)	(0.371)
White	0.519*	0.515*	0.520*	0.995*	1.013*
	(0.252)	(0.248)	(0.250)	(0.483)	(0.494)
Record * white	0.033	0.031	0.033	0.048	0.051
	(0.442)	(0.258)	**(0.262)**	(0.495)	(0.510)

*$p < .05$, **$p < 0.01$, ***$p < .001$

Note: We drew 100,000 simulations of a dataset with 300 observations. These data are drawn from a dataset of 600, represented by the "Actual" model. Lowest standard error for a coefficient is bolded

numbers per race. This results in 75 of each race-arrest combination. To compare matched and unmatched designs of similar sample sizes, we also simulate the matched case by randomly selecting 150 pairs among the 300, while again keeping race balanced.

We repeated this exercise 100,000 times to approximate the sampling distribution. Thus, for each simulation, the value of the coefficients is recorded as an observation in the sampling distribution. The standard deviation of these 100,000 values provides the standard error. In what follows, we focus primarily on the comparative sizes of the standard errors, and not the coefficients or their statistical significance. We restrict our discussion to the standard errors (see Uggen et al. 2014 for the interpretation of the coefficients in our original audit).

Table 6.4 displays the results of these simulations. The lowest standard error for a given model is indicated by bold type. Three models are shown in the rows: one with the effect of the treatment, one that adds an employer-level randomized variable, and the interaction of both. We begin by comparing the unmatched logit model (shown in the first column) to the matched logit model with cluster-corrected standard errors (shown in the second and third columns). Model 1 confirms our sample size calculations for cases of high concordance: the matched sample has a lower standard error according to the cluster-corrected model than the unmatched sample. Model 2 reveals an interesting finding: the employer-level effect for White testers actually has a lower standard error for the unmatched case. In the case of high con-

cordance then, the cluster-level effect is more efficient in the unmatched case than in the matched case.

Figure 6.3 depicts the sampling distributions for the simulated Model 2's, with histograms graphing the frequency of the various simulated coefficient values across 100,000 draws of 300 observations. The lower standard error is reflected in the tighter distributions for White in the unmatched case, and for the arrest record in the matched case. Turning to the matched hierarchical model, the coefficients and standard error are larger. These differences likely emerge because, with a binary outcome, the hierarchical model represents the unit-specific model, whereas the clustered standard errors represent the population averaged model. While technically less efficient, the coefficients and standard error for the hierarchical model are proportional to the cluster correction, such that the same inferential conclusions (i.e., *p*-values) are reached.

Although randomization results in no covariation between the treatment and the cluster-level effect, there is still the potential for a significant interaction between these two measures. For an unmatched design, this interaction reflects four unique treatment/control categories, as each combination is randomly assigned to a single employer. For the matched case, this cross-level interaction represents the difference-in-difference between the treatment and control at employers that were randomly assigned a race pair. Regarding efficiency for Model 3 in Table 6.4, we find that the interaction is more efficient for the matched design when concordance is high, which also remains the case for the main effect of the record treatment. The standard errors for the main effect of the employer-level race effect are virtually identical regardless of matching. We similarly show the sampling distribution for Model 3 in

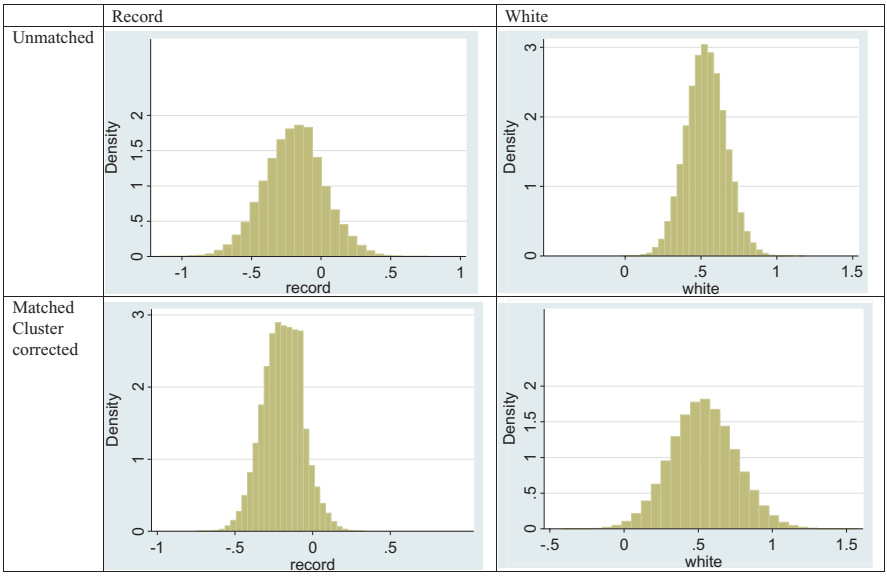

Fig. 6.3 Simulated sampling distributions for Table 6.4, Model 2

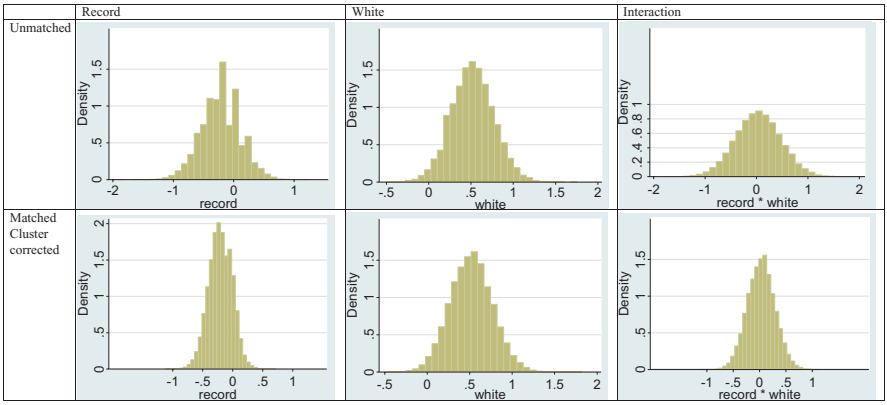

Fig. 6.4 Simulated sampling distributions for Table 6.4, Model 3

Fig. 6.4. As expected, there are tighter distributions in the matched case for the arrest record treatment and the interaction, but nearly identical distributions for the main effect of White.

We next consider the case of low concordance. As demonstrated above, the unmatched case is more efficient according to sample size calculations. This result is intuitive: for matching to matter, the experimental unit (e.g. employers) must respond at least somewhat similarly to the matched pairs (e.g., testers). Lacking such an empirical dataset, we created a mock dataset with low concordance. However, the example of employers as the experimental unit is inadequate in this case, as it would imply that at least 50% of employers called back one tester, but not the other (and would also include cases where the employer called back both). Such callback rates would be unrealistically high, based on every published audit of employers of which we are aware (see Vuolo et al. 2016 for a summary).

By contrast, audits of *landlords* typically have much higher callback rates than employer audits and provide a published example of an unmatched design (Lauster and Easterbrook 2011). Given this, our mock example assumes we have two apartment-seekers (testers) who ask to see an advertised housing unit from a landlord (experimental unit). The outcome of interest is whether or not they were called back by the landlord to tour the unit (considered an affirmative response). The experimental manipulation is the presence of children, signaled by one tester stating they were interested in the unit for their family (treatment) and the other tester not mentioning any family (control). To the extent that landlords view children as a risk (Desmond 2016), we would expect some level of discrimination. Children are hardly disqualifying, however, as there are many likely instances in which a landlord might prefer a family to a single individual. Finally, in contrast to our empirical example of job hunting, we expect landlords to be more eager to show units to prospective tenants, relative to employers' tendencies to call back prospective employees. As before, race is used as a cluster-level effect in the matched case, meaning that same-race pairs inquire about a single unit.

Table 6.5 Hypothetical distribution with lower concordance (0.25) of callbacks by mention of family for each paired audit of landlords

Total		Family mentioned		
		Callback	No callback	Total
No family mentioned	Callback	$n_{11} = 19$	$n_{10} = 135$	$n_{1+} = 154$
		$p_{11} = .063$	$p_{10} = .450$	$p_{1+} = .513$
	No callback	$n_{01} = 90$	$n_{00} = 56$	$n_{0+} = 146$
		$p_{01} = .300$	$p_{00} = .187$	$p_{0+} = .487$
	Total	$n_{+1} = 109$	$n_{+0} = 191$	$n = 300$
		$p_{+1} = .363$	$p_{+0} = .637$	$p_{++} = 1$
African American		Family mentioned		
		Callback	No callback	Total
No family mentioned	Callback	$n_{11} = 5$	$n_{10} = 62$	$n_{1+} = 67$
		$p_{11} = .033$	$p_{10} = .413$	$p_{1+} = .447$
	No callback	$n_{01} = 43$	$n_{00} = 40$	$n_{0+} = 83$
		$p_{01} = .287$	$p_{00} = .267$	$p_{0+} = .553$
	Total	$n_{+1} = 48$	$n_{+0} = 102$	$n = 150$
		$p_{+1} = .320$	$p_{+0} = .680$	$p_{++} = 1$
White		Family mentioned		
		Callback	No callback	Total
No family mentioned	Callback	$n_{11} = 14$	$n_{10} = 73$	$n_{1+} = 87$
		$p_{11} = .093$	$p_{10} = .487$	$p_{1+} = .580$
	No callback	$n_{01} = 47$	$n_{00} = 16$	$n_{0+} = 63$
		$p_{01} = .313$	$p_{00} = .107$	$p_{0+} = .420$
	Total	$n_{+1} = 61$	$n_{+0} = 89$	$n = 150$
		$p_{+1} = .407$	$p_{+0} = .593$	$p_{++} = 1$

While there are many distributions among the four cells that we could have chosen, for the sake of an example, we chose the population distribution in Table 6.5 with a concordance of $p_{CC} = .25$. From the top of the table depicting the whole sample, 45% of those who did not mention family were invited to tour the housing unit when the family was not, while 30% of those who mentioned a family were offered a tour when those who presented as single were not. This amounts to a difference of $p_{10} - p_{01} = .15$. For the concordance, we assumed that landlords called back neither inquiry for a tour (p_{00}) 18.7% of the time, and both inquiries (p_{11}) 6.3% of the time. These specific values of p_{11} and p_{00} produce marginal values of $p_{1+} = .513$ and $p_{1+} = .363$, which again represent the only values that would be observed in the unmatched case. The lower two panels of Table 6.5 distribute these case over the landlord-level effect of race, producing values that imply racial discrimination, but with low concordance. We recognize that this hypothetical example represents a case where there are strong preferences for either the treatment or control (as might be the case, for example, in retirement communities or college campuses, in which housing is segregated by family status), but emphasize that this is only an illustrative case. To offer another hypothetical example, a similarly divided response could

Table 6.6 Logistic model simulations for case of low concordance

	Unmatched	Matched cluster correction		Matched hierarchical model	
	Simulated	Actual	Simulated	Actual	Simulated
Model 1					
Family	−0.617***	−0.614**	−0.616**	−0.614***	−0.616**
	(0.114)	(0.205)	(0.205)	(0.167)	(0.204)
Model 2					
Family	−0.627***	−0.622**	−0.626**	−0.622**	−0.624**
	(0.116)	(0.208)	(0.208)	(0.168)	(0.208)
White	0.463*	0.459***	0.462***	0.459**	0.462***
	(0.210)	(0.115)	**(0.116)**	(0.168)	(0.116)
Model 3					
Record	−0.545***	−0.540	−0.544	−0.540*	−0.546
	(0.176)	(0.292)	(0.294)	(0.240)	(0.294)
White	−0.541*	0.537*	0.539*	0.537*	0.539*
	(0.235)	(0.234)	(0.235)	(0.233)	(0.235)
Family * white	−0.150	−0.161	−0.161	−0.161	−0.159
	(0.232)	(0.416)	(0.419)	(0.336)	(0.418)

*$p < .05$, **$p < .01$, ***$p < .001$

Note: We drew 100,000 simulations of a dataset with 300 observations. These data are drawn from a dataset of 600, represented by the "Actual" model. Lowest standard error for a coefficient is bolded

occur if testers signaled their partisan political affiliation when applying for jobs. Just as children may be favored or disfavored by landlords, some employers would have a preference for Republicans and others would discriminate against them.

Table 6.6 presents simulated models analogous to those presented above. Looking across the columns to identify the lowest standard errors (again shown in bold), the results clearly indicate that the most efficient estimators in the low concordance case are opposite of those observed in the high concordance case. In this case, the two matched modeling approaches yield almost identical results and lead to the same conclusions. In this simulation, the hierarchical and clustered standard error approaches are much more similar, likely because the unit-specific and population average models converge when there is little effect of the unit (as with low concordance). With low concordance, the lowest standard errors for the family treatment effect and the interaction belong to the unmatched design. Clearly, these decisions have implications for the ability to detect a significant effect, precisely the point of a priori power analyses. Any matching efficiency benefit for the main treatment of interest (i.e., the treatment condition researchers would vary within an experimental unit) disappears when the concordance is low. As further evidence of the importance of concordance, the White race effect in Model 2 of Table 6.6 has the lowest standard error when a matched design is used (whereas it was lowest for the unmatched design in Model 2 of Table 6.4). Figures 6.5 and 6.6 reiterate this efficiency for the matched design by again displaying the simulated sampling distribution histograms.

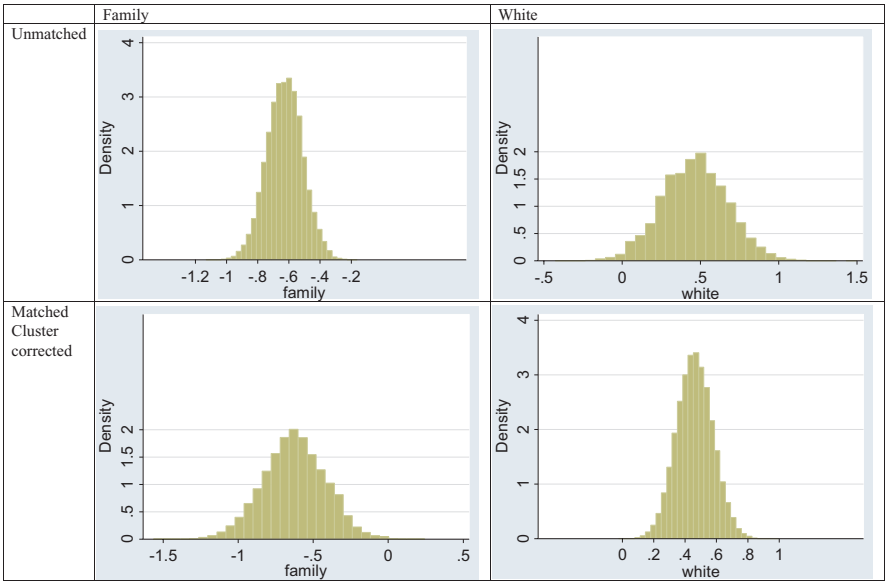

Fig. 6.5 Simulated sampling distributions for Table 6.6, Model 2

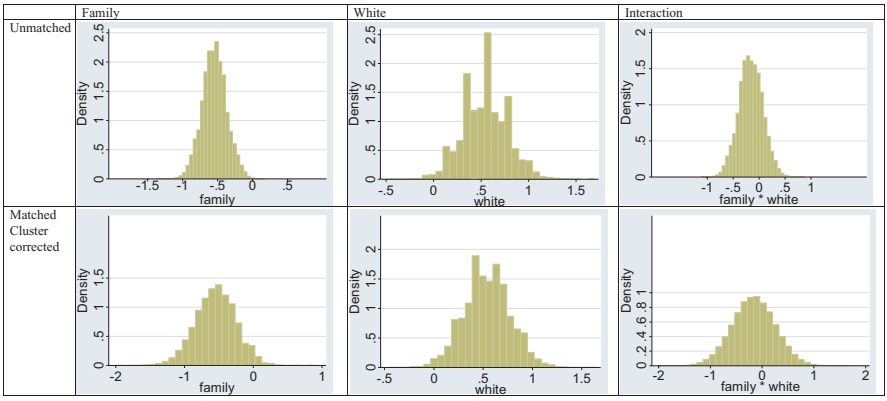

Fig. 6.6 Simulated sampling distributions for Table 6.6, Model 3

6.3 Substantive Considerations to Matching

Until this point, we have considered the benefits of matched and unmatched approaches in purely statistical terms, specifically power and efficiency. In this section, we discuss substantive considerations that may lead researchers to prefer one approach over the other. That is, statistical considerations must be weighed within the context of conducting a real-world social experiment in which non-statistical contingencies typically arise.

The first consideration stems from the number of experimental units to be sampled relative to the total number of observations. In the matched design, two testers are sent to n experimental units, such that the number of observations is $2n_m$ but the total experimental units to be sampled remains n_m (m for matched). In the unmatched design, n_u represents the number of observations per group, such that both the number of observations and the total experimental units to be sampled is $2n_u$ (u for unmatched). This result does not imply that matched designs take twice as many experimental units as an unmatched design because, as shown in detail above, they do not require the same sample size n to detect the same difference in the population. That is, $n_m \neq n_u$, except where the horizontal line crosses the curve in Figs. 6.1 and 6.2. Rather, researchers should calculate the required n for each design, keeping in mind that, if the unmatched design is more efficient with lower n, they will need to sample $2n_u$ units.

The efficiency comparison is calculated on the difference between n_m and n_u, but the need to consider the availability of $2n_u$ units to sample is a substantive consideration, not a statistical one. Whether this is of concern depends, in part, on the total population of experimental units. In correspondence studies of employers that send only a fictional application with no live tester and use several geographic locations (e.g., Wright et al. 2013; Tilcsik 2011; Bertrand and Mullainathan 2004), the population and subsequent sample is typically quite large, such that this is rarely a concern. In a single labor market, however, the population may be more limited. Moving away from employers, one can imagine a case where the required $2n_u$ could exceed the elements of the population. For example, in an audit of admission to medical schools, there are only so many elements of the population. Thus, despite the efficiency gains in the unmatched case, the matched case still might be preferred because it requires fewer experimental units to be drawn from this limited population.

Another limiting factor concerns resources, such as when the audit uses live testers (e.g., Pager 2003 on job applicants) or when the audits cost money (e.g., Stewart and Uggen 2016 on college admissions). In such a case, the researcher typically only has a budget delineated for a predetermined number, or is seeking funds for a given number. With live testers or costly audits, the total number of observations matters greatly because the same funds will be spent to send testers to $2n$ tests regardless of whether they are sent to n_m and n_u employers. Given resource constraints, the number of proposed total observations is often more restrictive due to the prohibitive cost of compensating live testers or paying for applications. Thus, the likelihood of exceeding the elements of the population when sampling is likely low. Here, we would recommend not using the substantive consideration of $2n_u$ compared to n_m. Rather, given the overall lower sample size when resources are limited, the ability to detect a statistically significant effect should take precedence, such that the efficiency comparison of n_m and n_u should be more important.

