

# Measuring and Explaining Discrimination

by

Charles Crabtree

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Political Science)  
in the University of Michigan  
2019

Doctoral Committee:

Professor Christian Davenport, Co-Chair  
Professor Matt Golder, Co-Chair  
Professor Robert Axelrod  
Professor Vince Hutchings  
Professor Beatriz Magaloni  
Professor Kiyoteru Tsutsui

Charles Crabtree  
ccrabtr@umich.edu  
ORCID iD: 0000-0001-5144-8671

© Charles Crabtree 2019

# Dedication

*To my wife, without whom this would not have been either possible or worth it.*

# Acknowledgments

By submitting this dissertation, I bring to a close my graduate school career. The years I have spent in graduate school count among the most trying in my life, but also among the most rewarding. Through my Ph.D. studies, I have met some of the best and smartest people I have ever known, many of whom are thanked below. From them I have learned to be a better reader, writer, researcher, colleague, and friend. My interactions with these people are the true reward of my graduate studies.

The most enjoyable aspect of graduate school for me has been discussing research with colleagues at conferences, mini-conferences, workshops, or CP Group meetings. Sometimes these conversations resulted in joint projects, sometimes they did not. Regardless of the outcome, these conversations have nourished and sustained me. To all those who have joined me in these chats over the years, I thank you.

I am particularly grateful to the person who initially shared these conversations with me, Holger L. Kern. My first mentor, he taught me to think like a social scientist. Just as importantly, he gave me the confidence to believe that I could complete my degree and join the professoriate. Our collaborations have provided me with a constant education, and our joint service activities have helped me overcome my feelings of professional insecurity. I owe him so very much for his continued friendship and mentorship.

I've been fortunate to have many mentors in graduate school. Many of my coauthors have taken on this role at various points in my studies. I'm particularly thankful for the personal support that Quintin Beazer, Dan Butler, Courtenay Conrad, Chris Sullivan, John Holbein, Indriði Indriðason, Michael Gaddis, Doug Lemke, Ben Newman, Amanda Murdie, Steve Pfaff, and other collaborators offered throughout the job market season. They made a difficult year far easier.

The members of my dissertation committee have been my primary mentors in the past several years. From each of them, I have received treasured lessons about research, the profession, and life more generally. Beatriz Magaloni has been very supportive of my work, particularly my research on policing. Moving forward, I hope to follow her model of policy-focused research. Vince Hutchings has helped encourage my research on race and ethnic politics in the American context, and has supported me in making important connections within that community. In my career, I hope to mirror his devotion to mentorship, and his generosity to others.



Bob Axelrod has left an indelible mark on my career, shaping my views of how researchers should work and what they should study. He has generously read many papers that I've written — no small or pleasant task — and has provided deep insights into how the pieces relate to each other and why I probably study what I do. He has provided innumerable ideas about my own research. He has increased my ambitions and has given me the confidence to believe that I can achieve them. I am very fortunate to be one of his students.

Kiyoteru Tsutsui has deeply influenced my career and life as well. Through his example, he has taught me how to lead a project, a center, and an intellectual community. He connected me to many new intellectual networks and has worked tirelessly to help me obtain my career goals. Not only has he mentored me professionally, but he's become a trusted personal advisor and valued friend. I look forward to seeing our collaborations and friendship deepen with time.

To my co-chairs, I am especially grateful. Christian Davenport helped me feel at home in the department after transferring from Pennsylvania State University. He became my advisor at a difficult and rather late point in my graduate school career. Despite this, he has invested as much in my research and personal and professional development as any of his other students. He has pushed me to consolidate my work around a few key themes, to take risks theoretically, and to set loftier goals. I look forward to many future conversations about research and the profession in the years to come as we continue working together on projects dealing with state violence.

Like Holger Kern, Matt Golder has been a through line in my graduate school career. From my time as a first-year student at Pennsylvania State University throughout my years at the University of Michigan, Matt has been a dedicated advisor, offering unceasing support for me and my work. He has taught me how to write, spending hours at a time working with me on one or two paragraphs. He has taught me how to think and has shown me through his own example that the best social scientists think like social scientists all the time.

It is difficult to write about Matt for long without also writing about Sona Golder. As members of the CP Group often joke, to have one as your advisor is to have both as your advisors. Like Matt, Sona has helped me think like a social scientist, providing keen insights into strategic behavior. She has also been a constant source of good humor and kindness, particularly in times when I doubted my work, my abilities, or my future.