An additional substantive consideration concerns the possibility of being discovered or "caught" when conducting an audit that relies on deception. Were such an audit to become known to the experimental units (see, e.g., Gaddis's 2015 discussion of educational credentials in his pilot) or the public (e.g., if a college admissions

audit was showcased in the *Chronicle of Higher Education* when researchers were still in the field), results could be biased or contaminated by this information. Moreover, researchers are often interested in testing more than two treatment levels, as is often the case with race and ethnicity (e.g. Ghoshal and Gaddis 2015; Pager et al. 2009), or multiple treatments, such as criminal record and race (Pager 2003; Uggen et al. 2014), race and skill (Bertrand and Mullainathan 2004), and gender and parenthood (Correll et al. 2007) in employer audits. There are two strategies that have typically been employed, both in matched designs. Using the example of employers, first, a researcher can send all treatment combinations to a single employer (Bertrand and Mullainathan 2004; Pager et al. 2009). Second, a researcher can randomly assign the first treatment to the employer (i.e. the cluster) and then send both the treatment and control of the second treatment to each employer (Pager 2003; Correll et al. 2007; Uggen et al. 2014). This choice often hinges on the possibility of being discovered or caught doing the audit, as sending many applications that are too similar on all other characteristics (a necessity to isolate the treatment) could arouse suspicion. But as we demonstrate in Tables 6.3 and 6.5, this choice also has efficiency implications that hinge on the expected degree of concordance, as those two treatments at the between- and within-cluster levels exhibit opposite efficiency and the interaction of the two treatments must also be considered. Thus, this decision is consequential, directly affecting a researcher's ability to detect a statistically significant effect for each treatment.

We want to emphasize a third strategy that would minimize the chance of being discovered while conducting an audit: utilizing an unmatched design when it is expected to be the more efficient design. In this approach, no single employer would be confronted with two applications that look so similar as to raise suspicion. There are certainly scenarios where suspicion could still be aroused, for example, if a researcher does not realize that two establishments share an owner and manager. But this would occur in the matched case as well, and would likely be even more detectable as there would be multiple applications at both sites. An empirical example of this strategy is in Rivera and Tilcsik's (2016) audit of law firms, where they sent a single application to reduce the risk of discovery. Notably, however, this was likely the less efficient strategy, as the modal response was overwhelmingly for no applicant to be invited to an interview (which would likely have also been the case had a matched pair been sent). Thus, any gains in efficiency may have been more than offset by the greater likelihood of detection.

To this point, we have only described the case of two treatment/control levels. Whenever researchers want to send multiple treatments to a single experimental unit, the odds of detection typically increase. One strategy in such cases involves sending subsets of the various possible treatments to a single employer (Wright et al. 2013; Ghoshal and Gaddis 2015). An unmatched design, however, is even less likely to be discovered in the field. But what of efficiency? Our results above concerning the .5 concordance threshold also apply in the case of more than two treatment/control categories, whose corresponding matched statistical test is known as Cochran's Q (Cochran 1950). The formula we derived (Vuolo et al. 2016) from Donner and Li (1990), shows that the same size requirements for the unmatched

Pearson's chi-squared test are related to a matched test via a weight that measures intraclass correlation, and is applicable regardless of the number of treatments. Thus, even if one expands the table to $2 \times m$, where m is the number of treatment categories, the preference between matched or unmatched in terms of efficiency still depends on whether the concordance is above or below 0.5. An unmatched design may not only be more efficient, but it may also reduce the chance of being detected while conducting an audit.

Finally, we emphasize the importance of quality randomization, as departures from randomization are even more problematic in an unmatched design. Why is this the case? Experiments are typically considered the gold standard of research for making causal claims. The randomization process renders the influence of other covariates ignorable (Quillian 2006; Pager 2007). Proper randomization in experiments, however, is demanding, even though the advantages of the method are predicated on it (Berk 2005). If the randomization process is compromised or incomplete, the result would be correlation between the treatment and observed or unobserved covariates, which would limit or altogether prevent the researcher from making the causal inferences that motivated the study. This is not the case when randomization is done well (except by random chance). While certainly not preferable to a well-conducted experiment, in the case of the matched design with incomplete randomization, researchers can fall back on fixed effects to estimate the causal influence of the treatment via the comparison of the outcomes between the two treatments at a single experimental unit (Winship and Morgan 1999; Halaby 2004). With bad randomization in an unmatched design, however, there are no post-hoc remedies to prevent the influence of covariates on the treatment effect, except classic non-causal covariate adjustment. Regardless, we emphasize the need for quality randomization. In a perfect case, sampled units should be pulled completely randomly from the population, and then randomly assigned a treatment category. For live testers, rotating the treatment among the testers is of the upmost importance so that treatments are not confounded with tester effects. Further, all treatments must be simultaneously conducted throughout the experiment, as seasonality (e.g., Schwartz and Skolnick 1962) or an exogenous shock such as a recession (e.g., Vuolo et al. 2017) could alter the outcomes. Such a shock could be correlated with a given treatment, if that treatment was more likely to be assigned at a certain point in the data collection. In the end, quality randomization should always be a priority, which would make this substantive consideration unnecessary.

6.4 Conclusions and Recommendations

Although matched designs are often touted for their efficiency over unmatched designs, we demonstrated that for a dichotomous outcome, this conclusion is not always justified (see Hedberg and Ayers 2015 for an argument concerning continuous outcomes in a paired t-test). Rather, the degree of concordance dictates whether the matched or unmatched design is more efficient. In a situation where concordance is above .5 in the population, the experimental unit itself is exerting an effect

because the majority of employers gave the same response for each test, regardless of treatment or control condition. In this case, a matched design is preferable in terms of efficiency, based on the "more important" treatment that was varied within an experimental unit and its interaction with any randomly assigned cluster-level treatment. When concordance is below .5 in the population, the experimental unit is exerting a smaller effect because employers view the two applicants differently. In this case, an unmatched design is more efficient. We caution, however, that there are cases in which the unmatched design is more efficient, even when the concordance is above .5. And of course, substantive considerations are of utmost importance in creating the research design, as discussed above.

We conclude with recommendations for researchers to parse out this difficult a priori decision in the real world, building and expanding upon those in our past work (Vuolo et al. 2016). Most importantly, researchers should complete an anticipated version of Table 6.2 at the design stage so that they can calculate the appropriate sample size and make an informed decision between a matched and unmatched design. We recommend making several versions of this table in order to establish expected lower and upper boundaries. As in all power calculations, this information can come from two sources. First, past studies of a similar treatment can be used. Our second and preferred recommendation is to also conduct a small pilot. We note that all sample size calculations are based only on the *proportions* in each of the four cells (and the resultant marginals in the unmatched case). Even a small pilot can assist in filling out those proportions and providing bounds.

If the calculated sample sizes for matched and unmatched designs are close or overlap to a great degree within the bounds used, we recommend taking into account the substantive considerations discussed above. When sample size calculations are close in both matched and unmatched designs, the matched design may be preferable for maintaining the possibility of estimating fixed effects (if randomization is compromised) or if twice as many experimental units for the unmatched design may exceed available elements of the population. On the other hand, if one is testing many treatment levels, the unmatched design may be preferable in the interest of not being discovered conducting the audit. Resource constraints may also make the more efficient design preferable, regardless of how close the calculation is.

The matched audit design has become very popular, seemingly becoming the norm due to a perceived efficiency gain and historical momentum(Gaddis 2018; Lahey and Beasley 2018). Unmatched audit designs are less common, but are beginning to appear, as shown in Table 6.1. In most published studies thus far, the outcome has relatively high concordance, in part because the most common response among employers (the most commonly used experimental unit) is not calling either applicant back. Thus, from an efficiency perspective, researchers will likely continue to prefer matched designs. As audits of other types of experimental units become increasingly common, however, the degree of expected concordance is likely to vary substantially (see, e.g. Lauster and Easterbrook's 2011 audit of landlords and Wright et al.'s 2015 audit of prospective church members), with the efficiency implications we demonstrate through our hypothetical example. As audits expend considerable resources and yield important causal inferences, our results

here show that the resultant decision – to match or not to match – should be a central consideration in the design of social experiments.

Acknowledgments The illustrative empirical data used in this article come from a study conducted in partnership with the Council on Crime and Justice and supported by the JEHT Foundation and the National Institute of Justice [grant number 2007-IJ-CX-0042]. We are grateful to Laura DeMarco for the example of a landlord audit and Rob Stewart for the example of college admissions, with each coming from their respective dissertations. The R functions referenced herein, including instructions for use, are publicly available on the first author's website.

References

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1*, 417–433.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*, 991–1013.

Chen, D., & Peace, K. E. (2011). *Clinical trial data analysis using R*. Boca Raton: CRC Press.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37*, 256–266.

Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *The American Journal of Sociology, 112*, 1297–1339.

Dalgaard, P. (2008). *Introductory statistics with R* (2nd ed.). New York: Springer.

Decker, S. H., Ortiz, N., Spohn, C., et al. (2015). Criminal stigma, race, and ethnicity: The consequences of imprisonment for employment. *Journal of Criminal Justice, 43*, 108–121.

Desmond, M. (2016). *Evicted: Poverty and profit in the American city*. New York: Crown.

Donner, A., & Li, K. Y. R. (1990). The relationship between chi-square statistics from matched and unmatched analyses. *Journal of Clinical Epidemiology, 43*, 827–831.

Evans, D. N., & Porter, J. R. (2015). Criminal history and landlord rental decisions: A New York quasi-experiment study. *Journal of Experimental Criminology, 11*, 21–42.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*, 1451–1479.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. *The Annals of the American Academy of Political and Social Science, 660*, 282–299.

Ghoshal, R., & Gaddis, S. M. (2015). *Finding a roommate on craigslist: Racial discrimination and residential segregation*. Available at SSRN: https://ssrn.com/abstract=2605853

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507–544.

Hedberg, E. C., & Ayers, S. (2015). The power of a paired t-test with a covariate. *Social Science Research, 50*, 277–291.

Hipes, C., Lucas, J., Phelan, J. C., et al. (2016). The stigma of mental illness in the labor market. *Social Science Research, 56*, 16–25.

Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City Community, 10*, 351–372.

Kramer, M. S. (1991). Clinical biostatistics: An overview. In H. Troidl, W. O. Spitzer, D. S. Mulder, et al. (Eds.), *Principles and practice of research: Strategies for surgical investigators* (2nd ed., pp. 126–143). New York: Springer.

Kugelmass, H. (2016). 'Sorry, I'm not accepting new patients': an audit study of access to mental health care. *Journal of Health and Social Behavior, 57*, 168–183.

Lahey, J., & Beasley, R. (2018). Technical aspects of correspondence studies. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Lauster, N., & Easterbrook, A. (2011). No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. *Social Problems, 58*, 389–409.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153–157.

Michel, E. (2016). Discrimination against queer women in the U.S. workforce: A resume audit study. *Socius, 2*, 1–13.

Pager, D. (2003). The mark of a criminal record. *The American Journal of Sociology, 108*, 937–975.

Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science, 609*, 104–133.

Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market a field experiment. *American Sociological Review, 74*, 777–799.

Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*, 262–289.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology, 32*, 299–328.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available via http://www.R-project.org

Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review, 81*, 1097–1131.

Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston: Brooks/Cole.

Schwartz, R. D., & Skolnick, J. H. (1962). Two studies of legal stigma. *Social Problems, 10*, 133–142.

Shih, W. J., & Aisner, J. (2016). *Statistical design and analysis of clinical trials: Principles and methods*. New York: CRC Press.

Stewart, R., & Uggen, C. (2016, November 17). *A modified experimental audit of criminal records and college admissions*. Paper presented at the American Society of Criminology Meetings, New Orleans.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *The American Journal of Sociology, 117*, 586–626.

Uggen, C., Vuolo, M., Lageson, S., et al. (2014). The edge of stigma: An experimental audit of the effects of low-level criminal records on employment. *Criminology, 52*, 627–654.

Vuolo, M., Uggen, C., & Lageson, S. (2016). Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes. *Sociological Methods & Research, 45*, 260–303.

Vuolo, M., Uggen, C., & Lageson, S. (2017). Race, recession, and social closure in the low wage labor market: Experimental and observational evidence. *Research in the Sociology of Work, 30*, 141–183.

Wallace, M., Wright, B. R. E., & Hyde, A. (2014). Religious affiliation and hiring discrimination in the American south: A field experiment. *Social Currents, 1*, 189–207.

Widner, D., & Chicoine, S. (2011). It's all in the name: Employment discrimination against Arab Americans. *Sociological Forum, 26*, 806–823.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659–706.

Wright, B. R. E., Wallace, M., Bailey, J., et al. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. *Research in Social Stratification and Mobility, 34*, 111–126.

Wright, B. R. E., Wallace, M., Wisnesky, A. S., et al. (2015). Religion, race, and discrimination: A field experiment of how American churches welcome newcomers. *Journal for the Scientific Study of Religion, 54*, 185–204.

Part III
Nuance in Audit Studies: Context, Mechanisms, and the Future

Chapter 7
Opportunities and Challenges in Designing and Conducting a Labor Market Resume Study

William Carbonaro and Jonathan Schwarz

Abstract In this chapter, we summarize the results of an audit study that we conducted in the city of Chicago. Our study examined how race, high school credentials, and academic grades were related to call backs for jobs. We briefly describe the design and results of our study, and then discuss numerous broader issues about audit studies. Our main goal is to help researchers who plan to conduct similar studies in the future by highlighting and reflecting upon challenges, obstacles, and unexplored opportunities in our work. We conclude with several recommendations for future researchers who plan to use an audit design to study labor market stratification.

Keywords Education · Race · Discrimination · Sorting · Signaling

7.1 Introduction

The relationship between schooling and earnings is one of the best-documented findings in the social sciences (Psacharopoulos and Patrinos 2004). However, as Bills (2003) noted, there are relatively few studies that examine how schooling affects hiring decisions in the labor market. In this chapter, we describe the main findings and contributions of an audit study that we conducted to address this gap in the literature. Our study examined how the educational characteristics of job applicants were related to the likelihood of receiving a call back for a job. We also present numerous important lessons that emerged in the course of conducting our research. In particular, we discuss three main issues. First, we discuss the opportunities and challenges involved with conducting audit studies in different labor markets, with particular attention to the issue of external validity. Second, we highlight the obstacles involved with designing an audit study under conditions of rapid change, due to

W. Carbonaro (✉) · J. Schwarz
University of Notre Dame, Notre Dame, IN, USA
e-mail: wcarbona@nd.edu

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_7

technological innovation. Finally, we describe the challenges involved with generating a random sample of jobs in audit studies, and the implications of non-random sampling for internal validity. Our goal in this chapter is to share our substantive findings, but more importantly, to reflect upon our experience in the field so that other scholars who are interested in similar questions can improve the design of their own studies.

Perhaps the greatest lesson we learned in our review of extant literature and the execution of our own audit study is the following: there is considerable variation both within and across labor markets and the form, process, and "decision rules" of audit studies should reflect the peculiarities of the labor market under investigation. From this lesson, we derive three recommendations for future audit studies. First, new audits should look to established methods to inform their own approach, but not default to strict replication of previous studies in new markets. Second, when designing and fielding audit studies, researchers should always focus on the "prototypical" applicant's and employer's experiences, practices, and overall frame of mind. Finally, audits may be most effective when integrated with other research techniques that can inform both study design and findings.

7.2 Understanding How Schooling Matters in Hiring

We begin by describing the overall design and main findings of our audit study (Carbonaro and Schwarz 2012). Prior experimental studies have found that characteristics such as race, residence, a criminal record, and college quality affect a job applicant's chances of receiving a job callback (e.g., Bertrand and Mullainathan 2004; Gaddis 2015; Pager 2003; Pager et al. 2009). However, there are relatively few studies that specifically focus on how different educational credentials matter for hiring (Deterding and Pedulla 2016; Gaddis 2015; see also Gaddis 2018 for a review). In particular, there are no studies that focus on job applicants who possess no more than a high school degree.

Our paper sought to address two important limitations in the literature on education and labor market inequality. First, few studies (see below) directly examine how educational characteristics affect *hiring* decisions. This is a rather surprising limitation in the literature on labor market sorting, given the prominence of sorting and signaling theories, both of which focus primarily on the hiring decision (Bills 2003; Weiss 1995). Second, researchers typically compare individuals with different *levels* of education (e.g., Bertrand and Mullainathan 2004), but very few have examined inequalities among job seekers with the same level of education. This is an important gap in the literature, because, after winnowing the applicant pool based on differing levels of education, employers typically must use secondary screens to eliminate applicants with the same educational level. Gaddis (2015) conducted an audit study and found that applicants with credentials from more prestigious 4 year colleges were more likely to receive job callbacks than applicant with degrees from less prestigious schools. Deterding and Pedulla (2016) used an audit design to examine whether employers preferred applicants from different 2-year

institutions, and found no difference between call back rates for applicants with nonprofit and for-profit 2-year institutions.

Our study contributes to the literature on labor market sorting by focusing on job seekers whose highest level of education is a high school degree. Prior research based on interviews with employers in urban areas suggested that characteristics of an applicant's high school are sometimes a useful cue about the quality of an applicant (Kirschenman and Neckerman 1991; Moss and Tilly 2001; Wilson 1996). We hypothesized that two characteristics of an applicant's high school degree might affect his/her chance of receiving a call back for a job: the racial composition, and academically selectivity of the applicant's high school. We expected that employers would prefer applicants from more racially mixed high schools (relative to homogenous, majority-minority schools), and from more academically selective high schools. We also hypothesized that academic grades in high school might affect one's chances of getting a job by signaling an applicant's "trainability" to employers. Finally, we hypothesized that the effect of an applicant's credentials and academic background would vary by his/her race. Our expectation was that black applicants would experience greater benefits from a stronger academic record than white applicants, because the positive signal of academic success would offset the negative racial stigma that many applicants experience.

Our field experiment (described below) investigated the following three research questions:

1. Do the racial composition and academic selectivity of a job applicant's high school affect his/her chances of being hired?
2. Does a job applicant's academic record in high school (i.e., grades and class rank) affect his/her chances of being hired?
3. Do school characteristics and an applicant's academic record have the same effect on one's chances of being hired, for both black and white job applicants?

To address these questions, we designed an audit study in which we uploaded fictitious resumes to two job search web sites for job in the Chicago area. Our study follow the basic design of Bertrand and Mullainathan's (2004) influential audit study, but it differs in several important respects. Bertrand and Mullainathan (2004) varied their applications by submitting high quality and low quality resumes to the same posting. In contrast, we made our applicants as similar as possible on all dimensions except for those related to our research questions. Most importantly, all of our applicants had the same level of education (a high school degree) while Bertrand and Mullainathan (2004) allowed the education level of their job applicants to vary. Thus, we held education level constant, and varied other features of the academic backgrounds of our fictitious applicants. Finally, rather than building a bank of actual resumes and alternating key variables, we designed and piloted resumes that would qualify applicants for entry level positions in sales, administrative services, and retail sectors of the labor market.

Our main findings are reported in Table 7.1. An applicant's race had a strong effect on labor market success: black applicants were only half as likely to receive a callback as white applicants (7.3% vs. 12%). Applicants with high grades (3.7

Table 7.1 Callbacks rates for applicants with different demographic characteristics, grades, and high school characteristics

Category	Number of applications	Number of callbacks	Callback rate
Demographic characteristics			
Sex			
Female±	977	102	10.44%
Male	975	88	9.03%
Race			
White±	997	120	12.04%
Black	955	70	7.33%***
Race by sex			
White female±	502	66	13.15%
White black male	495	54	10.91%
Black female	475	36	7.58%***
Black white male	480	34	7.08%***
Student grades §			
High±	562 (1037)	39 (74)	6.94% (7.14%)
Medium	514	64	12.45%**
Low	401	52	12.97%**
High school characteristics			
Racial composition			
Mixed race high school±	973	99	10.17%
Black high school	979	91	9.29%
Academic selectivity			
Selective high school±	977	91	9.31%
Neighborhood high school	975	99	10.15%
High school types			
Mixed-race, selective H.S.±	486	44	9.05%
Predominantly minority, selective H.S.	491	47	9.57%
Mixed-race, neighborhood H.S.	487	55	11.29%
Predominantly black, neighborhood H.S.	488	44	9.02%
Overall	1952	190	9.73%

Note: § Students in the 'high' group are restricted to the cases where the three grade conditions were rotated weekly. The numbers in parentheses combine all of the 'high' grade cases, before and during the manipulation of 'grades' in the study
Statistical tests were dependent sample t-tests, with each reference category denoted by '±'. **p < .01 ***p < .001

GPA) were only half as likely (7% vs. 13%) to receive job callbacks relative to applicants with medium (3.0 GPA) and low grades (2.3 GPA). Neither the academic selectivity nor the racial composition of an applicant's high school were related to his/her chance of receiving a callback. Finally, our findings also showed that the

effects of academic grades and school characteristics on receiving a callback did not vary by the applicant's race. None of the variables in our study changed the effect of an applicant's race on job callbacks in our analysis.

In the remainder of the chapter, we will reflect upon challenges, obstacles, and missed opportunities in doing our study. We will conclude with several recommendations, based on our experiences, for future researchers who plan to examine educational inequality using experimental methods.

7.3 External Validity Issues: Selecting a Research Site

Prior theory and research guided the main research questions and overall design of our audit study. One limitation that we identified among audit studies that examine labor market sorting was an exclusive focus on large urban labor markets. In planning our project, we decided that our study could make a contribution by conducting the same audit study in both a smaller and larger labor market. At one level, we believed that an audit study of a smaller labor market would increase the external validity of our findings, and reveal whether research on large labor markets can be generalized to smaller labor markets.

We hypothesized that labor market size would be particularly important in studying high school credentials. We suspected that differences between high schools would be stronger and less ambiguous signals in smaller labor markets because employers could more easily recognize the contours of the educational landscape when it is less vast. Thus, our original study included plans for a comparison of two research sites, one small, and one large: South Bend, IN and Chicago, IL. As we explain below, we were only able to conduct our audit study in Chicago, due to numerous unforeseen obstacles that arose and made our planned study of hiring in the South Bend area impossible to complete.

7.3.1 Selecting a Small City

Based on our prior knowledge and experience, we believed that South Bend would serve as an excellent site for an audit study of a smaller city. South Bend is a city with about 100,000 residents, and it is part of a larger metro area (MSA, South Bend-Elkhart-Mishawaka) of 320,000. The city of South Bend is demographically diverse with a racial composition in 2010 that was roughly 60% white, 25% black, and 13% Latino. The median household income for the city is approximately $32,000, and 16.7% of the city's population falls below the federal poverty line. The five largest employers in South Bend are (in descending order): the University of Notre Dame, Beacon Medical Group, South Bend Community School Corporation, AM General, and Saint Joseph Regional Medical Center. Thus, South Bend is a

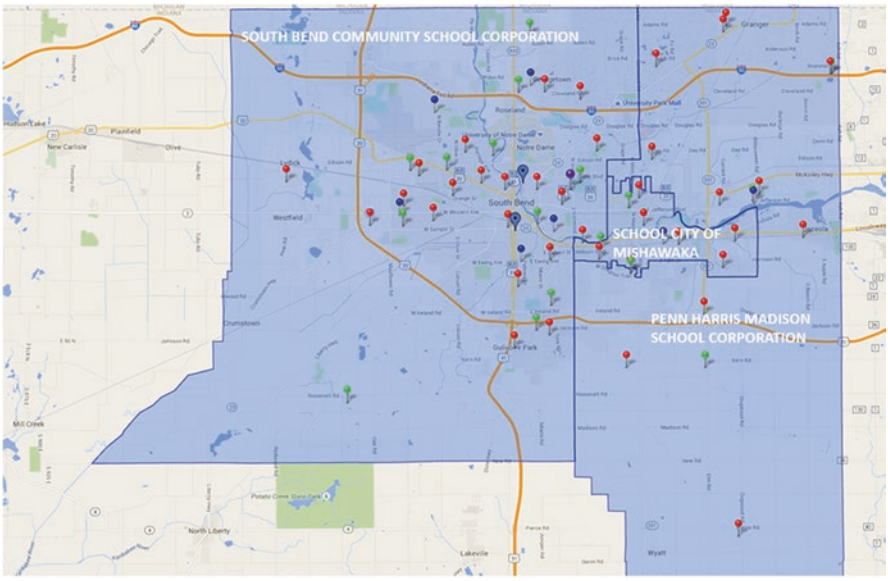

Fig. 7.1 Map of south bend schools and local school districts
Note: School district boundaries are indicated by the dark blue lines on the map. Each pin in the map represents a school, where elementary schools are red, middle schools are green, and high schools are blue.

former manufacturing hub that is slowly shifting toward a post-industrial economy dominated by service sector jobs.

It is important to note that South Bend is geographically contiguous to two other communities: Mishawaka and Granger. (See Fig. 7.1 for reference.) The population of Mishawaka is only half as large as South Bend, with roughly the same median household income, but half the poverty rate. However, Mishawaka is much less demographically diverse than South Bend, with whites making up 85% of the population, alongside small proportions of blacks (7%) and Latinos (4.5%). Granger lies to the east of South Bend and is smaller than Mishawaka (with a population of 30,000). It is extremely racially homogenous, with 94% of its residents classified as white. It is also relatively affluent with a median household income of about $80,000. Although South Bend, Mishawaka, and Granger are technically separate municipalities, it is important to emphasize that they form one large, connected geographic and economic unit. One can easily cross from South Bend into Granger or Mishawaka without recognizing that one has left the city. The three municipalities form one large labor market, with many residents living in one area, but working in another (e.g., people who live in Granger but work in South Bend, etc.).

For our purposes, one theoretically important feature of the South Bend area is that each of these three localities has its own school district (see Fig. 7.1), and

unsurprisingly, each district's demographic profile reflects the make-up of its community. The South Bend Community School Corporation (SBCSC) is the largest and most demographically diverse district, with one-third white, one-third black, and one-fifth Latino students. Roughly 75% of students in SBCSC receives a free or reduced priced lunch (FRPL). Comparatively, the School City of Mishawaka is much more racially homogenous (80% of students are white), but similar in family background (about 65% FRPL). Finally, Granger is served by the Penn-Harris-Madison (PHM) district. Nearly all (about 95%) of students attending PHM schools are white, and roughly 10% of students are FRPL.

In the South Bend community, PHM is widely considered the highest quality district, SBCSC is considerably less reputable, and School City of Mishawaka, slightly above SBCSC. Test scores and graduation rates in the three districts are certainly consistent with these general rankings. Indeed, one of the main arguments in favor of living in Granger is the perceived higher quality of the school system.

In short, the race and class composition of the three districts is consistent with judgments about school quality in the South Bend area. Given these strong differences in perceptions of school quality in the community, we hypothesized that job applicants who attended high school in these different districts would be evaluated differently by local employers. We were particularly interested in whether black students would disproportionately benefit from graduating from the less diverse districts with more white students. The South Bend labor market context seemed quite similar to the labor market described Eaton (2001) in her interview study of the METCO program in Boston, and the surrounding suburbs, although on a much smaller scale. Eaton reported that employers were very eager to hire to black job applicants who were participants in the METCO program, because these students held diplomas from highly regarded white suburban schools (outside of Boston). We concluded that South Bend would be a useful case for testing this qualitative finding, particularly since South Bend's labor market is considerably smaller than Boston's, and most employers would likely be very familiar with the distinctions between high schools in these three local districts.

7.3.2 Finding a Larger City for Comparison

The next stage in planning our project was to select a large city for comparison with our small city. We chose Chicago as our second site for our field experiment for several reasons. First, it is geographically close to South Bend, and macroeconomic conditions tend to be fairly similar in the two cities. Second, Chicago also has a very large public school district, with about 400,000 students, and nearly 200 public high schools. In contrast with South Bend, the sheer size of the Chicago Public School system makes it much more difficult for employers in Chicago to distinguish among public high schools. Based on his field research, Rosenbaum (2001) argued that some Chicago employers formed specific working relationships with certain high schools, but were less familiar with others. Other qualitative studies suggested that

employers used credentials from Chicago Public Schools as negative screen in hiring (e.g., Kirschenman and Neckerman 1991, and Wilson 1996). Given the socioeconomic and demographic makeup of the Chicago school district (which is largely minority and lower income), it is likely that many Chicago employers did not have firsthand experience with Chicago Public Schools (CPS), either as students, or parents with children in CPS. Thus, Chicago would be an excellent case to compare with our study of South Bend.

Third, Bertrand and Mullainathan (2004) studied Chicago (as well as Boston) in their much cited study, so we believed that it would useful to build on their analyses by collecting new data, and asking similar, yet distinct, questions. Finally, Chicago offered an opportunity to study a stratified public school district, where students could either attend a selective program (with admission criteria) or a neighborhood school. These selective high schools were few in number (about ten), and might stand out to employers as more recognizable signals of applicant quality. Thus, our planned study could compare one school system with distinctions in quality between districts (South Bend), and another with distinctions within the same district (CPS).

In our study, we signaled high school characteristics to employers through the names of our high schools. In the case of race composition, we selected predominantly black schools that were named after famous African Americans. For academically selective high schools, we choose magnet schools that had "college prep" in the school's name. In retrospect, it would have been useful to conduct some unstructured interviews with employers and inquire about their familiarity with and knowledge of CPS high schools before collecting our data. Doing so would have helped us highlight differences between schools that might have been meaningful to employers. However, in hindsight (and as we discuss more fully below), it is not entirely clear which employers we should have contacted to conduct these interviews. The sampling frame for employers who are seeking high school educated applicants is by no means self-evident.

7.3.3 Unforeseen Difficulties in Studying Our Small City

It was relatively easy to conduct our field experiment in Chicago. There were ample numbers of jobs to apply for by submitting resumes through two on-line job search engines. Unfortunately, numerous unanticipated problems arose in conducting our audit study of the labor market in the South Bend area. First, the on-line search engines that we used to generate our sample of jobs in Chicago produced very few jobs each week in South Bend (typically, less than ten, many of which were "reposted" from week-to-week). It quickly became apparent that we could never generate a sufficient sample of jobs from these search engines for our study of the South Bend labor market. We also quickly recognized that the largest employers in the area (e.g., the University of Notre Dame, and local area health care providers) had their own application intake systems. These systems did not allow for us to apply for jobs by simply uploading a resume and answering pro forma questions. Instead, we

had to provide much more detailed information, and often included requests for social security numbers (SSN) (presumably for background checks). Since we could not provide fake social security numbers, these options were foreclosed to us.

Big box retailers (such as Target, etc.) and department stores (such as Sears) were another potential source of job openings that we considered in our South Bend sample. These retailers also used intake systems that involved either walking in the store and filling out a job application on a computer or submitting a comprehensive online application. Once again, these employers required personal information (such as a Social Security number) from applicants that we could not provide.

Even if we were able to successfully complete an application for one of these retailers, it was not immediately clear that these applications were comparable to those posted through on-line search engines. Rather than posting a position when it became available and soliciting applications for that position, these retail stores regularly accepted applications for employment. Applicants were asked to "check a box" for jobs that interested them in a variety of fields (inventory/stock, cashier, customer service, etc.). When there was a need for a new cashier, for instance, the manager in charge of cashiers would presumably look at recent applicants who expressed an interest in working the cash register, and contact promising applicants. As researchers, we would not know when and for which jobs our applications might have been considered or how long each application would be retained. In other words, we would submit "one application" that might be reviewed independently by half a dozen hiring managers. This lack of clarity about how many times an application may be under active consideration complicates how we, or any researcher, might calculate an "application" or callback rate. Finally, other large local businesses (such as grocery chains) still relied upon "walk-in" applications, which entailed an entirely different methodology than our resume study.

Ultimately, all of these limitations caused us to abandon our plans to conduct an audit study in South Bend. Our initial plan was to standardize both the design of our experimental procedures and our sampling procedures across our two sites, but we ultimately realized that pursuing one of these goals undermined the other. An audit study of South Bend would have required different sampling and design features than we implemented in our study of Chicago. Doing so, would have made it impossible to directly compare the South Bend findings with those from Chicago.

In the end, Chicago became our only source of data, which forced us to abandon the goal of comparing two distinct labor markets to extend the external validity of research. This was unfortunate because we strongly suspected that school effects on call backs would be fairly weak in Chicago – which is indeed what we found. Perhaps we would have found something different if we had studied South Bend, but we would be left with an "apples-to-oranges" comparison. Ultimately, we recognized that both the sampling and design features of audit studies are best adapted to local labor market conditions, in order to maximize experimental realism from the perspective of the employer and job applicant. However, this may have to come at the expense of the researcher's ability to generalize across multiple research sites.

7.4 Understanding the Hiring Process

One major strength of audit studies, when they are conducted as field experiments (rather than lab experiments), is that they are high in mundane realism. More specifically, researchers are studying "real life," meaning that their data are drawn from the social situations and social actors that are directly relevant to their research questions. Audit studies are also high in experimental realism: participants perceive the experimental conditions to be real and meaningful. Indeed, these two features of the audit study made it an attractive method for examining our research questions.

In devising our study, we drew upon research that suggested that employers actively screened job applicants based on meaningful cues that reduced uncertainty about each applicant. Much of the research that influenced our thinking about the design of our project was drawn from qualitative research (mostly interviews) with employers who were involved in the hiring process (Bills 1988; Kirschenman and Neckerman 1991; Rivera 2011, 2012, 2015; Rosenbaum 2001; Wilson 1996). These studies suggested many different ways in which employers used numerous different screens to select job applicants. Regarding the use of educational credentials as screens, fieldwork is somewhat mixed. Kirschenman and Neckerman (1991) and Wilson (1996) found that employers used information about high school credentials as meaningful screens that helped them evaluate job applicants. In contrast, Rosenbaum (2001) suggested that employers cared little about either academic grades, or distinctions among high schools.

The aforementioned studies helped us design our study because they provided a model of employer decision-making that shaped our hypotheses. However, we also recognized some important limitations in these qualitative studies that our project could address. It is reasonable to be skeptical of interview data by employers, because (a) employers may intentionally mislead interviewers by providing socially desirable responses (or intentionally alter responses for other reasons), and (b) employers may not fully understand at a conscious level what factors affect their decision-making regarding job applicants (Pager and Shepherd 2008; Quillian et al. 2006). Thus, we believed that data from an audit study could contribute to the literature by establishing whether an applicant's educational background is a useful screen for employers. It should be noted that our findings indicated that our skepticism was at least partly warranted: academic grades *did* matter to employers, although school characteristics did not.

One major limitation in the literature that we failed to fully appreciate was that many of the studies on employer behavior in hiring are now quite dated, with much of the data collected in the late 1980s and late 1990s. The near ubiquity of computer hardware and access to information through the internet has likely changed how employers evaluate job applicants (see Cappelli 2012). We can think of at least two ways in which computer technology and access to the internet may have affected employer behavior in hiring.

First, the emergence of internet job posting services have greatly expanded the number of job applicants for a given opening (Cappelli 2012). Increased numbers of

job applicants for openings have fundamentally changed how the screening and hiring process works. At the very least, employers likely spend less time reading each applicant's materials, and this alone should affect the evaluation process. Time constraints should make screening more common, and employers will likely use cruder (and fewer) screens to weed out applicants. Cappelli (2012) reported that many large employers have replaced human evaluators to screen job applicants, and instead have relied upon computer software with algorithms that screen applicants based on several weighted criteria.

These changes add a whole new level of complexity to designing an audit study. How do we know whether a human has actually laid eyes upon our fictitious applicants' resumes? How common is the use of software to screen applicants for high school level jobs? What types of parameters are included in software algorithms that might give some applicants advantages over others? The research literature on employers provides little guidance on these questions. In addition, among the employers who "called back" our applicants, there was some clear variation among employers in their screening procedures. Some employers clearly undertook a "mass processing" approach to screening job applicants, sending highly routinized and standardized responses. Other employers seemed less affected by technology, and followed a more traditional "personal" approach.

We also observed unexpected geographic variation in application screening (see also Besbris et al. 2018 and Gaddis and Ghoshal 2015 for more on geography). The research design for our study was predicated on the belief that applications for high-school-level service jobs are reviewed locally. Telephone area codes from which we received callbacks, however, indicate that this was not always the case. In order to accurately measure geographically-oriented phenomenon, audit studies should take into account both automation, which removes human review and technical advances that have moved application review or decision-making out of the local setting.

Second, access to the internet makes it easier for employers to reduce uncertainty when evaluating job applicants. For example, after the applicant pool has been winnowed, it is possible that final decisions about call backs include internet searches to reduce lingering uncertainty about applicant characteristics. Street addresses can be entered into Google maps, which with a street view, might be useful to employers in their decision making. Likewise, if an employer is unfamiliar with a given high school, s/he may use an internet search to learn more about it.

Given our experience, an important first step in designing audit studies is to first interview employers (particularly personnel are responsible for evaluating job applicants) to understand how technology has affected the hiring process. Based on this field work, we suspect our experimental design could have been better aligned with the practices that employers report using in their evaluation of applicants for jobs.

Qualitative fieldwork can also help researchers better understand and explain their findings. Audit studies are powerful designs for detecting causal effects, but ultimately they remain "black boxes" whereby the processes that generate the observed effects (or non-effects) remain invisible to the researcher. This issue had particular relevance for interpreting several of our reported findings.