What the Golders have done for me during my meandering graduate school journey cannot be overstated. They have provided emotional support, intellectual community, and friendship whenever needed, no matter the hour or day. They have created a nurturing and stimulating scholarly community, and invited me and their other students to participate in it. They have invited my wife and I into their home.<sup>1</sup> I will spend much of my career trying to emulate their work as researchers, teachers, mentors, and friends. I'm excited to continue working with them as I begin my professional career, and to continue learning

---

<sup>1</sup>A shout-out to Sean Golder, an exceptionally talented young man, who never seemed to mind the time that I and other students needed from his parents.

from these beloved friends.

Turning from the professional to the personal, I am grateful for my group of friends at the University of Michigan. Graduate school thrusts many challenges upon those who aim to complete it. Among others, Kevin Cope, Jesse Crosson, and Anil Menon have helped me circumnavigate and sometimes pass straight through the obstacles that have crossed my road to degree completion. I am deeply grateful for their friendship, and I look forward to having them as colleagues for many years to come. I am particularly thankful for the kindnesses of Kevin and his wife, Mila Versteeg. Their constant support throughout graduate school made dark days much lighter.

My parents, Janiece and Steve, have brightened my days as well. They have supported my education at all turns, even when such support has entailed considerable sacrifice. They have nurtured my curiosity, encouraged me through every point of my academic journey, and patiently waited as I found my way. The wisdom that they shared over early morning camp breakfasts, picnics, and home-cooked dinners helped light my path. In ways both large and small, they light it still.

The largest share of my gratitude belongs to Volha Chykina, my wife, my colleague, and my better in nearly all things. To quote Browning, she is the "nobler of us two." Her contributions to this dissertation, to my graduate school experience, and to my life deserve a tome larger than this document. Without her, I would not have the wonderful life I now enjoy. Without her, that wonderful life would mean very little. I look forward to closing this part of our life together, and to beginning the next one.

# Table of Contents

<b>Dedication</b>	<b>ii</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>List of Appendices</b>	<b>xiv</b>
<b>Abstract</b>	<b>xvi</b>
<b>I Introduction</b>	<b>1</b>
<b>Chapter 1. Studying Discrimination</b>	<b>2</b>
<b>II Methods</b>	<b>11</b>
<b>Chapter 2. Audit Studies in Political Science</b>	<b>12</b>
2.1 Introduction . . . . .	12
2.2 Basics of Audit Studies . . . . .	17
2.2.1 Be precise about the question and the sample . . . . .	18
2.2.2 Develop the instrument (or message) to maximize the likelihood that it reflects a commonly encountered communication . . . . .	19
2.2.3 Hold confounding factors constant . . . . .	20
2.3 Maximizing External Validity . . . . .	20
2.3.1 Include typical requests in the instrument . . . . .	21
2.3.2 Use different aliases . . . . .	22
2.3.3 Use multiple requests and names . . . . .	23
2.3.4 Use reasonable email or postal services . . . . .	24
2.3.5 Check the final instrument . . . . .	25
2.4 Additional Design Considerations . . . . .	26

2.4.1	Internal validity . . . . .	26
2.4.2	SPAM Concerns . . . . .	27
2.4.3	Email tracking . . . . .	27
2.5	Ethical and Other Considerations . . . . .	27
<b>Chapter 3.</b>	<b>Email Audit Studies</b>	<b>31</b>
3.1	Introduction . . . . .	31
3.2	Guide to Implementation . . . . .	34
3.2.1	Experimental Design and Sample Selection . . . . .	34
3.2.2	Email Address Collection . . . . .	36
3.2.3	Covariate Collection . . . . .	38
3.2.4	Treatment Randomization . . . . .	38
3.2.5	Email Delivery . . . . .	38
3.2.6	Pre-Implementation . . . . .	40
3.2.7	Implementation . . . . .	42
3.2.8	Outcome Collection . . . . .	43
3.2.9	Analysis . . . . .	45
3.3	Extending Audit Studies . . . . .	45
3.4	Discussion . . . . .	46
<b>Chapter 4.</b>	<b>Name Selection in Audit Studies</b>	<b>47</b>
4.1	Last Name Selection . . . . .	49
4.2	Data and Results . . . . .	51
4.3	Discussion . . . . .	53
4.4	Conclusion . . . . .	54
<b>III</b>	<b>Applications</b>	<b>56</b>
<b>Chapter 5.</b>	<b>Persistent Bias Among Local Election Officials</b>	<b>57</b>
5.1	Experiment Design . . . . .	58
5.1.1	Treatment assignment and implementation . . . . .	60
5.2	Results . . . . .	61
5.2.1	Evidence of implicit discrimination . . . . .	63
5.2.2	Awareness of experiment . . . . .	65
5.3	Conclusion . . . . .	66
<b>Chapter 6.</b>	<b>Does Religious Bias Shape Access to Public Services? A Large-Scale Audit Experiment Among Street-Level Bureaucrats</b>	<b>68</b>
6.1	Background and conceptual framework . . . . .	72
6.2	Previous experimental research on discrimination . . . . .	74
6.3	Potential sources of religious discrimination in American public education . . . . .	76