For example, we can only speculate *why* educational credentials were unrelated to callback rates in Chicago. One possibility is that employers recognized and understood differences in the high school attended by our applicants, but simply found this information unimportant in hiring. It is also possible that employers either (a) did not distinguish the race-composition and academic selectivity of our high schools from the names on our resumes, or (b) made hiring decisions long before ever learning anything about the applicant's high school. Ideally, a manipulation check (where we confirmed whether or not an employer experienced the "treatment" as we intended) could have helped sort out these different possibilities, but it was logistically impossible to incorporate this feature into our design. This issue seems particularly important when interpreting "null" findings in audit studies: without a manipulation check, one cannot distinguish between null findings that are the product of social processes, and those that are due to poor design.

Likewise, we are forced to speculate regarding our finding regarding the effect of GPA, was somewhat counterintuitive: academically strong students experienced a sizable penalty in job callback rates. Our explanation for this finding – that academically strong applicants seem overqualified, and therefore risky hires – was plausible, but ultimately, it remains ad hoc. Interviews with employers could have helped us interpret this finding, but as noted above, employer's explanation of their own practices should not be taken at face value.

7.5 Sample Selection Issues: Which Jobs to Sample?

One of the most vexing and complicated aspects in designing our audit study involved devising clear and consistent rules about which jobs to include in our sample. We have already described our decision to exclude certain jobs where employers used specific on-line intake systems, as well as "walk-in" applications at large retail and grocery stores. These were decisions that were driven by pragmatic and logistical challenges that we could not resolve. However, it is important to recognize that these decisions surely affected both the external validity, and possibly the internal validity of our findings.

One rule that we imposed in drawing our sample of jobs was to adhere to employer requirements regarding credentials and work experience. For example, if an employer posted a job ad that "required" a 4 year degree, or more work experience than our applicants listed on their resumes, we would not include this job in our sample. These constraints eliminated many job postings from our sample, and greatly reduced the overall number of jobs that our applicants could apply for in a given week. We think that the accuracy with which we replicated the job-search experience of applicants in the labor market under investigation far outweighs the loss in sample size.

The rationale behind these rules regarding which jobs to include in our sample was straightforward: we wanted our sample to reflect the portfolio of job opportunities available to individuals with only a high school degree, who are unable to adjust

their credentials or experience in response to job postings. However, in imposing these rules to generate our sample, we admittedly were extrapolating based on our own ideas regarding how high school graduates search for jobs. Upon reflection, if we had data on the job applications submitted by actual job seekers, we would surely see variation in job search strategies among our respondents. Some job seekers would be more expansive in their approach than our rules permitted, and others less so. Accordingly, Pager and Pedulla (2015) found substantial variation in job search strategies among applicants, as well as racial difference in search strategies. Ultimately our sample of jobs was a function of the Chicago labor market, our ability to apply online, and the selection rules we put in place. We hoped that our sample would include a diverse array positions that would reflect the heterogeneity of the labor market available to high school graduates, but we question whether auditors can be fully confident that their research design meets this goal.

While external and internal validity are often treated as separate issues, it is important to recognize that sample selection can lead to biased estimates of causal effects (Elwert and Winship 2014). For example, we found that black applicants were roughly half as likely to receive a job callback as white applicants. Can we conclude from this finding that the population of white applicants with a high school degree in Chicago is twice as likely to get a job callback as otherwise identical black applicants? The answer depends on the actual job search behavior of high school graduates in Chicago. It is possible that our sample restrictions caused jobs where racial screening is common to be over-represented in our sample, thereby upwardly biasing the estimate of racial preference in our study. (An under-representation of jobs using racial screening is possible too, which would lead to a downwardly biased estimate.) Unfortunately, we have no way to externally validate whether our sample of jobs is biased (let alone the magnitude of that bias) since the sampling frame is unobserved. It reassuring that in-person audit studies of sub-baccalaureate job seekers (Pager et al. 2009), which likely sampled different types of jobs than our study, reached similar conclusions regarding the effect of race on call backs.

Given the importance of this potential problem of sample selection bias, it is surprising how little attention this topic has received in the literature. When designing audit studies, researchers must consider how their sampling procedures might create a biased sample of jobs that might undermine the internal validity of their estimates. One approach to minimizing possible bias due to sample selection would be to collect data on application behaviors from a sample of job seekers in the population of interest. In our case, it would have been very helpful to know what types of jobs job seekers in our population of interest actually pursued. This information could have shaped the rules that we devised for sampling jobs, and increased the external and internal validity of our findings. In addition, such data would have been extremely helpful in designing our proposed audit study of South Bend. The modest number of on-line postings in South Bend clearly indicated a broader approach to sampling jobs was needed for our proposed study, and interviews with local job seekers could have been enormously helpful to us in finding new strategies for overcoming the challenges that we faced. Indeed, we assumed that job search and application strategies were directly comparable in large and small labor markets, but that

was an empirical claim that we could have investigated by collecting data, and adapting our experimental design accordingly as early in data collection as possible.

7.6 Recommendations for Future Research

The findings from our audit study made several contributions to the literature on the importance of education and race in the labor market. We found that for high school graduates seeking high school level jobs, race and academic grades affected an applicant's chances of getting a job callback. The racial composition and academic selectivity of an applicant's high school was unrelated to job callbacks, and the effects of our treatment variables did not vary by an applicant's race. However, as described in this chapter, there were limitations in our study we could not resolve. To conclude, we share two main recommendations for researchers who want to conduct audit studies on labor market sorting processes.

First, researchers should consider how time and place affect the social processes that they are studying. Researchers often view "external validity" through the lens of "place": findings that are generalizable should be consistent across different social contexts and geographic units. Indeed, one of the motivations for our study was to examine *whether* place mattered by comparing data from a large and small labor market. As we explained in this chapter, we learned an unexpected lesson about how place "matters": job search strategies in smaller labor markets differ substantially from large labor markets, and different sampling and experimental design features may be needed to conduct research in labor markets that differ in size and scale. Researchers are typically less concerned with "time" as a component of external validity, but we also appreciated its importance in reflecting upon our study. The emergence of the internet and computer technology has revolutionized how employers gather information and screen job applicants. Field studies that predate this new era are less informative, and likely describe a much smaller share of the labor market than they used to.

These challenges of generalizing across both time and place lead us to recommend that researchers who want to conduct audits studies first perform extensive field work in their proposed research sites. Interviews, observations, or surveys can help the researcher to understand both sides of the hiring process, from the perspective of the job searcher and the employer. Researchers should then use this information to design audit studies that capture the processes that are representative of the on-the-ground conditions faced by employers and job seekers. In doing so, researchers will maximize the external and internal validity of their findings.

Second, researchers must be more attentive to issues of random sampling when conducting audit studies that focus on job opportunities. The population of jobs in a labor market is difficult to calibrate, given that some parts of the labor market are hidden from view (due to jobs that are not advertised). Sampling becomes even more complicated when studying subgroups of a population (i.e., applicants who

have a given level of education) who are likely to apply for some job openings, but not others. As we noted in our discussion above, this issue is not simply a problem of external validity; sampling bias can potentially undermine the internal validity of an experiment as well.

Once again, we recommend that researchers who conduct audit studies in labor markets should plan to conduct exploratory research in their field sites. In practice, rules about sampling jobs should reflect the actual experiences of job seekers. These strategies surely will vary by time and place, and researchers must seek to capture this variation as much as possible in the design of their studies. In doing so, researchers will produce audit studies with greater external and internal validity, and thereby making increasingly valuable contributions to the field.

References

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013.

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2018). The geography of stigma: experimental methods to identify the penalty of place. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Bills, D. (1988). Credentials and capacities: Employers conceptions of the acquisition of skills. *The Sociological Quarterly, 29*, 234–260.

Bills, D. B. (2003). Credentials, signals, and screens: Explaining the relationship between schooling and job assignment. *Review of Educational Research, 73*(4), 441–469.

Cappelli, P. (2012). *Why good people can't get jobs: The skills gap and what companies can do about It*. Philadelphia: Wharton Digital Press.

Carbonaro, W., & Schwarz, J. (2012). *Does where you go matter?: An audit study of high school diplomas and labor market outcomes.* Paper presented at the annual meetings of the American Education Research Association.

Deterding, N. M., & Pedulla, D. S. (2016). Educational authority in the "open door" marketplace: Labor market consequences of for-profit, nonprofit, and fictional educational credentials. *Sociology of Education, 89*(3), 155–177.

Eaton, S. (2001). *The other Boston busing story: What's won and lost across the boundary line*. New Heaven: Yale University Press.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology, 40*(1), 31–53.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. *The Annals of the American Academy of Political and Social Science, 660*(1), 282–299.

Kirschenman, J., & Neckerman, K. (1991). We'd Love to Hire Them, But…': The meaning of race to employers. In C. Jencks & P. E. Peterson (Eds.), *The urban underclass* (pp. 203–232). Washington, DC: The Brookings Institution.

Moss, P., & Tilly, C. (2001). *Stories employers tell: Race, skill, and hiring in America*. New York: Russell Sage Foundation.

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108*(5), 937–975.

Pager, D., & Pedulla, D. S. (2015). Race, self-selection, and the job search process 1. *American Journal of Sociology, 120*(4), 1005–1054.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*(1), 181–209.

Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review, 74*(5), 777–799.

Psacharopoulos, G., & Patrinos, H. A. (2004). Returns to investment in education: A further update. *Education Economics, 12*(2), 111–134.

Quillian, L., Cook, K. S., & Massey, D. S. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology, 32*(1), 299–328.

Rivera, L. A. (2011). Ivies, extracurriculars, and exclusion: Elite employers' use of educational credentials. *Research in Social Stratification and Mobility, 29*(1), 71–90.

Rivera, L. A. (2012). Hiring as cultural matching the case of elite professional service firms. *American Sociological Review, 77*(6), 999–1022.

Rivera, L. A. (2015). Go with your gut: Emotion and evaluation in job interviews. *American Journal of Sociology, 120*(5), 1339–1389.

Rosenbaum, J. (2001). *Beyond college for all: Career paths for the forgotten half*. New York: Russell Sage Foundation.

Weiss, A. (1995). Human capital vs. signalling explanations of wages. *The Journal of Economic Perspectives, 9*(4), 133–154.

Wilson, W. J. (1996). *When work disappears: The world of the new urban poor*. New York: Alfred A. Knopf.

Chapter 8
The Geography of Stigma: Experimental Methods to Identify the Penalty of Place

Max Besbris, Jacob William Faber, Peter Rich, and Patrick Sharkey

Abstract The United States remains a spatially segregated nation by many measures including race, income, wealth, political views, education, and immigration status. Scholars have, for many years, grappled with questions stemming from spatial inequality and have come to recognize the neighborhood in which an individual lives as a socially organizing unit of space, predictive of many individual-level outcomes. The mechanisms that underlie the relationship between neighborhoods and outcomes for residents, however, remain relatively underexplored. In this chapter, we show how the use of audits and field experiments can help uncover one such mechanism—place-based stigma in social interactions. Specifically, we describe the methodology of a previous study (Besbris M, Faber JW, Rich P, Sharkey P, Effect of neighborhood stigma on economic transactions. Proc Nat Acad Sci 112:4994–4998, 2015) that revealed how signaling residence in a poor community of color negatively affected sellers' ability to attract buyers in a classified marketplace. We focus on the study's operationalization of neighborhoods and show how future research can use non-individual-level treatment characteristics such as units of space. Doing so helps us better understand the causal relationship between space and individual-level outcomes, as well as better parse the effects of individual-level variables versus non-individual-level variables, which are often conflated in non-experimental research. We close by suggesting the implementation of field experiments in testing for effects at other geographic scales, such as metropolitan area, state, region, country, or continent.

Keywords Spatial stigma · Experimental design · Socio-spatial inequality

M. Besbris (✉)
Rice University, Houston, TX, USA
e-mail: mb89@rice.edu

J. W. Faber · P. Sharkey
New York University, New York, NY, USA

P. Rich
Cornell University, Ithaca, NY, USA

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_8

8.1 Introduction

Researchers studying the effects of residential environments have become increasingly aware of the limitations of both the methods generally used to study "neighborhood effects" as well as the basic question that motivates much of the literature. *Whether* neighborhoods matter for individual life chances, we argue, is an exceedingly narrow and underspecified question. Instead, research should focus on the ways in which neighborhoods matter, or the questions of when, where, why, for whom, and to what extent are individual or group outcomes affected by their local context (Sharkey and Faber 2014). At the same time, developments in the use of field experiments to estimate the impact of discriminatory behaviors and attitudes have focused primarily on stigma and discrimination at the level of individuals or groups (see Gaddis 2018). Experimental methods have rarely been used to understand the spatial dimensions of inequality, or the geography of stigma.

In this chapter, we show the potential for experimental audits and field experiments to test whether place-based discrimination contributes to spatial foundations of inequality (Galster and Sharkey 2017). Specifically, we draw on a previous study (Besbris et al. 2015) to illustrate how field experiments can operationalize ecological variables. We discuss best practices for signaling various aspects of place—and for designing effective signals in experimental research more broadly.

Our understanding of how neighborhoods affect inequality has largely been limited by methodological constraints (Sampson 2008). Despite advances in quasi-experimental techniques, observational studies are rarely able to provide causal evidence that isolates the effect of residential context on individual outcomes (Cheshire 2012; Ludwig et al. 2008; Mayer and Jencks 1989). Such studies face skepticism that differential outcomes across individuals who reside in different neighborhoods reflect unobserved confounders rather than an actual effect of neighborhood context. This methodological impasse calls for creative alternative approaches capable of producing unbiased tests of neighborhood effects. As we elucidate throughout this chapter, field experiments and audits use randomized conditions within "real world" environments to overcome the problem of selection bias while simultaneously focusing attention on discriminatory social behavior. Because field experiments and audits isolate specific variables and test for their impact on a given outcome (Baldassarri and Abascal 2017; Bertrand and Duflo 2016), they provide a methodological opportunity to move beyond older debates in the research literature on neighborhood effects. However, taking advantage of this opportunity requires a great deal of attention to constructing and testing signals of place which may not be as easily communicated as individual level characteristics like race or gender.

We focus on one particular mechanism through which place may affect individual outcomes: spatial stigma. Spatial stigma refers to the process by which individuals who reside in neighborhoods marked by poverty, crime, and/or racial isolation are thought to be less desirable interactional partners (Besbris 2015; Besbris et al. 2015; see also Goffman 1963; Link and Phelan 2001). We outline how spatial stigma might operate and review non-causal evidence of its existence. We then describe

how field experiments can measure the presence and magnitude of spatial stigma. We summarize the two existing field studies that have examined this phenomenon, focusing on how they operationalize place and the potential problems when in creating signals of place in audits and experiments more broadly. We then conclude by outlining the limitations of experimental methods, reflecting on what the existing experimental studies help us understand about neighborhood effects research more broadly, and proposing future lines of work for experimental and observational studies of place.

8.2 Spatial Stigma

As previously theorized, negative spatial or neighborhood stigma is generated when a particular place has a reputation for crime, disorder, poverty, and/or racial isolation. Such a possible set of characteristics devalues the place in relation to affluent, white, and otherwise more advantaged places. The residents of these differentiated places may come to embody the negative characteristics of their neighborhoods and, as a result, may experience suspicion, mistrust, and undesirability in their interactions with others when their residential origin is revealed (Anderson 2011; Bauder 2002; Wacquant 2008). Similar to other forms of stereotype, the consequences of spatial stigma arise when negative perceptions of a place are attached to individuals, leading to systematic disapproval, discrimination, and/or exclusion (Fiske 1998; Link and Phelan 2001).

For people to act upon spatial stigma and experience its consequences, there must be recognizable variation across geographic areas. In the U.S., residential segregation by race and income produces patchworks of neighborhoods distinct not only in their demographic composition but also in their concentrations of advantage and disadvantage (Logan and Stults 2011; Massey and Denton 1993; Reardon and Bischoff 2011; Wilson 1987). Race and class composition of neighborhoods typically also correlates with other community-level attributes, such as the quality and density of local institutions (e.g. schools, churches, municipal services), commercial activity, job opportunities, environmental conditions, (dis)amenities, property values, and the quality of public life (Clark 1991; Ellen 2000; Harris 1999; Sampson 2012). A long tradition of scholarship on housing preferences has demonstrated that people perceive neighborhoods through the "prism of race" (Krysan and Bader 2007), using racial composition as a direct or indirect measure of neighborhood conditions (Charles 2003; Emerson et al. 2001; Harris 1999; Krysan and Farley 2002). The judgments individuals make about neighborhood quality affects choices about where they live and—importantly—who they live near, contributing to continued residential segregation (Krysan et al. 2014).

If people make place-specific judgements about where to live, they might also make judgements about the people who live in one neighborhood versus another. Assumptions about residents of unfamiliar neighborhoods may be impacted by the fact that segregated neighborhoods reinforce segregated social networks (Sampson

and Sharkey 2008), leading to fewer connections between residents from different communities. Social psychologists have shown that in cases of limited inter-group contact people more often apply generalized stereotypes that ultimately reinforce social distance (Pettigrew 1998; Sigelman and Welch 1993). Thus, segregated metropolitan areas where many people lack nuanced information about communities beyond their immediate surroundings (Bader and Krysan 2015) provide a context for social actors to impose narratives about neighborhoods and the people who live there (Anderson 1999, 2011; Jones and Jackson 2012; Small 2004; Wacquant 2008; Wilson 1987). Moreover, dominant conceptions of black neighborhoods in particular as ghettos may shape how individuals interact with residents from majority black neighborhoods (Anderson 2012, 2015). This dynamic puts residents from racially-isolated, high-poverty neighborhoods at a potential disadvantage when interacting with strangers from outside their community.

Spatial stigma may be more pronounced in areas where crime (and especially violent crime) is concentrated. The geography of crime and race often overlap in U.S. cities, such that white and minority Americans live in what Peterson and Krivo (2012) describe as "divergent social worlds." This may lead to exaggerated perceptions of criminality and danger regarding non-white or poor neighborhoods (Liska et al. 1982; Quillian and Pager 2001; Sampson 2012; Sharkey et al. 2016), which are categorically avoided by outsiders whose fears are stoked by media representations of rampant crime (Chiricos et al. 1997, 2000). These fears, in turn, could dissuade individuals from hiring, dating, educating, or transacting with residents of isolated minority communities, regardless of the resident's individual characteristics. Anderson (2012, 2015) argues that these risks are highest for black Americans who are presumed to live in ghettos.

The disinclination of individuals to interact with those from different neighborhoods may also emerge from a simpler dynamic: geographic proximity. As the scale of segregation grows to broader geographies, such as between cities and places rather than neighborhood blocks (Lichter et al. 2015), any distance penalty added to the perceived cost of an interaction will have a negative impact regardless of intention or any place-based stigma. Growing income segregation (Reardon and Bischoff 2011) also compounds the social distance between individuals of different economic strata with geographic distance.

Geographic boundaries similarly structure the perceived distance between places. Natural boundaries, such as lakes and rivers, as well as constructed boundaries, such as highways and railroad tracks, create physical obstacles that residents of one neighborhood may need to cross or circumvent in order to reach another neighborhood. Bureaucratic and other symbolic boundaries, such as municipal borders, may also increase the social (and financial) costs of interactions across places. In fact, scholars have used natural (Card and Rothstein 2007), constructed (Ananat 2011), and municipal (Cutler and Glaeser 1997) boundaries as instrumental variables for racial segregation because they tend to segment space and separate groups.

In sum, various factors—physical, bureaucratic, symbolic—segregate neighborhoods that are differentiated by demography and exposure to crime. The spatial stigma hypothesis posits that people living in neighborhoods associated with

poverty, crime, and racial isolation may face negative stereotypes and discrimination from strangers when they are forced to reveal their residential location or origin. This may result in lost job opportunities, suspicion by law enforcement, or mistrust in market transactions. The converse may be true as well: residents of affluent and white communities benefit from positive stereotypes, which manifest as more favorable social interactions. Through all of these pathways, the stigma of place may be an important mechanism through which neighborhood segregation reinforces social inequality (Ellen and Turner 1997; Galster 2012; Harding et al. 2011; Jencks and Meyer 1990; Neckerman and Kirschenman 1991; Small and Feldman 2012; Sharkey and Faber 2014). Despite the strong theoretical support for this concept, few previous studies have estimated the effects of neighborhood stigma, in part because it is difficult to disentangle from other forms of disadvantage.

8.3 The Challenges of Measuring Spatial Stigma

Field experiments and audits are critical tools for evaluating the spatial stigma hypothesis. Within specific social settings between strangers—a job application, online dating message, or classified advertisement, for instance—researchers can randomly manipulate a place-based signal while holding all other characteristics constant. They verify whether spatial stigmatization occurs by measuring variation in the rate of favorable responses between place signals. In this sense, field experiments and audits provide researchers with a falsification test of the "null" hypothesis—i.e., that spatial stigma does not occur independently from other forms of discrimination. Field experiments and audits are less well suited to determine precise lower or upper bounds on the effects of spatial stigmatization. And while the use of randomization provides such studies with potentially high internal validity, their conclusions may be constrained to the specific forms of social interaction they test. Nonetheless, because audit studies measure observed actions in real-world situations, they have a distinct advantage over survey methods for measuring discrimination. Specifically, field experiments and audits avoid bias due to social desirability behavior of survey respondents, who may report behavior that they think the researcher (or they themselves) normatively prefer, rather than how they actually behave when they must make trade-offs and experience the consequences of their decisions (Pager and Quillian 2005). Despite this methodological strength, the use of field experiment and audit methods to evaluate the spatial stigma hypothesis presents unique challenges that require careful consideration in design and implementation.

One challenge of measuring spatial stigma in an experimental context is the choice about an appropriate mode of interaction the researchers will control. In an in-person interaction, it is far easier to signal an individual's race or gender than an individual's home address. As such, examining the existence of spatial stigma is extremely difficult via in-person audits. Field experiments that entail correspondence provide a much easier venue because personal letterhead as well as official documents

often contain an individual's address. For example, resumes for jobs, applications for credit cards, and judicial processing documents all usually have an individual's home address or other indicators of their residential location. Interactions in online marketplaces, dating services, and other web-based communities may even require a user to identify where they live or provide such information to others via GPS capabilities. Addresses, zip codes, and neighborhood names, however, are not necessarily strong or clear signals. Residents of a particular city may not know the names and locations of every street or which neighborhoods correspond to which zip codes. Even in correspondence-based field experiments, communicating place of residence is still more difficult than signaling race or gender, which can be done using racially- and gender-identifiable names (Gaddis 2015, 2017a, b).

A second challenge for measuring spatial stigma is the need to effectively capture the local schemas that people use to cognitively map their city. Non-experimental research provides strong support for the claim that residents catalogue and label different parts of the cities in which they live (Anderson 2011, 2012; Hunter 1974; Jones and Jackson 2012; Bader and Krysan 2015; Suttles 1972), yet these cognitive mappings do not necessarily correspond to administrative designations of place such as census tracts, zip codes, or political boundaries. Indeed, employers may have particular reactions to neighborhood names but not street addresses (Wilson 1996:116). So while an application that lists an address may not be screened initially, an applicant may be rejected when they mention their particular neighborhood of residence during a later interview.

The context of the interaction may also activate different ways of interpreting and cataloging space. For example, administrative designations such as school districts may play a role in how people divide space when they are looking to buy a home (Lareau 2014) but they may be less important when a business owner is looking for a storefront to rent. People may also use geographic and physical divides such as railroad tracks, highways, or major thoroughfares as distinct spatial boundaries. Social factors like a local place's average income or racial makeup certainly shape how people define spatial boundaries as well. Furthermore, collective understandings of neighborhood boundaries update over time due to demographic changes (e.g. gentrification), new or demolished housing stock, improvements to public transportation, a shifting geography of crime (or perceptions of crime), and a number of other factors (Ehrenhalt 2012; Hwang 2016). The challenge of effectively signaling place requires researchers to draw on other forms of data when considering how to communicate place; ethnographic and interview data on how individuals map their surroundings should be particularly helpful.

A third and related complication to estimating the effects of spatial stigma is the fact that social phenomena operate at a diverse set of intersecting and overlapping geographies. For example, a police officer may carry geographically-narrow stereotypes of individuals based on the specific blocks on which they reside, which have developed over the course of time spent on a beat. Or an employer may prefer job candidates from one high school catchment area over another—a less granular analysis of space. The relevant spatial unit, therefore, may depend on the phenomena under study (Sharkey and Faber 2014) and assumptions about the local knowledge

of respondents. As a result, experimental studies must have sound reasoning for their selection of local designation. Small pilot studies can often indicate if the target population is recognizing a particular neighborhood signal.

8.4 Experimental Design: Examples from Two Studies

The various challenges in experimental evaluation of spatial stigma require careful considerations of design. The type of behavior analyzed and the local scale of spatial meaning in any given project should inform how analysts make specific methodological decisions. To illustrate important design decisions by example, we discuss how spatial stigma was operationalized in two studies. To our knowledge, these are the only two studies that have ever used field experiments to directly test for the existence of spatial stigma.

The first study tested for spatial stigma by responding to help-wanted advertisements in Chicago and Boston and varying the address on the resume to signal either advantaged or disadvantaged neighborhood of residence (Bertrand and Mullainathan 2004). Across all job applicants, the authors found that living in a whiter, more educated, or higher-income neighborhood increased the likelihood of receiving a call back. Interestingly, they found the same effect across resumes using both identifiably white and black names. In other words, they provide evidence that spatial stigma exists and that it acts similarly for whites and blacks. While the main focus of the study was to identify racial discrimination in the labor market, the addition of a neighborhood signal highlights the ability of field experiments to test for non-demographic sources of discrimination such as spatial stigma.

Betrand and Mullainathan (2004) operationalize neighborhood using contact info on applicant resumes. The researchers randomly assigned fake addresses with real zip codes to every resume, drawing from every possible zip code within Chicago or Boston (p. 996). The authors utilize probit regression models to measure the relative change in likelihood of call-back as the characteristics of zip codes (racial composition, education, and income) changes. One worry, of course, is that employers will not recognize the signal, as zip codes are not necessarily part of individual employers' everyday cognitive schemas of the city (see Wilson 1996). However, the researchers were able to show that the zip code signal was received by employers since it produced differential outcomes. Had the response rate not significantly varied across zip codes, the researchers could have concluded that their operationalization of neighborhood quality was a poor one or that place of residence was not a factor that employers cared about. Yet they did find a difference in response rate, and because field experiments allow for causal claims, they can be certain that the zip code of where an applicant lived mattered for their chances of a call back.

The second study, which we authored and therefore highlight in greater depth here, examined whether advertisements for second-hand iPhones posted from advantaged (i.e. affluent and predominantly white) or disadvantaged (i.e. impoverished and black or Latino) neighborhoods in markets across the U.S. received the

same number of responses from buyers (Besbris et al. 2015). We used an existing online market for second-hand goods, enabling us to gather large amounts of data quickly. We chose 12 markets in large urban areas to reflect the geographic and racial diversity of cities in the U.S. The study found that advertisements signaling disadvantaged neighborhoods received 16 percent fewer responses than those signaling advantaged neighborhoods, providing a strong verification of spatial stigmatization in action.

How were neighborhoods across 12 cities chosen? We began by drawing upon multiple sources of information to name and select advantaged and disadvantaged neighborhoods (see p. 4995). First, census tract-level data on poverty and racial composition were aggregated to boundaries specified by the world's most visited real estate website, Zillow.com. We chose to use a real estate website because it provided more plausible real-world neighborhood boundaries than census tract boundaries, which are often the preferred source in neighborhood effects research (see Sharkey and Faber 2014). While all neighborhood names and boundaries are, in a sense, artificial impositions, we assumed that a popular real estate website's designations were more reflective of individual residents' general understandings than the Census Bureau's and more widely understood than zip codes. Furthermore, Zillow.com provided a systematic tool for naming neighborhoods across all cities in the sample. Second, to confirm our assumptions and verify the names we gathered from Zillow.com, we searched LexisNexis for recent news (including in print and online) that used the neighborhood names we found on Zillow. These searches provided evidence that local media used these neighborhood names, adding strength to our assumption that they might reflect the parlance of local residents. Furthermore, we cross-referenced neighborhood names with the terms "poverty," "homicide," "crime," and "theft" to identify whether local news sources portrayed these neighborhoods as disadvantaged. Finally, we confirmed that the neighborhood names we selected from Zillow would not be unusual or unrecognized by searching for them in the local listings of the online market itself. This strategy of triangulation—using multiple sources including census data, data from Zillow.com, data from local news sources, and data from the market under investigation itself—allowed for a confident assumption that the names we chose were not only recognizable to participants but also identifiable as advantaged or disadvantaged places. Table 8.1 provides a list of neighborhoods and cities from the study.

After selecting neighborhoods associated with advantage and disadvantage for each city in the study, we designed an experiment to isolate the phenomenon of spatial stigmatization in the online marketplace. Specifically, we posted advertisements that randomly signaled a seller's residence using short sentences. To avoid repetitive advertisements, which could produce a negative time effect by conditioning buyers to recognize our posts and ignore them, we generated several versions of advertisement title, text language, and price that were assigned randomly. Table 8.2 provides examples of actual posts. Importantly, the signal of neighborhood origin was added to the end of the text, but was also accompanied by information about desired meeting location. We randomly varied two types of meeting location—willing to meet in buyer's neighborhood, or willing to meet at a central place—to

Table 8.1 Neighborhoods selected for the field experiment in Besbris et al. 2015

City	Neighborhood	Classification	Poverty rate	Selected racial composition
Atlanta	Midtown	Advantaged	9.1%	70.2% white
	Oakland City	Disadvantaged black	35.4%	87.5% black
Baltimore	Canton	Advantaged	11.8%	75.4% white
	West Baltimore	Disadvantaged black	37.9%	83.7% black
Boston	Back Bay	Advantaged	9.7%	86.0% white
	Dorchester	Disadvantaged black	18.8%	45.8% black
Chicago	Lincoln Park	Advantaged	11.6%	82.5% white
	North Lawndale	Disadvantaged black	41.8%	91.7% black
Los Angeles	Century City	Advantaged	9.7%	76.8% white
	Crenshaw	Disadvantaged black	25.3%	68.9% black
NY Brooklyn	Cobble Hill	Advantaged	4.3%	71.2% white
	Bedford-Stuyvesant	Disadvantaged black	29.6%	77.3% black
NY Manhattan	Upper East Side	Advantaged	6.0%	81.2% white
	East Harlem	Disadvantaged Latino	35.5%	56.6% Latino
Philadelphia	Fox Chase	Advantaged	8.9%	78.9% white
	Nicetown	Disadvantaged black	32.2%	93.8% black
	Juniata	Disadvantaged Latino	39.3%	52.1% Latino
Phoenix	Ahwatukee Foothills	Advantaged	6.1%	73.3% white
	Central City	Disadvantaged Latino	44.2%	64.4% Latino
San Antonio	North Central	Advantaged	3.8%	74.0% white
	Southwest San Antonio	Disadvantaged Latino	38.8%	92.2% Latino
Seattle	Madrona	Advantaged	4.4%	74.8% white
	Leschi	Disadvantaged black	18.1%	36.2% black
	International District	Disadvantaged Asian	43.1%	49.0% Asian
Washington DC	Dupont Circle	Advantaged	11.1%	73.6% white
	Anacostia	Disadvantaged black	31.6%	97.1% black

Neighborhood boundaries derived from Zillow.com. Tract-level data aggregated to neighborhood using 2007–2011 American Community Survey data

Table 8.2 Examples of advertisements varying reference to seller neighborhood from Besbris et al. 2015

Advantaged neighborhood	Disadvantaged neighborhood
City: Atlanta	*City:* Atlanta
Price: $265	*Price:* $265
Heading: 16GB IPHONE 5 - ATT - BLACK - LIKE NEW!	*Heading:* 16GB IPHONE 5 - ATT - BLACK - LIKE NEW!
Advertisement text: 4 month old black iPhone 5 for sale. Includes original box, headphones, and charger. Perfect condition, no scratches. I live in Midtown and can meet downtown.	*Advertisement text:* 4 month old black iPhone 5 for sale. Includes original box, headphones, and charger. Perfect condition, no scratches. I live in Oakland City and can meet downtown.
City: Boston	*City:* Boston
Price: $405	*Price:* $405
Heading: AT&T Black iPhone 5 (16G)	*Heading:* AT&T Black iPhone 5 (16G)
Advertisement text: Like new iPhone 5 for sale - just a few months old. Comes with box and all items that were in the box. No scrapes or dents. I'm in Back Bay. Meet in your neighborhood.	*Advertisement text:* Like new iPhone 5 for sale - just a few months old. Comes with box and all items that were in the box. No scrapes or dents. I'm in Dorchester. Meet in your neighborhood.
City: Chicago	*City:* Chicago
Price: $320	*Price:* $320
Heading: iPhone black 5 16G (AT&T)	*Heading:* iPhone black 5 16G (AT&T)
Advertisement text: If you want a good deal on a basically new (no scratches, dents, etc.) iphone 5, this is it. You'll get all the things that were in the original box. Meet in the loop. I live in North Lawndale.	*Advertisement text:* If you want a good deal on a basically new (no scratches, dents, etc.) iphone 5, this is it. You'll get all the things that were in the original box. Meet in the Loop. I live in Lincoln Park.

Advertisements included randomly selected versions of heading, price, and advertisement text, and suggested meeting location (central location or in buyer's neighborhood). Advantaged and disadvantaged neighborhoods of the seller were assigned randomly according to Table 8.1, and the central location was specific to each city (for example: "the Loop" in Chicago). Prices were determined based on the median advertised price in each live market, updated each month of the field experiment. For additional information, see the Supporting Information of Besbris et al. 2015 (http://www.pnas.org/content/112/16/4994.full.pdf?with-ds=yes)

address the concern that response rates captured proximity and convenience concerns (or higher concentrations of buyers in some areas than others)—although, as we discuss above, geographic distance may be an important factor in shaping social distance and the perceived cost of cross-neighborhood interaction.

We did consider that our reference to specific places could be viewed as artificial. This could potentially suppress the number of responses, although any such effect would need to be correlated with the neighborhood preferences of respondents in order to bias our results. Specifically, without such a confounding relationship, it would not affect the difference in the average number of responses between advantaged and disadvantaged neighborhoods because both used identical versions of syntax in the advertisements. Nonetheless, to ensure that our posts would not be systematically ignored by real buyers, we ran a pilot to test our method and sampled

advertisements posted by others (i.e. actual sellers) from each of the 12 local markets. Nearly 60 percent of the advertisements we sampled indicated location. As such, we were confident that the method we used to signal neighborhoods was typical of other actors in that online community and not artificial. Again, brief pilots can provide researchers with some indication of how place and other characteristics are normally signaled in a particular context and if the signals to be used in the experiment are being received by the target group.

Field experiments and audits in online (and brick and mortar) marketplaces face a challenge unique to the setting: conditions of supply and demand may change rapidly due to external forces. For example, the release of a new version of a particular technology can dramatically reduce the desirability of the previous version. Indeed, we observed that the secondary market for the iPhone 5 declined over the course of our study both as measured by the number of other advertisements in each local market as well as the prices listed by those sellers. We addressed this time trend by including controls for market conditions and by adjusting the advertised price each month according to the median price of all other advertisements. In addition, we included statistical controls for time when evaluating the results. Nonetheless, it is possible that the nature of bias can change over time, as buyers and sellers adjust their behaviors in a dynamic marketplace. It would be difficult to detect if, for example, the magnitude of neighborhood stigma declined over time as buyers face dwindling supply of a particular good. As such, researchers may have to continually assess which aspects of a field experiment can be altered during data collection without changing the intended signal or compromising marketplace behavior.

To conclude, in both Bertrand and Mullainathan (2004) and Besbris et al. (2015), field experiments tested and found evidence for the existence of spatial stigma. Because this particular research design can isolate place of residence from other factors that correlate with it (e.g., education, race, income, proximity, or some other unobserved covariate), both studies provide valid causal evidence of a social mechanism through which neighborhoods matter in two economically consequential activities: the search for jobs and the sale of goods in a classified marketplace.

8.5 Limitations

Although online audits and field experiments have a number of beneficial qualities, there are limitations to these approaches. Most importantly, perhaps, is the broad challenge of capturing the complexity of the social world and, in particular, the ways in which multiple aspects of society often work in concert to shape outcomes. For example, although Besbris et al. (2015) identified a negative effect of mentioning a poor, black neighborhood in an online marketplace for used smartphones, the specific cause of this effect is unclear. Potential buyers in that market may be assuming that the seller from a poor, black neighborhood is a poor, black person, or simply a black person, or simply poor. The signal may also be priming concerns about

criminality, but this cannot be directly tested in the study. In such a spatially stratified society, specific individual and ecological characteristics are often bound together, which makes it difficult to interpret how the behavior of respondents in online audits and field experiments maps on to theories of disadvantage and bias. Therefore, researchers must be extremely careful in the kinds of claims they make based on experimental field studies that measure different outcomes across non-demographic characteristics. Increasingly, experimental research signals multiple variables (e.g. both race *and* gender) for an intersectional understanding of how various aspects of social life may work in concert to stigmatize individuals (for review, see Baldessarri and Abascal 2017). Research on spatial stigma should follow and use signals of place in combination with different demographic or non-demographic variables to better evaluate if place itself contributes to various outcomes or if it is simply acting as a proxy for other potential variables (see below).