6.3.1	Secularism as a basis for anti-religious bias . . . . .	76
6.3.2	Judeo-Christian nationalism and religious bias . . . . .	79
6.3.3	Civil religion and bias against non-believers . . . . .	81
6.4	Research design and data . . . . .	84
6.5	Empirical results . . . . .	91
6.5.1	Potential mechanisms: Perceived costs of intense beliefs . . . . .	93
6.5.2	Benchmarking to theoretical expectations . . . . .	94
6.5.3	Limitations . . . . .	96
6.6	Conclusion and implications . . . . .	97
<b>Chapter 7. Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions</b>		<b>101</b>
7.1	Testing the Effect of Information . . . . .	103
7.2	Results . . . . .	106
7.3	Limitations . . . . .	109
7.4	Suggestions for Future Audit Studies . . . . .	111
7.5	Conclusion . . . . .	113
<b>Chapter 8. How Public Opinion Shapes Discriminatory Policing</b>		<b>115</b>
8.1	Introduction . . . . .	115
8.2	Theory . . . . .	122
8.2.1	Supply of Unbiased Policing . . . . .	122
8.2.2	Demand for Unbiased Policing . . . . .	125
8.2.3	Police Responsiveness . . . . .	129
8.3	Empirics . . . . .	132
8.3.1	Sample . . . . .	134
8.3.2	Experimental Research Design . . . . .	138
8.3.3	Empirical Analysis . . . . .	141
8.4	Conclusion . . . . .	151
<b>IV Supplementary Materials</b>		<b>153</b>
<b>Appendices</b>		<b>154</b>
<b>Bibliography</b>		<b>291</b>

# List of Figures

Figure 3.1	The eight stages of a typical email audit study. . . . .	35
Figure 4.1	$P(R S, L)$ Plots . . . . .	52
Figure 5.1	Core Findings . . . . .	64
Figure 5.2	Rapidly slowing rates of response. The vertical axis plots the cumulative number of responses, split by group identity of sender; the horizontal axis plots time since sending. Election Day and NASS emails are noted with vertical dashed lines. Responses follow a clear diurnal rhythm, and patterns of bias appear rapidly. . . . .	66
Figure 6.1	Email to principals . . . . .	99
Figure 6.2	Estimated treatment effects based on model in Table B.2 . . . . .	100
Figure 8.1	Theoretical Expectations about Discriminatory Policing - Concern about Public Demand . . . . .	131
Figure 8.2	OLS Models with Data from Elected Officials . . . . .	145
Figure 8.3	Perceptions of Public Demand for Racial Equality and Discrimination . . . . .	148
Figure 8.4	OLS Models with Data from Law Enforcement Administrators . . . . .	150
Figure A.1	CONSORT Document . . . . .	157
Figure A.2	A likely view of our stimulus in local election officials' email inboxes. Subfigure (a) presents the view in gmail, (b) outlook, and (c) yahoo inboxes. . . . .	159
Figure A.3	A likely view of our stimulus, once opened, in local election officials' email inboxes. Subfigure (a) presents the view in gmail, (b) outlook, and (c) yahoo inboxes. . . . .	160
Figure A.4	The number of emails sent is marked on the y-axis, and the time (in UNIX seconds, in the UNIX epoch) are plotted on the x-axis. Note the 30 minute gap in sending. Here, we waited to ensure that emails were making it to officials' inboxes, before green-lighting the remainder of the production email run. . . . .	180