Most social interactions are multi-stage, and field experiments typically only address one step—often an initial one—in the process. A decision to hire an employee, for example, may involve the review of a resume, followed by a phone call, and concluded by an in-person interview and a check of references. Typically, due to resource constraints and the protection of human research subjects, online audits are limited to the study of one stage in an interaction and, perhaps, the least consequential stage. Resume-based field experiments (Bertrand and Mullainathan 2004; Gaddis 2015; Pedulla 2016) are useful in understanding discrimination early on in the hiring process, but are incapable of fully capturing how biases contribute to employment disparities. Internet-based field experiments in the housing market (Ahmed and Hammarstedt 2008; Hanson and Hawley 2011; Hogan and Berry 2011) suffer from a similar limitation in that a listing agent or landlord may agree to show a house to a minority homeseeker with no intention of actually renting to her. Furthermore, experimentally testing the presence of bias at later stages of either of these processes (e.g. as part of a background or credit check) would be impossible.

This limitation is especially important to note in the context of investigating neighborhood stigma. Not only are field experiments limited in their ability to capture the cumulative effect of place of residence in a given interaction (i.e. neighborhood of residence may not matter in the initial job application but does become a relevant signal at a later screening stage), but there is increasing evidence that neighborhood effects compound over time (Chetty et al. 2016; Sharkey and Elwert 2011; Wodtke et al. 2011). As a result, the *cumulative* effect of spatial stigma is unknowable, even when using more precise measures like field experiments.

A related limitation in experimental studies of social stigma stems from the fact that multiple mechanisms may simultaneously affect perceptions in social interactions. Ideally, one could test stigmatization across neighborhoods that vary by one specific trait—such as crime—but have the same poverty levels and racial makeup. However, because crime, poverty, and racial/ethnic groups are so highly segregated in many American cities, and because there is such a strong class gradient across racial/ethnic groups, it is nearly impossible to compare neighborhoods that are only different on one axis. For example, geographically concentrated white poverty does not exist in many American cities. Furthermore, the poorest, predominantly white

area within a city is often not nearly as poor or racially homogenous as communities of color in that same city. This is why, in our previous work, "disadvantaged neighborhoods" are all non-white (Besbris et al. 2015).

An online experiment may also alter the community in which it takes place. For example, flooding a small market with fake advertisements may reduce trust in the market among those who are using it for its intended purpose. Similarly, posing as a potential employee or romantic partner may exact substantial costs—time, emotional and psychological commitment, etc.—on those who are evaluating potential matches. Spillover effects of online studies must be considered as a potential violation of the Stable Unit Treatment Value Assumption (SUTVA) (Morgan and Winship 2007) and efforts must be taken to avoid harming the integrity of the community and its members. In Besbris et al. (2015), for instance, we limited the frequency of advertisements to twice per week, and utilized a generalized randomization technique—rather than a matched-pair audit study—that constrained the number of observations collected and the amount of statistical power available to address nuanced research questions, such as heterogeneity in spatial stigmatization across different cities.

Another limitation stems from the fact that it is often difficult, if not impossible, to gather data on the respondent population in online studies. Many online interactions are characterized by anonymity and in studies in which non-response can be just as informative as response (e.g. Besbris et al. 2015), there is no way to know which users did not initiate communication. Furthermore, conducting post-audit surveys of respondents likely increases the risk of contaminating the market in ways discussed above. Yet without data on participants, it is impossible to verify whether inferences drawn from online audits and field experiments reflect broader population dynamics or, instead, behavior specific to the members of the online community studied since market participants may not be representative of the populations of neighborhoods, cities, or any larger geography. While not a threat to internal validity of such studies, this external validity concern is substantial. More broadly, the lack of post-hoc data in many field experiments and audits limits they types of conclusions that can be drawn—especially if the findings are null. If we had found no difference in response rate between ads from advantaged or disadvantaged neighborhoods, we would not know if the results were due to a poor operationalization of our treatment or if the variable itself did not matter for how buyers made their choices. This risk can be mitigated with the type of triangulation we performed when selecting our neighborhoods as well as with pilot phases of experiments and post-hoc data collection (e.g., interviews with participants).

Even separating geographic proximity from other potential mechanisms of stigma is difficult, as poor black and Latino areas are typically not near affluent white communities. We previously tried to address this potential issue by signaling willingness to travel to respondents' neighborhoods or a central location and by choosing neighborhoods that were relatively close to one another, a downtown area, or transit hub. However, the extent to which individuals are disinclined to interact with others who do not live nearby is itself a cause and consequence of segregation, so it is theoretically unclear whether distance is a confounding variable or a causal pathway.

8.6 Three Areas for Future Research

Audit studies and field experiments are uniquely positioned to advance understanding of how different dimensions of advantage and disadvantage, at the level of individuals and places, can influence social and economic interactions and outcomes. More specifically, we believe that field experiments focusing on the impact of places have substantial potential to generate new insights into the spatial dimensions of inequality and the mechanisms underlying neighborhood effects. Building on the research described in this chapter, we have identified three areas of research that are crucial for moving toward a more complete understanding of spatial stigma.

First, field experiments carried out in different geographic settings and focusing on different forms of interactions are essential to understanding where and when spatial stigma may become salient. In our own study, we examined the effect of neighborhood disadvantage in online markets for interpersonal exchanges of smartphones, but the particular conditions of the market for this item almost certainly influence the potential impact of spatial stigma. Does place of residence matter for exchanges that do not involve personal, face-to-face interaction? If place of residence matters for interpersonal economic exchanges and job applications, does it matter for college admissions or promotion within a given firm?

The close connection between place of residence and race suggests that spatial stigma may exist in situations where race has been shown to affect outcomes, such as romantic partnerships (Robnett and Feliciano 2011; Torche and Rich 2017) or assessments by market intermediaries like real estate agents (Besbris 2016; Besbris and Faber 2017; Yinger 1995) or mortgage lenders (Faber 2013). Testing for the impact of spatial stigma in these and other sites of stratification is necessary to develop a broader theory of how residential context can advantage or disadvantage individual residents and understand when, where, and for whom place of residence acts as a filtering and sorting heuristic. Using field experiments to test for spatial stigma across situations and interactions therefore fills both theoretical and empirical gaps.

Second, and related to the previous point, study designs can be developed to assess the interactions between individual (or group) disadvantage and spatial disadvantage, and to attempt to disentangle the relative influence of each. While a number of studies have explored intersectionality in specific arenas, for example, by testing for both race effects and sexuality effects in the market for jobs (Pedulla 2014) or both parental status and sexuality in the market for housing (Lauster and Easterbrook 2011), minimal research has considered the interaction of individual disadvantage and spatial disadvantage.[1] As noted previously, signaling place of residence along with other variables can help isolate the effects of the place itself. This is especially needed in the study of spatial stigma because place of residence is

[1] Although Bertrand and Mullainathan (2004) found spatial stigma operating similarly across black and white job applicants, it is possible that spatial stigma may produce different results across races in other areas of social life like mate selection.

so tightly linked to socio-economic status, race, education, and a host of other characteristics that may also affect decisions across situations.

Three of the authors are conducting an additional study designed to disentangle the impact of spatial disadvantage from individual race and ethnicity by responding to advertisements for smartphones while signaling both race/ethnicity and residential location. This design will help determine whether the impact of neighborhood disadvantage is partially or fully explained by assumptions made about the race/ethnicity of the individual taking part in the transaction.

Examining multiple dimensions of stigma in the same study also allows for tests of interaction effects. The implicit and explicit associations that individuals make regarding delinquency, intelligence, sexual proclivity, and other behavioral traits are rarely race- or gender-neutral, so certain race and gender combinations may accentuate neighborhood stigma, while others may moderate it. For example, does the negative effect of the stigma associated with a poor, black, and high-crime neighborhood operate for white women? Conversely, do black men garner the same interactional benefit from residing in an affluent white community? These and related questions will shed light on the nature of racial and gender inequality in the context of a highly segregated society.

Third, the effect of spatial stigma must be examined at multiple levels of analysis. Although our examples in this chapter focused on relatively small geographies (i.e. neighborhoods), stigma may operate at other spatial units in ways that create advantage or disadvantage. Country of origin, while often conflated with race, religion, and language, may also communicate cultural affinity, wealth, or political leaning in particular interactions. Given current public debates in the United States regarding immigration from Latin America and the Middle East, nationality bias may be particularly strong. Americans may also carry stereotypes about individuals from different regions within the country, which manifest as interactional bias—particularly in combination with race. And within metropolitan areas, people living in urban areas may be perceived differently than people living in suburbs. Affiliation with even smaller locations within individual cities, such as specific public housing projects, or, conversely, luxury residential developments, may also influence interactions with the police, teachers, employers, or potential romantic partners.

Understanding the various, context-dependent roles of these layered geographies, as well as the ways in which they interact with race, gender, age, and other characteristics is a challenging task. In addition to audits and field experiments, the theory and suggestions outlined in this chapter can be extended to other methodological approaches. Qualitative work is particularly well positioned to investigate the role (or roles) played by spatial stigma at the interactional level. In combination with experimental approaches, ethnography and interviews can help elaborate interactional mechanisms—such as spatial stigma—shaping the geography of inequality.

References

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the internet. *Journal of Urban Economics, 64*, 362–372.

Ananat, E. O. (2011). The wrong side of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics, 3*, 34–66.

Anderson, E. (1999). *Code of the street: Decency, violence, and the moral life of the inner city*. New York: Norton.

Anderson, E. (2011). *The cosmopolitan canopy: Race and civility in everyday life*. New York: Norton.

Anderson, E. (2012). The iconic ghetto. *The Annals of the American Academy of Political and Social Sciences, 642*, 8–24.

Anderson, E. (2015). The white space. *Sociology of Race and Ethnicity, 1*, 10–21.

Bader, M. D., & Krysan, M. (2015). Community attraction and avoidance in Chicago: What's race got to do with it? *The Annals of the American Academy of Political and Social Science, 660*, 261–281.

Baldassarri, D., & Abascal, M. (2017). Field experiments in the social sciences. *Annual Review of Sociology, 43*. https://doi.org/10.1146/annurev-soc-073014-112445.

Bauder, H. (2002). Neighborhood effects and cultural exclusion. *Urban Studies, 39*, 89–93.

Bertrand, M., & Duflo, E. (2016). *Field experiments on discrimination* (NBER Working Paper #22014).

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991–1013.

Besbris, M. (2015). Stigma. In F. F. Wherry & J. B. Schor (Eds.), *The Sage encyclopedia of economics and society* (pp. 1532–1534). Thousand Oaks: Sage.

Besbris, M. (2016). Romancing the home: Emotions and the interactional creation of demand in the housing market. *Socio-Economic Review, 14*, 461–482.

Besbris, M., & Faber, J. W. (2017). Investigating the relationship between real estate agents, segregation, and house prices: Steering and upselling in New York State. *Sociological Forum, 32*. https://doi.org/10.1111/socf.12378.

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2015). Effect of neighborhood stigma on economic transactions. *Proceedings of the National Academy of Science, 112*, 4994–4998.

Card, D., & Rothstein, J. (2007). Racial segregation and the black-white test score gap. *Journal of Public Economics, 91*, 2158–2184.

Charles, C. Z. (2003). *Won't you be my neighbor? Race, class, and residence in Los Angeles*. New York: Russell Sage Foundation.

Cheshire, P. (2012). Are mixed community policies evidence based? A review of research on neighbourhood effects. In M. van Ham, D. Manley, N. Baily, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 267–294). New York: Springer.

Chetty, R., Hendren, N., & Katz, L. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review, 106*, 855–902.

Chiricos, T., Hogan, M., & Gertz, M. (1997). Racial composition of neighborhood and fear of crime. *Criminology, 35*, 107–132.

Chiricos, T., Padgett, K., & Gertz, M. (2000). Fear, TV news, and the reality of crime. *Criminology, 38*, 755–786.

Clark, W. A. V. (1991). Residential preferences and neighborhood racial segregation: A test of the Schelling segregation model. *Demography, 28*(1), 19.

Cutler, D. M., & Glaeser, E. L. (1997). Are ghettos good or bad? *The Quarterly Journal of Economics, 112*, 827–872.

Ehrenhalt, A. (2012). *The great inversion and the future of the American City*. New York: Knopf Doubleday Publishing Group.

Ellen, I. G. (2000). *Sharing America's neighborhoods*. Cambridge: Harvard University Press.
Ellen, I. G., & Turner, M. A. (1997). Does neighborhood matter? Assessing recent evidence. *Housing Policy Debate, 8*, 833–866.
Emerson, M. O., Chai, K. J., & Yancey, G. (2001). Does race matter in residential segregation? Exploring the preferences of white Americans. *American Sociological Review, 66*, 922–935.
Faber, J. W. (2013). Racial dynamics of subprime mortgage lending at the peak. *Housing Policy Debate, 23*, 328–349.
Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In L. Gardner, D. Gilbert, & S. T. Fiske (Eds.), *The handbook of social psychology* (pp. 357–411). Oxford: Oxford University Press.
Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*, 1451–1479.
Gaddis, S. M. (2017a). How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*, 469–489.
Gaddis, S. M. (2017b). Racial/ethnic perceptions from Hispanic names: Selecting names to test for discrimination. *Socius, 3*, 1–11.
Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.
Galster, G. C. (2012). The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications. In M. van Ham, D. Manley, N. Baily, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 23–56). New York: Springer.
Galster, G. C., & Sharkey, P. (2017). Spatial foundations of inequality: A conceptual model and empirical overview. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 3*(2), 1–33.
Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs: Prentice Hall.
Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in U.S. cities. *Journal of Urban Economics, 70*, 99–114.
Harding, D., Gennetian, L., Winship, C., Sanbonmatsu, L., & Kling, J. (2011). Unpacking neighborhood influences on education outcomes: Setting the stage for future research. In G. Duncan & R. Murnane (Eds.), *Whither opportunity? Rising inequality, schools and children's life chances Russell* (pp. 277–296). New York: Sage.
Harris, D. R. (1999). 'Property values drop when black move in, because…' racial and socioeconomic determinants of neighborhood desirability. *American Sociological Review, 64*, 461–479.
Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & Community, 10*, 351–372.
Hunter, A. (1974). *Symbolic communities*. Chicago: University of Chicago Press.
Hwang, J. (2016). The social construction of a gentrifying neighborhood: Reifying and redefining identity and boundaries in inequality. *Urban Affairs Review, 52*, 98–128.
Jones, N., & Jackson, C. (2012). 'You just don't go down there': Learning to avoid the ghetto in San Francisco. In R. Hutchinson & B. D. Haynes (Eds.), *The Ghetto: Contemporary global issues and controversies* (pp. 83–110). Boulder: Westview Press.
Jencks, C., & Mayer, S. (1990). The social consequences of growing up in a poor neighborhood. In L. Lynn & M. McGeary (Eds.), *Inner city poverty in the United States* (pp. 111–186). Washington, DC: National Academies Press.
Krysan, M., & Bader, M. (2007). Perceiving the metropolis: Seeing the city through the prism of race. *Social Forces, 86*, 699–733.
Krysan, M., & Farley, R. (2002). The residential preferences of blacks: Do they explain persistent segregation? *Social Forces, 80*, 937–980.
Krysan, M., Crowder, K., & Bader, M. D. M. (2014). Pathways to residential segregation. In A. Lareau & K. Goyette (Eds.), *Choosing homes, choosing schools* (pp. 27–63). New York: Russell Sage.

Lareau, A. (2014). Schools, housing, and the reproduction of inequality. In A. Lareau & K. Goyette (Eds.), *Choosing homes, choosing schools. Russell* (pp. 169–206). New York: Sage.
Lauster, N., & Easterbrook, A. (2011). No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. *Social Problems, 58*, 389–309.
Lichter, D. T., Parisi, D., & Taquino, M. C. (2015). Toward a new macro-segregation? Decomposing segregation within and between metropolitan cities and suburbs. *American Sociological Review, 80*, 843–873.
Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology, 27*, 363–385.
Liska, A. E., Lawrence, J. J., & Sanchirico, A. (1982). Fear of crime as a social fact. *Social Forces, 60*, 760–770.
Logan, J. R., & Stults, B. (2011). *The persistence of segregation in the metropolis: New findings from the 2010 census* (Census Brief prepared for Project US2010). http://www.s4.brown.edu/us2010 1–25
Ludwig, J., Liebman, J. B., Kling, J. R., Duncan, G. J., Katz, L. F., Kessler, R. C., & Sanbonmatsu, L. (2008). What can we learn about neighborhood effects from the moving to opportunity experiment? *American Journal of Sociology, 114*, 144–188.
Massey, D. S., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Cambridge: Harvard University Press.
Mayer, S. E., & Jencks, C. (1989). Growing up in poor neighborhoods: How much does it matter? *Science, 243*, 1441–1445.
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal evidence*. New York: Cambridge University Press.
Neckerman, K. M., & Kirschenman, J. (1991). Hiring strategies, racial bias, and inner-city workers. *Social Problems, 38*, 433–447.
Pager, D., & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review, 70*, 355–380.
Pedulla, D. (2014). The positive consequences of negative stereotypes: Race, sexual orientation, and the job application process. *Social Psychology Quarterly, 77*, 75–94.
Pedulla, D. (2016). Penalized or protected?: Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*, 262–289.
Peterson, R. D., & Krivo, L. J. (2012). *Divergent social worlds*. New York: Russell Sage Foundation.
Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.
Quillian, L., & Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology, 107*, 717–767.
Reardon, S. F., & Bischoff, K. (2011). Income inequality and income segregation. *American Journal of Sociology, 116*, 1092–1153.
Robnett, B., & Feliciano, C. (2011). Patterns of racial-ethnic exclusion by internet daters. *Social Forces, 89*, 807–828.
Sampson, R. J. (2008). Moving to inequality: Neighborhood effects and experiments meet social structure. *American Journal of Sociology, 114*, 189–231.
Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect*. Chicago: University of Chicago Press.
Sampson, R. J., & Sharkey, P. (2008). Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography, 45*, 1–29.
Sharkey, P., & Elwert, F. (2011). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology, 116*, 1934–1981.
Sharkey, P., & Faber, J. W. (2014). Where, when, why, and for whom do residential contexts matter? Moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology, 40*, 559–579.

Sharkey, P., Besbris, M., & Friedson, M. (2016). Poverty and crime. In D. Brady & L. M. Burton (Eds.), *The Oxford handbook of the social science of poverty* (pp. 623–636). Oxford: Oxford University Press.

Sigelman, L., & Welch, S. (1993). The contact hypothesis revisited: Black-white interaction and positive racial attitudes. *Social Forces, 94*, 781–795.

Small, M. L. (2004). *Villa Victoria: The transformation of social capital in a Boston barrio*. Chicago: University of Chicago Press.