Figure A.5	On the x-axis are the minutes elapsed since the first time the local election officials opened our stimulus, until the time that we received a response from that election official. On the y-axis are the cumulative number of responses that have been received in that duration of time. . . . .	182
Figure A.6	Flexible estimates of HTE across Arab/Muslim population (left plot) and the three-level factor that indicates (0.0) zero arab/muslim population; (1.0) Less than 1% Arab/Muslim population; and (2.0) 1% or more Arab/Muslim population. . . . .	190
Figure B.1	Estimated probabilities of reply based on model in Table B.2 . . . . .	221
Figure B.2	Comparison between sample and NCES population I . . . . .	222
Figure B.3	Comparison between sample and NCES population II . . . . .	223
Figure B.4	Comparison between sample and NCES population III . . . . .	224
Figure B.5	Treatment effect heterogeneity I . . . . .	225
Figure B.6	Treatment effect heterogeneity II . . . . .	226
Figure C.1	Race of Officials by Race of Constituents . . . . .	228
Figure C.2	Treatment Email 1 . . . . .	229
Figure C.3	Site Screenshot - Research Summary Page 1 . . . . .	230
Figure C.4	Site Screenshot - Research Summary Page 2 . . . . .	231
Figure C.5	Site Screenshot - Research Summary Page 3 . . . . .	232
Figure C.6	Site Screenshot - Home Page . . . . .	233
Figure C.7	Survey . . . . .	234
Figure C.8	Black-sounding Constituent Names . . . . .	236
Figure C.9	White-sounding Constituent Names . . . . .	237
Figure D.1	Number of Policing Articles Published in Political Science Journals	244
Figure D.2	The Geographic Distribution of Political Science Research on Policing	246
Figure D.3	Principal-Agent Scenarios . . . . .	249
Figure D.4	Chain of Delegation . . . . .	251
Figure D.5	Email Invitations . . . . .	253
Figure D.6	Paragraph Ratings . . . . .	255
Figure D.7	Survey Timeline . . . . .	256
Figure D.8	Map of Respondents . . . . .	258
Figure D.9	Law Enforcement Administrator Descriptive Statistics . . . . .	260
Figure D.10	Law Enforcement Administrator Descriptive Statistics (Continued)	261
Figure D.11	Elected Official Descriptive Statistics . . . . .	263
Figure D.12	Elected Official Descriptive Statistics (Continued) . . . . .	264
Figure D.13	Treatment Recollection . . . . .	266
Figure D.14	Description of Public Interest . . . . .	268
Figure D.15	Initial Black Last Names . . . . .	270
Figure D.16	Black Last Names . . . . .	273

Figure D.17 Black Names . . . . .	274
Figure D.18 Latino Names . . . . .	275
Figure D.19 White Names . . . . .	276
Figure D.20 ALES Homepage . . . . .	280
Figure D.21 ALES About Page . . . . .	281
Figure D.22 ALES Survey Page . . . . .	282
Figure D.23 ALES People Page . . . . .	283
Figure D.24 ALES Contact Page . . . . .	284
Figure D.25 Site Evaluations . . . . .	285
Figure D.26 Dependent Variables for Law Enforcement Administrators . . . . .	287
Figure D.27 Dependent Variables for Elected Officials . . . . .	288



# List of Tables

Table 5.1	Response Rates by Experimental Condition . . . . .	62
Table 7.1	Email Replies from Elected Municipal Officials . . . . .	108
Table 7.2	Email Replies from Elected Municipal Officials . . . . .	109
Table A.1	Local Election Officials excluded prior to randomization . . . . .	156
Table A.2	Features manipulated for random assignment of messages to regis- trars of voters. . . . .	163
Table A.3	Question FE Model . . . . .	165
Table A.4	Blocking . . . . .	170
Table A.5	Response Rates by Experimental Condition . . . . .	171
Table A.6	Causal Estimates . . . . .	173
Table A.7	Robust to Logit and Probit Specification . . . . .	174
Table A.8	Robust to Logit and Probit Specification . . . . .	175
Table A.9	Robust to Pilot Exclusion . . . . .	177
Table A.10	Robust to Pilot Exclusion . . . . .	178
Table A.11	Cox Proportional Hazards Models . . . . .	184
Table A.12	No Difference in Estimates in Interference States . . . . .	185
Table A.13	TEH - Communities by Minority Share . . . . .	187
Table A.14	TEH - Communities by Ethnicity . . . . .	188
Table A.15	TEH - Arab Communities . . . . .	191
Table A.16	Name Score Table . . . . .	192
Table B.1	Balance . . . . .	215
Table B.2	Parameter estimates (probit) . . . . .	216
Table B.3	Parameter estimates (probit) controlling for blocks . . . . .	217
Table B.4	Parameter estimates (OLS) . . . . .	218
Table B.5	Parameter estimates with MA omitted (OLS) . . . . .	219
Table B.6	Parameter estimates from population-weighted sample (WLS) . . . . .	220
Table D.1	Top 10 Most Relevant Stemmed Terms by Topic . . . . .	247
Table D.2	Surnames and Occurrence Across Racial Groups . . . . .	271
Table D.3	New Names and Racial Conditions . . . . .	272
Table D.4	OLS Results - Elected Officials . . . . .	289

Table D.5 OLS Results - Law Enforcement Administrators . . . . .	290
--	-----

# List of Appendices

<b>Appendix A. Appendix for ‘Persistant Bias Among Local Election Officials’</b>	<b>154</b>
A.1 Email Scraping . . . . .	154
A.2 Email Server Construction . . . . .	158
A.3 Email Back-End Considerations . . . . .	158
A.4 Mailer Content . . . . .	162
A.5 Pilot . . . . .	164
A.6 No Question Effects . . . . .	165
A.7 Name Selection . . . . .	166
A.8 Blocking . . . . .	168
A.9 Nonparametric Results . . . . .	171
A.10 Fixed Effects Models . . . . .	172
A.11 Robust to Link Function . . . . .	174
A.12 Pilot Inclusion . . . . .	176
A.13 Email Send Timing . . . . .	179
A.14 Time to Response . . . . .	181
A.15 No Damage from Spillover . . . . .	183
A.16 Limited District Characteristic Heterogeneity . . . . .	186
A.17 Names and Assessment of Racial and Ethnic Group . . . . .	192
<b>Appendix B. Appendix for ‘Does Religious Bias Shape Access to Public Services? A Large-Scale Audit Experiment Among Street-Level Bureaucrats’</b>	<b>206</b>
B.1 Treatment effect heterogeneity . . . . .	206
B.2 Generalizing impact estimates to NCES universe . . . . .	209
B.3 Ethics . . . . .	210
<b>Appendix C. Appendix for ‘Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions’</b>	<b>227</b>
C.1 Census and ICMA Data . . . . .	227
C.2 Treatment Email . . . . .	228
C.3 Research Summary . . . . .	230
C.4 Aliases in Experiment . . . . .	236

C.5	List of Questions Used in Emails . . . . .	238
<b>Appendix D. Appendix for ‘How Public Opinion Shapes Discriminatory Policing’</b>		<b>241</b>
D.1	Quantitative Analysis of Policing Literature . . . . .	241
D.2	Principal Agent Description . . . . .	248
D.3	Chain of Delegation . . . . .	250
D.4	Email to Law Enforcement Administrators . . . . .	252
D.5	Email Pre-test . . . . .	254
D.6	Timeline for Surveys . . . . .	256
D.7	Location of Respondents . . . . .	257
D.8	Law Enforcement Administrator Respondent Details . . . . .	259
D.9	Elected Officials Respondent Details . . . . .	262
D.10	Treatment Pretest . . . . .	265
D.11	Public Demand for Equality Treatment . . . . .	267
D.12	Name Selection . . . . .	269
D.13	Elected Official Vignettes . . . . .	277
D.14	Surveys . . . . .	278
	D.14.1 ALES Website . . . . .	278
D.15	Dependent Variables . . . . .	286
D.16	Results - Elected Officials . . . . .	289
D.17	Results - Law Enforcement Administrators . . . . .	290

# Abstract

How can we measure discrimination? What drives it? How can we reduce it? My dissertation addresses these important questions. In the first part, I provide methodological guidance on how to conduct audit studies. In Chapter 2, I offer the first comprehensive guide to conducting audit studies in political science. In Chapter 3, I provide the first introduction that I am aware of to conducting audit studies via email. These chapters provide advice about how researchers can improve existing audit study designs and implement them with increased efficiency. In Chapter 4, I address one of the most important methodological issues involved in audit studies — name selection. I demonstrate that the probability of a name denoting a race varies considerably across contexts and that this is more of a problem for some names than others. This suggests limitations for (1) the generalizability of audit study findings and (2) the interpretation of geography-based conditional effects.

In the second part of my dissertation, I use audit and survey experiments to better understand racial, gender, and religious discrimination in America. The first set of chapters build on past audit studies by not only measuring discrimination but also attempting to identify its causes. In Chapter 5, I measure discrimination by county election officials during the 2016 election cycle, showing that the bias toward Latinos observed during the 2012 election has persisted. I also show that Arab/Muslim Americans face an even greater barrier to communicating with local election officials, and that this bias appears driven by

implicit discrimination. I find no evidence of bias toward Blacks, however, indicating that discrimination against groups in political contexts might be sample dependent.