Small, M. L., & Feldman, J. (2012). Ethnographic evidence, heterogeneity, and neighbourhood effects after moving to opportunity. In M. van Ham, D. Manley, N. Baily, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 57–78). New York: Springer.

Suttles, G. D. (1972). *The social construction of communities*. Chicago: University of Chicago Press.

Torche, F., & Rich, P. (2017). Declining racial stratification in marriage choices? Trends in black/white status exchange in the United States, 1980 to 2010. *Sociology of Race and Ethnicity, 3*, 31–49.

Wacquant, L. J. D. (2008). *Urban outcasts: A comparative sociology of advanced marginality*. New York: Wiley.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago: University of Chicago Press.

Wilson, W. J. (1996). *When work disappears: The world of the new urban poor*. New York: Knopf.

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of longterm exposure to concentrated disadvantage on high school graduation. *American Sociological Review, 76*, 713–736.

Yinger, J. (1995). *Closed doors, opportunities lost*. New York: Russell Sage.

Chapter 9
Emerging Frontiers in Audit Study Research: Mechanisms, Variation, and Representativeness

David S. Pedulla

Abstract Audit studies have gained popularity in the social sciences, producing important insights about discrimination and bias across a range of social statuses, such as race and gender. Yet, important questions persist about *why*, *when*, and *where* discrimination and bias emerge. In this chapter, I suggest that tackling these issues is a central task of audit studies and discuss emerging frontiers of audit study research that are attempting to address these pressing issues. First, audit studies can contribute to our understanding of *why* discrimination occurs by incorporating strategies to uncover the mechanisms that drive the empirical patterns observed in the data. Second, audit studies can provide insights about *when* and *where* discrimination and bias occur by paying attention to theoretically important variation in average treatment effects and clarifying the representativeness of a given set of findings. Throughout, I present evidence from recent audit study research that pushes the boundaries on each of these frontiers and discuss potential paths forward to continue to advance the design, implementation, and contribution of this method to social science research.

Keywords Audit studies · Mechanisms · Research design · Representativeness

Audit studies have gained popularity in the social sciences over the past decades, particularly among researchers examining discrimination and bias. In part, the growth in the use of this method stems from one of its core strengths: enabling scholars to combine high levels of internal validity – causally linking an independent variable with a dependent variable – with the examination of "real world" behavior. Audit studies have been used to investigate discrimination in multiple institutional domains, including housing markets (Lauster and Easterbrook 2011), lending

I thank Maria Abascal, Jacob Avery, Mike Bader, Alex Murphy, S. Michael Gaddis, and an anonymous reviewer for comments on an earlier version of this chapter. The usual disclaimer applies.

D. S. Pedulla (✉)
Department of Sociology, Stanford University, Stanford, CA, USA
e-mail: dpedulla@stanford.edu

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9_9

markets (Ross et al. 2008), online classified markets (Besbris et al. 2015), and labor markets (Pager et al. 2009a).[1] This methodological approach has also been used to investigate disparate treatment along multiple axes of social experience, including race (Bertrand and Mullainathan 2004), gender (Neumark 1996), sexual orientation (Mishel 2016), age (Lahey 2008), employment history (Pedulla 2016), having a criminal record (Pager 2003), educational background (Deming et al. 2016; Deterding and Pedulla 2016), class background (Jackson 2009; Rivera and Tilcsik 2016), and a host of other social positions (see Gaddis 2018 for a review).

By documenting the often biased behaviors of employers, landlords, and other actors, audit studies have contributed important knowledge about the persistent barriers faced by members of different social groups. Yet, the focus of many audit studies on establishing *whether* there is a direct effect of a given social status on a particular outcome has meant that many (although certainly not all) audit studies leave unanswered questions about the *why*, *where*, and *when* of discrimination and bias. However, these issues are of central theoretical interest to scholars across the social sciences. Designing audit studies that are better able to shed light on why discrimination occurs, where discrimination occurs, and when discrimination occurs will help to ensure that this powerful methodological tool continues to build social scientific knowledge.

While not exhaustive, this chapter will examine three frontiers of audit study research that can assist in answering questions about why, where, and when discrimination and bias take place by focusing on: (1) mechanisms, (2) variation, and (3) representativeness. Before moving forward, I offer a few caveats. While the audit study literature examines multiple institutional contexts, such as housing and lending markets, this chapter will focus on the labor market. However, many of the issues addressed in the labor market context will be applicable to other domains, which I will briefly discuss toward the end of this chapter. Additionally, audit studies of the labor market have been conducted all over the world. While I will draw on some international examples, much of the scholarship discussed below is drawn from research conducted in the United States. I will begin by discussing the issues related to mechanisms and then move on to address variation and representativeness.

9.1 Uncovering the Mechanisms of Discrimination and Bias

A clear benefit of audit studies is that the results are characterized by high levels of internal validity, enabling researchers to make causal claims. The logic of audit studies, discussed in more detail in other parts of this volume, is relatively straightforward. In the labor market, for example, the researcher sends nearly

[1] Audit studies need not only examine treatment in markets. For example, Milkman et al. (2012) use an audit study to examine faculty members' responses to prospective doctoral students that varied in their race, ethnicity, and gender.

identical, fictitious job applications to a set of job openings.[2] Everything is held constant about the applications, with the exception of the social status or characteristic of interest (e.g., race, gender). By randomly assigning which application receives the "treatment" characteristic and which one does not, iterating over a broad set of job postings, and tracking employers' responses (e.g., "callbacks") to each application, the researcher is then able to estimate the differential treatment of job applicants based on the characteristic of interest. The differences in callback rates between the treatment and control group are then interpreted as a measure of discrimination or bias.

While a powerful technique for estimating causal effects, audit studies often have difficulty identifying the mechanisms that drive the empirical patterns that are observed.[3] In other words, audit studies often leave outstanding questions about why discrimination or bias occurred. This can leave a "black box" of speculation between the independent and dependent variables. For example, if an audit study finds that African Americans face discrimination during the hiring process, it could be due to stereotypes that lead them to be perceived as less competent than white applicants, that they are perceived as less of a "fit" with the organization than white applicants, or some other mechanism altogether. In other words, audit studies of the hiring process are well equipped to establish that certain social statuses or social characteristics advantage or disadvantage job applicants. However, audit studies are less often designed to shed theoretical light on the question of *why* those types of disparities occur.

While there are many reasons that we may care about mechanisms, including the development of mid-range theory, mechanisms are also important in the design of remedies to address bias and discrimination. If African Americans face discrimination due to employers' perceptions of their competence, for example, interventions to reduce discrimination would look different than if employers were less likely to hire black workers because they were not seen as a good "fit" for the company.

Below, I outline three ways that recent scholarship has attempted to address the challenge of why discrimination occurs by identifying the mechanisms that are at work: (1) measuring or manipulating mechanisms within the audit study, (2) combining audit studies with lab or survey experiments, and (3) collecting supplemental qualitative data.

9.1.1 Measuring or Manipulating Mechanisms

In thinking about opening up the "black box" of mechanisms in audit study research – answering the *why* question – one path forward is for researchers to measure or experimentally manipulate key theoretical mechanisms. This approach

[2] This can be done via electronic or paper applications as well as with in-person actors. This distinction will be discussed in more detail, below.

[3] When I refer to mechanisms in this context, I am referring to the theoretical construct or constructs that connect an independent and dependent variable with one another.

requires additional work for the researcher – either adding experimental cells to the research design or finding and merging additional contextual data with the audit study results. However, this strategy can prove valuable in thinking about the underlying processes driving outcomes in an audit study of labor market discrimination.

To illustrate this point, let us explore the hypothetical example mentioned above: racial discrimination against African Americans in the labor market. Imagine that one's theoretical argument is that African American job applicants are discriminated against because they are stereotyped by employers as less competent than white job applicants (Moss and Tilly 2001). A researcher could test for actual discrimination in the labor market using audit study methods, sending fictitious applications to real jobs and varying the race of each applicant. As existing research has shown, this approach is likely to lead to evidence of discrimination against black job applicants (Bertrand and Mullainathan 2004; Pager et al. 2009a).

How might we test for the proposed competence mechanism? One approach is to try to gain traction on this mechanism by actually measuring variation in the demand for competence among different employers. This strategy could involve, for example, coding each of the job postings in the audit study for employers' requests for high levels of competence or language related to the competence construct. If the "competence stereotype" mechanism is correct, then one might hypothesize that racial discrimination would be more severe among the job postings that emphasized competence.

An example of this approach – measuring the mechanism – can be found in Tilcsik's (2011) audit study of discrimination against gay men in the labor market. He argues that a key mechanism that may drive labor market discrimination against gay men is that they are stereotyped as effeminate. To test this argument, Tilcsik (2011) coded the job postings in the audit study for language that emphasized stereotypically male, heterosexual traits. He then used moderation analysis in a regression framework to examine whether discrimination against gay men varied by whether employers emphasized these stereotypical heterosexual and masculine attributes. Indeed, that is what he found. Gay men faced relatively stronger discrimination among this set of employers (Tilcsik 2011). Thus, these analyses provide useful information that stereotypes about gay men as effeminate likely play an important role in shaping the challenges they face during the hiring process.

Of course, this approach is not without limitations. The employers who utilize particular types of stereotypical language in their job postings, for example, may differ from those who do not on a host of observable and unobservable characteristics. Thus, it may be some other aspect about these employers – not the language in the job posting – that is driving the effect of interest. Nevertheless, this approach has the potential to provide valuable insights.

Coming back to our hypothetical example, a second approach to gaining traction on competence as a potential mechanism driving discrimination against African Americans would be for the researcher to manipulate the competence construct experimentally. In this case, an additional axis of variation would be built in to the experiment where the researcher would manipulate signals about the competence of the applicant. This could be done through language in the cover letter, the content of

the resume itself, or, in some contexts, through language in a reference letter. If the competence stereotype mechanism is correct, then racial discrimination should be weaker in the cases where a high level of applicant competence is signaled.

Kaas and Manger (2011) utilize this approach in their study of ethnic labor market discrimination. In their audit study, the authors sent applications for student internships in Germany, randomly assigning the applicants either a Turkish-sounding name or a German-sounding name. Additionally, they manipulated whether the application was sent with a reference letter providing favorable information about the applicant's personality. They theorize that these reference letters should assist in mitigating ethnic discrimination by providing positive information about the Turkish applicants. That is precisely what they find. While there is significant discrimination against Turkish applicants, that discrimination disappears among the applications where the positive reference letters were present. Here, the researchers are able to provide evidence of a key mechanism at play in leading to discrimination in the labor market by manipulating the presence or absence of that construct in the experimental design.

While a powerful tool for examining mechanisms, this approach can result in at least two additional challenges for the researcher. First, the additional manipulations require larger samples to ensure that the study is adequately powered to detect the effects of interest. This may present some challenges, depending on the resources available to the researcher. Second, employers may devalue signals of competence, or other attributes, from particular groups. Thus, in our hypothetical example, a strong competence signal for an African American applicant may carry less weight than a strong competence signal from a white applicant. Therefore, the competence signal may not close the gap between white and black applicants, but could still be a mechanism that is at play in driving discrimination against African Americans.

9.1.2 Combining Audit Studies with Survey or Lab Experiments

Another way to assist with identifying the mechanisms that drive audit study findings is to pair audit studies with other types of experiments, such as lab or survey experiments. These other types of experiments enable the researcher to measure some of the mechanisms that are unable to be captured in audit studies.

For example, a researcher could first conduct an audit study of labor market discrimination by sending matched pairs of applications to apply for real job openings, varying only the race of the applicant. Then, the researcher could use the same application materials in a lab experiment, asking participants to make recommendations about which applicant to interview or hire, which would serve as a proxy for the outcome measure from the audit study.[4] The researcher would likely anticipate

[4] A distinct, but related, strategy is to contact the employers that were targeted in an audit study and collect information about their attitudes or beliefs. Indeed, utilizing this approach, Rooth (2010)

finding less favorable outcomes for African American job applicants in both the audit study and the lab experiment. However, in the lab experiment, there would be a key additional component. The subjects in the lab experiment would also be asked to evaluate the applicants they reviewed along a host of measures designed to capture perceptions of the applicant's competence. These competence perceptions could then be used in a mediation analysis to determine if they explain the racial disparities in interview or hiring recommendations by race. If they do, the evidence from the lab experiment could be used to bolster the theoretical argument about the mechanism that may be driving the racial discrimination found in the audit study.

Existing scholarship has deployed this strategy. In their article examining the "motherhood penalty" in the labor market, Correll et al. (2007) combined evidence from a lab experiment and an audit study. A key component of the theoretical argument in their paper is that perceptions of mothers as less competent and less committed drive, at least in part, their disadvantage in the labor market. To test this argument, Correll et al. (2007) empirically demonstrate in the lab-experimental component of their research that, indeed, perceptions of competence and commitment assist in explaining the motherhood penalty in terms of promotion likelihood, being recommended for management training, being recommended for hire, and salary recommendations. These findings from the lab experiment about the mediating role of competence and commitment perceptions are then useful in interpreting their finding that a motherhood penalty exists in their audit study of actual job openings (Correll et al. 2007).

In some of my own research (Pedulla 2016), I combine audit study techniques with a survey-experimental design to examine how histories of non-standard and mismatched employment (e.g., part-time work, temporary agency employment, and skills underutilization) affect men's and women's employment opportunities. The audit study demonstrates that men face severe penalties for both part-time work and skills underutilization, whereas women only face penalties for skills underutilization. I also conducted a survey experiment with hiring decision-makers at U.S-based firms to explore how perceptions of applicants' competence and commitment could assist in explaining the findings from the field experiment (Pedulla 2016). This supplementary survey-experimental data provides evidence that competence and commitment perceptions do play an important role in explaining the effects of non-standard and mismatched employment and, thus, assists in interpreting the findings from the audit study.

While the aforementioned studies present compelling evidence about the benefits of combining audit studies with lab- and survey-experimental data to uncover the underlying mechanisms at work, this approach is not without its challenges. In some cases, what respondents reveal in the lab or survey context may not align with their behaviors in the "real world." For example, Pager and Quillian (2005) find that employers who indicated a high likelihood of hiring formerly incarcerated individuals in a survey context were no more likely to actually hire formerly incarcerated

finds that employers' scores on the Implicit Association Test (IAT) are correlated with the treatment of Arab-Muslim job applicants in the Swedish labor market.

individuals in an actual audit study. Additionally, the task of evaluating job applicants in a lab or survey context likely differs in important ways from the evaluation of applicants in the actual labor market. In the lab or survey context, the stakes are often lower in terms of the consequences of the hiring decision. Additionally, the pool of applicants in a lab or survey experiment, against which the experimentally manipulated applicant is compared, may differ in important ways from the pool of applicants in an audit study. Thus, there are important cases where the approach of combining audit studies with survey or lab experiments may not be possible or may not mirror the empirical findings in the audit study. However, this approach can be useful in painting a more holistic picture of a given labor market process under conditions when survey or lab results align with audit study findings.

9.1.3 Collecting Qualitative Data

While the above section outlined how lab and survey experiments can supplement audit study data, qualitative data can also complement audit study data and assist in teasing apart the key mechanisms that are at play. Qualitative data of multiple sorts – interviews with employers as well as the experiences of testers in in-person audits – can enhance our understanding of the mechanisms that drive audit study findings.

Following up on our example from above, imagine that a researcher's argument is that employers' stereotypes about black workers as less competent than white workers drive their discrimination against this group of applicants. While audit study data is unable to document employers' stereotypes, qualitative data can help to understand what stereotypes employer may hold about black workers. Indeed, a series of qualitative studies with employers have documented employers' stereotypes of black workers' competence, skill, ability, and motivation (Moss and Tilly 2001; Waldinger and Lichter 2003). While interview data of this sort can be complicated to interpret because employer may not be forthright about their stereotypes (or even be aware of their own biases) and stereotypes do not necessarily drive behavior, interviews with employers could provide complementary evidence to scholars about how employers think and talk about workers of different races (Pager and Karafin 2009).

Recent audit studies have deployed these techniques in meaningful ways. Rivera and Tilcsik (2016), for example, utilize an audit study to examine the consequences of social class background on the likelihood that male and female applicants receive callbacks for jobs at 316 large law firms in the United States. They find that higher-class male applicants receive a higher callback rate than lower-class male, upper-class female, and lower-class female applicants. Yet, the audit study data are limited in their ability to tease apart the underlying mechanisms that may drive this complex empirical pattern. Thus, to understand the potential mechanisms driving their findings, Rivera and Tilcsik (2016) also interviewed lawyers with hiring experience. They found that the lawyers they spoke with expressed concerns about the commitment of higher-class women, mentioning potential competing demands of family

life. Thus, these interviews provide insights about the mechanisms that may be at play in the audit study context.[5]

Gaddis (2015) utilizes a different type of qualitative data to complement his audit study findings: emails from employers that were intended for colleagues, but accidentally were sent to the job applicant and which Gaddis received when coding employers' responses to the fictitious job applications. Gaddis's (2015) audit study aimed to understand how elite versus less selective college credentials intersect with an applicant's race to produce discrimination. Out of the 13 cases where employers accidentally sent emails to the "applicant" (e.g., Gaddis himself), five of them mentioned the elite university that the applicant attended: Harvard, Duke, or Stanford. As Gaddis writes: "These accidental e-mails provide some limited qualitative insight into the importance employers place on a degree from an elite university" (2015, p. 1471). Researchers using audit study techniques can be well-served by thinking of innovative ways to collect qualitative data that can shed light on the key processes of interest.

Many of the audit studies discussed thus far utilize correspondence methods – submitting electronic or paper applications for job openings – rather than in-person applications. Yet, when audit studies utilize in-person applicants, often referred to as "testers," these individuals can provide meaningful qualitative data that can assist in understanding the mechanisms that drive audit study findings. Pager et al. (2009a) implemented an in-person audit study of racial discrimination in New York City's low-wage labor market. While their audit study findings documented severe racial discrimination, they also utilized qualitative data from testers' experiences – and, specifically testers' experiences interacting with employers – to tease apart important underlying processes leading to racial exclusion. For example, they document a set of cases where "the same deficiencies of skill or experience appear to be more disqualifying for the minority job seekers" (p. 789). Additionally, they identify a process that they refer to as "job channeling," whereby job candidates were encouraged to apply for a different job than the one they had applied for or inquired about. The qualitative experiences of the testers indicate that the job channeling process is racialized in important ways, with minorities being likely to be channeled into worse jobs and whites being channeled into better jobs. Thus, the qualitative data from testers about their experiences applying for jobs were able to add theoretical detail to the audit study findings.

9.2 Identifying Key Axes of Variation

The strategies outlined above can assist audit studies in identifying the mechanisms that may drive discrimination and bias, providing traction on questions about *why* discrimination exists. In this section, I move on to examine how audit studies can

[5] Rivera and Tilcsik (2016) also conducted a survey experiment and found that similar mechanisms were at play in the survey experiment and the interviews.

tackle important questions about *when* and *where* discrimination occurs. I focus on variation by individual characteristics as well as place or geography. However, there are certainly other types of variation that can be and have been explored with audit studies, such as differences across occupations (Booth and Leigh 2010). I then present some additional areas that may be useful for future scholarship to examine to assist with identifying theoretically important variation in labor market discrimination.

9.2.1 Variation by Individual-Level Characteristics

Experiences of the social world are rarely one-dimensional. Individuals occupy multiple social positions – such as race and gender – that may intersect in shaping how they are treated (Collins [1990] 2000; McCall 2005). Thus, it is important for scholars to think about how social group memberships interact with one another. Coming back to the hypothetical example from earlier – racial discrimination against African Americans – it is possible that racial discrimination may vary by gender (Browne and Misra 2003). In other words, the effects of race on hiring outcomes may not be consistent for men and women due to the differing stereotypes that employers hold about black men and black women (Moss and Tilly 2001). Thus, paying attention to heterogeneity by individual characteristics can reveal important theoretical distinctions between social groups. Certainly, researchers have built these features into some of their audit studies. Pager (2003) examined the joint effects of race and having a criminal background. Gaddis (2015) explored the effects of race and college selectivity. Pedulla (2016) addressed how employment histories, such as part-time work or temporary agency employment, intersect with gender.

More research, however, needs to be done to explore how distinct social positions aggregate to produce labor market advantage and disadvantage. This line of inquiry is important for at least two reasons. First, different status positions can and do combine in counter-intuitive and complicated ways. In the lab and survey context, for example, scholars have found that while white men face penalties for being gay, gay black men are actually evaluated more positively than straight black men (Pedulla 2014; Remedios et al. 2011). Thus, just because a status position has a negative effect for one group does not mean it will have a negative effect for all groups. It is important for researchers working in this area to develop audit studies that explicitly test for how different status positions combine with one another.

Second, when the consequences of different social statuses are examined in separate audit studies, it is challenging to compare the effect sizes of discrimination for the different groups. Given that the details of the design of each audit study are distinct, any differences in effect sizes in separate studies – for example, examining the consequences of gender in one audit and race in another audit – could be the result of the status positions themselves or could be artifacts of the audit study design. Simultaneously including multiple social groups in a single audit study

opens the opportunity to directly compare the effects of different statuses to one another and potentially enables researchers to calibrate the severity of the discrimination faced by members of different social statuses.

9.2.2 Variation by Place or Geography

In addition to examining the interactive nature of individual-level characteristics, understanding how discrimination varies across space can provide important theoretical insights. However, the geographic specificity of many audit studies – applying for openings in one or a small number of locations – can make it difficult to know whether things would look similar in other labor markets. Does racial discrimination look the same in Tulsa as it does in New York City? Probably not. Perhaps the racial composition of a local labor market shapes discrimination or possibly the population density or size of the urban space influence how race operates. However, it is difficult to know exactly how it may be different without empirical data to compare outcomes in different contexts.

One strategy to address this issue is to include multiple geographic locations within the same audit study. This is increasingly possible with the heightened use of the Internet for hiring (Gaddis 2015; Pedulla 2016; Kroft et al. 2013). But, for some audit studies – especially in-person audits – it may be quite difficult to run the study in multiple locations at the same time. In these cases, however, it may be possible to run similar studies in different locations at two different points in time. This strategy was used by Pager (2003), who implemented an in-person audit to assess hiring discrimination based on race and having a criminal record in Milwaukee. After completing the Milwaukee study, she replicated and expanded the design in New York City and found similar results (Pager et al. 2009a, b).

Additionally, many of the major audit studies conducted over the past 15 years have focused on applying for jobs in major metropolitan areas, such as New York City, Los Angeles, Chicago, and Boston. While some audit studies have been conducted outside of these areas (see Wallace et al. 2014), less is known about how well the findings from major metropolitan areas about racial discrimination and other forms of bias in the labor market may extend to areas with less population density. Conducting audit studies in these areas presents a host of unique challenges – the possibility of tighter social network use for hiring, more limited ability to avoid detection of the experiment, etc. – but could contribute in important ways to audit study research on labor market discrimination.

Including geographic variation in one's audit study design also opens the possibility of examining theoretically relevant variation. Tilcsik's (2011) study of discrimination against gay men, for example, audited employers in seven different states. The geographic diversity of this study enabled an analysis of how state-level variation in legal protections and attitudes could assist in explaining discrimination. Indeed, Tilcsik's (2011) findings provide compelling evidence that discrimination against gay men varies by whether legal protections are in place and whether state-

level attitudes about gay rights are more progressive. Thus, analyzing variation has the potential to generate valuable theoretical insights about the forces that exacerbate and mitigate discrimination.

In a separate audit study that utilized geographic variation in an important way, Kroft et al. (2013) examine the effects of unemployment duration on a worker's likelihood of receiving callbacks for jobs. They submitted applications for job openings in 100 U.S. cities, experimentally varying the length of unemployment an applicant experienced leading up to their application. They were then able to examine whether the effects of unemployment duration varied by the tightness of the labor market in which the applications were submitted. Indeed, they find that the penalizing effect of longer unemployment spells is more severe in tight labor markets (Kroft et al. 2013). This provides useful theoretical information about when and why unemployment penalizes workers. Without a wide-ranging set of locations in the study, Kroft et al. (2013) would not have been able to generate this insight.

While examining variation by geographic location can be valuable theoretically, it also has limitations. Characteristics that vary by place – such as legal protections for gay workers or the unemployment rate – are not exogenously determined. Therefore, it is possible that some factor other than the measured variable is driving the variation in average treatment effects that is detected empirically. Even with this caveat, however, examining variation across geographic locations holds promise for furthering our understanding of the forces that shape discrimination and bias at the hiring interface.

9.2.3 Paths Forward for Identifying Key Axes of Variation

The possibilities outlined above – exploring variation by individual-level characteristics, as well as place or geography – are important. And, researchers are taking steps in these directions as new audit studies are implemented. There are, however, some additional avenues for future research that could be useful for understanding the ways that discrimination and bias vary across the labor market.

One potentially powerful area where there is currently limited research is in how processes of labor market discrimination vary over time. As economic and social conditions change, does discriminatory behavior systematically shift? While comparing estimates from different audit studies that were conducted at different points in time can be useful in addressing some of these questions, the lack of standardization across audit studies can make it difficult to compare estimates. Different research teams may use different names to signal race, application materials are likely distinct in terms of skills and background, and the occupations and labor markets under investigation differ between studies. Thus, important insights may be able to be generated by developing a standardized procedure for measuring discrimination in the labor market over time or having the same research team deploy the identical audit study at different moments in time. At the same time, this type of standardization can be difficult given that the way hiring occurs varies in important

ways over time. Common application procedures today will be outdated in the future, likely in just a few years. Thus, utilizing today's procedures a decade from now would likely signal something to employers about the applicant not understanding how the labor market operates, potentially confounding the findings. Nevertheless, utilizing audit studies to examine how various types of discrimination and bias vary, or remain consistent, over time could contribute important new insights to this area of scholarship.

An additional possibility for identifying interesting variation is merging audit study data with administrative data on employers. A potentially fruitful avenue along these lines could be to merge audit study data with the data collected by the Equal Employment Opportunity Commission (EEOC) through their EEO-1 reports. EEO-1 reports collect information about the race and gender composition of establishments (above a certain size threshold), broken down by major occupational categories. Thus, merging audit study data on racial discrimination with EEO-1 reports would enable researchers to examine questions such as: are establishments with more minority managers less likely to discriminate against minority job applicants? Addressing questions such as this would significantly enhance our understanding of the organizational-level correlates of discrimination.

Another potential path forward for combining data to examine important variation would be to simultaneously survey and audit employers. The survey data would collect detailed information about an employer's hiring practices, cultural environment, legal environment, as well as other dimensions of work life. The audit study would test for discrimination against particular groups: racial and ethnic minorities, women, parents, LGBTQ individuals, etc. By merging the employer survey data with the audit study data, researchers would be able to address challenging questions about how internal organizational dynamics shape the types of workers that they are more or less likely to hire. Scholars could address questions such as: Are organizations with supportive work-family policies more or less likely to discriminate against mothers? Or, does increased formalization of hiring reduce race and gender discrimination?

To date, these questions have been difficult to answer. Even in cases where survey data about employers' policies and practices can be linked to administrative data about the race and gender composition of the workplace (e.g., Kalev et al. 2006), it is difficult to know whether the composition of the workplace is due to selection processes on the supply side of the job matching process, demand side behavior, or some combination of the two. By pairing the audit method with an employer survey, researchers would be able to link detailed information about employers' policies and practices with direct measures of discrimination.

9.3 Clarifying Issues of Representativeness

The questions addressed in the previous section about when and where discrimination emerge are closely related to issues of the representativeness of audit studies. In other words, to what population do the findings from a particular audit study

generalize? This is often a difficult question to answer because the jobs applied for in an audit study are not necessarily drawn from a sampling frame of a known population of employers or companies. Thus, advances in the sampling procedures for audit studies could improve knowledge about to what population particular audit study findings generalize. In turn, this can assist in providing clearly demarcated scope conditions about the findings – improving our knowledge of *when* and *where* discrimination and bias occur.

Recent audit studies have tended to use national on-line job posting websites (e.g., Monster.com, Indeed.com, CareerBuilder.com) to draw a sample of job opening to which to submit applications. While these websites offer a broad set of job types from a wide range of companies, more detailed information about exactly which types of companies post jobs on these websites and which types of companies do not would be useful. Additionally, it could be useful for researchers to draw their sample of job openings from multiple job posting sites, potentially including some jobs that are posted only on a company's website.[6] Increasing the heterogeneity in the locations from which audit study samples are drawn could have positive consequences for the representativeness of audit study findings.

Another potential way to improve the representativeness of audit studies would be to draw the set of companies to audit from a representative national database. For example, establishments could be sampled from EEO-1 reports, which are required by the federal government for establishments that meet certain criteria (see Kalev et al. 2006). Thus, the establishments included in the study would be representative of a known population. Researchers could then audit these companies when job openings become available. Of course, this strategy presents challenges on at least two fronts. Some companies in the sample may not be hiring over the period of the audit. Additionally, tracking the job postings of a large number of companies to potentially audit would be highly labor intensive. However, the opportunity to be able to generalize audit study results to a known population of establishments or companies would enhance the contribution of this line of research.

9.4 Beyond the Labor Market

The above discussion centered on the labor market and, specifically, hiring discrimination. However, the issues of mechanisms, variation, and representativeness extend to other domains as well. In the housing market, for example, racial discrimination has been well-documented (Fix and Struyk 1993; Ewens et al. 2014), but identifying when, where, and why discrimination emerges in the housing market remains an important set of issues. Similar to the labor market, however, scholars

[6] Additional creative strategies are also possible for researchers trying to diversify the ways in which they find job openings. In addition to utilizing newspaper advertisements, for example, Lahey (2008) cold-called employers in the two cities where she conducted her audit study to identify job openings.

have begun to make inroads in these areas. For example, Zhao (2005) examines racial discrimination in the number of houses a broker shows to home-seekers and provides evidence consistent with a hypothesis that racial differences are due to the prejudice of white customers. Similarly, Hanson and Hawley (2011) find that racial discrimination against African Americans in the rental market is attenuated when the e-mail messages inquiring about an apartment signal a high social class background. Thus, whites' prejudice and the links between race and social class are likely key mechanisms in producing racial inequality in the housing market.

Thinking about variation at the intersection of housing and credit markets, Ross et al. (2008) examine racial and ethnic discrimination in mortgage lending in Chicago and Los Angeles. They find discrimination in Chicago, however, there is limited differential treatment by race in Los Angeles. Additionally, Ross et al. (2008) examine how this process varies with the characteristics of the lender. They find that large lenders and lenders with more applications from African Americans are less discriminatory than smaller lenders and lenders who receive applications from a primarily white clientele (Ross et al. 2008). These findings highlight the ways that racial and ethnic discrimination is often contingent, providing insights about when, where, and even why discrimination emerges.

Looking within a new domain of the housing market – the sharing economy, specifically the Airbnb platform – Edelman et al. (2017) find that African Americans are 16 percent less likely to be "accepted" by hosts than are whites. The authors also find that racial disparities are most severe among hosts who have not previously had an African American guest through Airbnb. Thus, they highlight important variation that may provide a point of intervention for Airbnb to target efforts for reducing racial discrimination. Additionally, the Edelman et al. (2017) study highlights the importance of using audit studies to understand discrimination and bias in different and emerging types of markets (see also Besbris et al. 2015). Additional scholarship that explicitly theorizes and tests for variation in the ways that discrimination plays out across different aspects of a related market – for example, the rental process, home buying, securing a mortgage, and the short-term rental "sharing" economy in the housing market – would significantly advance the contribution of audit study research to the social scientific understanding of discrimination.

9.5 Conclusion

Audit study research has become a central component of scholarship in the social sciences across disciplines, institutional domains, and areas of inquiry. Significant progress has been made in implementing these types of research designs and important knowledge about the contours of discrimination and bias has been generated from this line of research. Yet, questions about why, where, and when discrimination emerges can be challenging to address with audit studies.

In this chapter, I have discussed three of the emerging frontiers in audit study research that can assist in addressing these issues: (1) identifying mechanisms, (2)

examining variation, and (3) clarifying representativeness. While scholars have certainly started to include techniques for addressing these issues in their research designs, additional advances are needed in these areas to provide a more detailed and nuanced understanding of the set of barriers faced by different social groups in the labor market and other contexts. Specifically, these aspects of audit study research have important implications for theories of discrimination and bias. By focusing on mechanisms, variation, and representativeness, scholars can continue to develop and refine theoretical perspectives that help to clarify the ways that particular groups are excluded from access to employment, housing, and other aspects of social life.

In addition to the theoretical advances that these frontiers can provide, there are also policy-relevant insights that could be important for reducing discrimination. For example, identifying the organizational policies and practices that are associated with race and gender discrimination in hiring could assist companies with concrete strategies for diversifying their workforce. Similarly, if certain types of companies are more likely to discriminate, additional monitoring and enforcement resources targeted at those types of companies could be valuable. Thus, by pushing audit study methods forward, the evidence from research in this area may also shape public and organizational policies designed to reduce discrimination and bias.

Overall, the insights produced by audit studies have advanced social-scientific knowledge and policy in important ways. Yet, as scholars utilize this method moving forward, it will be important to think in new and innovative ways to identify mechanisms, document variation along key axes of differentiation, and more clearly understand the representativeness of the findings that are derived from this powerful methodological tool.

References

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*, 991–1013.

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2015). Effect of neighborhood stigma on economic transactions. *Proceedings of the National Academy of Sciences, 112*(16), 4994–4998.

Booth, A., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economic Letters, 107*, 236–238.

Browne, I., & Misra, J. (2003). The intersection of gender and race in the labor market. *Annual Review of Sociology, 29*, 487–513.

Collins, P. H. ([1990] 2000). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. New York: Routledge.

Correll, S. J., Benard, S., & In, P. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology, 112*, 1297–1338.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review, 106*(3), 778–806.

Deterding, N. M., & Pedulla, D. S. (2016). Educational authority in the 'open door' marketplace: Labor market consequences of for-profit, nonprofit, and fictional educational credentials. *Sociology of Education, 89*(3), 155–170.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*(2), 1–22.

Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *The Review of Economics and Statistics, 96*(1), 119–134.

Fix, M., & Struyk, R. J. (Eds.). (1993). *Clear and convincing evidence: Measurement of discrimination in America*. Washington, DC: Urban Institute Press.

Gaddis, S. M. (2015). Discrimination in a credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.), *Audit studies: Behind the scenes with theory, method, and nuance*. Cham: Springer International Publishing.

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics, 70*, 99–114.

Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology, 60*(4), 669–692.

Kaas, L., & Manger, C. (2011). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review, 13*(1), 1–20.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review, 71*(4), 589–617.

Kroft, K., Lange, F., & Notowidigo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics, 128*(3), 1123–1167.

Lahey, J. N. (2008). Age, women, and hiring: An experimental study. *Journal of Human Resources, 43*(1), 30–56.

Lauster, N., & Easterbrook, A. (2011). No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. *Social Problems, 58*(3), 389–409.

McCall, L. (2005). The complexity of intersectionality. *Signs: Journal of Women in Culture and Society, 30*(3), 1771–1800.

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science, 23*(7), 710–717.

Mishel, E. (2016). Discrimination against Queer women in the U.S. workforce: A resume audit study. *Socius: Sociological Research for a Dynamic World, 2*, 1–13.

Moss, P. I., & Tilly, C. (2001). *Stories employers tell: Race, skill, and hiring in America*. New York: Russell Sage Foundation.

Neumark, D. (1996). Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics, 111*(3), 915–941.

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108*(5), 937–975.

Pager, D., & Karafin, D. (2009). Bayesian bigot? Statistical discrimination, stereotypes, and employer decision making. *The Annals of the American Academy of Political and Social Science, 621*(1), 70–93.

Pager, D., & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review, 70*(3), 355–380.

Pager, D., Western, B., & Bonikowski, B. (2009a). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review, 74*, 777–799.

Pager, D., Western, B., & Sugie, N. (2009b). Sequencing disadvantage: Barriers to employment facing young black and white men with criminal records. *The Annals of the American Academy of Political and Social Science, 623*(1), 195–213.

Pedulla, D. S. (2014). The positive consequences of negative stereotypes: Race, sexual orientation, and the job application process. *Social Psychology Quarterly, 77*(1), 75–94.

Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*(2), 262–289.

Remedios, J. D., Chasteen, A. L., Rule, N. O., & Plaks, J. E. (2011). Impressions at the intersection of ambiguous and obvious social categories: Does gay + black = likable? *Journal of Experimental Social Psychology, 47*(6), 1312–1315.

Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals. *American Sociological Review, 81*(6), 1097–1131.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523–534.

Ross, S. L., Turner, M. A., Godfrey, E., & Smith, R. R. (2008). Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process. *Journal of Urban Economics, 63*, 902–919.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology, 117*(2), 586–626.

Waldinger, R. D., & Michael, I. L. (2003). *How the other half works: Immigration and the social organization of labor*. London/Berkeley: University of California Press.

Wallace, M., Wright, B. R. E., & Hyde, A. (2014). Religious affiliation and hiring discrimination in the American south: A field experiment. *Social Currents, 1*(2), 189–207.

Zhao, B. (2005). Does the number of houses a broker shows depend on a homeseeker's race? *Journal of Urban Economics, 57*, 128–147.

Index

S. M. Gaddis (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Methodos Series 14, https://doi.org/10.1007/978-3-319-71153-9