In Chapter 6, I examine religious discrimination among public-school principals, an important group of street-level bureaucrats. I emailed the principals of more than 45,000 public schools and asked for a meeting, randomly assigning the religious affiliation/non-affiliation of the family and family belief intensity. I find evidence of substantial discrimination against Muslims and atheists, particularly when their religious beliefs are high, as well as bias against ardent protestants and Catholics. These results suggest that one of the mechanisms driving discrimination is belief intensity.

The remaining chapters in the second part of the dissertation extend prior work by not only identifying discrimination in important contexts but also by attempting to reduce it. In Chapter 7, I provide the results from an adapted audit experiment designed to test whether making local officials aware of their possible biases could reduce discrimination. I find no evidence that my informational treatment influences discriminatory behavior, but that White, local, elected officials are less responsive to Black constituents. This is concerning as local government is often the level that most directly affects citizens' daily lives.

In Chapter 8, I investigate racial discrimination by the police. I argue that it depends in a conditional way on the extent of egalitarian views among the police *and* the public. To test my theory, I conduct a survey experiment with American law enforcement administrators and elected officials who oversee the police. Elected politicians exhibit less racial discrimination in law enforcement oversight when informed that the public supports racial equality in policing. Police, though, do not react to perceived public demand for egalitarianism. My results suggest that public attitudes toward racial equality influence police discrimination perhaps only indirectly.

# **Part I.**

## **Introduction**

# Chapter 1. Studying Discrimination

Employers pay female workers less (Macpherson and Hirsch, 1995; Cohn, 2000; Wolfers, 2006). United States legislators prefer White constituents (Butler, 2014; Grose, 2011; Butler and Broockman, 2011*a*). The police discriminate against Blacks (Baumgartner, Epp and Shoub, 2018; Epp, Maynard-Moody and Haider-Markel, 2014; Mauer, 2006; Tonry, 2011). Landlords are less likely to rent properties to Hispanics (Hanson and Santas, 2014; Ondrich, Stricker and Yinger, 1999; Ross and Turner, 2005). Universities are biased against Asian student applicants (Bunzel and Au, 1987; von Spakovsky, 2018). Society treats individuals with fringe political beliefs poorly (Wetherell, Brandt and Reyna, 2013). Schools violate the rights of religious minorities (Pfaff et al., 2018).

Statements like these — where individuals of one group are alleged to be treated worse because of some shared gender, racial, ethnic, political, or religious attribute — are very common in the news, political discourse, common conversation, and scientific journals.<sup>2</sup> They have served as the basis for many political positions, policy suggestions, social movements, and legal suits. They also color individual perceptions of important social and political institutions (Rocque, 2011; Schmitt, Branscombe and Postmes, 2003). Perhaps because of the normative and substantive importance of these declarations, their truth

---

<sup>2</sup>One indication of this can be found in Google search results. For example, as of March 25, 2019, a Google News search for the phrase ‘racial discrimination’ returns about 202,000 results, while a search for ‘gender discrimination’ returns about 122,000 results.



value remains fundamentally contested. For example, many politicians have argued that any evidence of discrimination is illusory, sometimes going so far as to claim that the group charged with discriminatory behavior is actually the target of negative biases (Sides, Tesler and Vavreck, 2018). How do we know who is right? How can we measure discrimination? What drives it? How can we reduce it? My dissertation addresses these broad questions and suggests many new avenues for future methodological and substantive work in this vital area of inquiry. Before summarizing my dissertation, I first discuss how researchers have previously attempted to identify discrimination and the limitations of these approaches.

Measuring discrimination is incredibly difficult (Council et al., 2004). The primary way that researchers have approached this task is by comparing descriptive statistics for different groups. This might be done by comparing raw rates, looking at cross tabulations, or by conducting basic bivariate statistical tests (Freedman, Pisani and Purves, 2007). To fix concepts, I will use the running example of researchers who are concerned about racial inequalities in the judicial system. These researchers might use a t-test to compare the rates at which Blacks and Whites are incarcerated, and interpret a statistically significant difference as evidence of discrimination. But how do we know that this difference is the result of discrimination? Members of those groups might differ in more ways than just their race. For example, Blacks might be born into worse socioeconomic conditions and therefore be more likely to commit crimes. If this is the case, then any differences might be the result of some observed or unobserved factor (Morgan and Winship, 2015).

To account for potential imbalances across groups, researchers have increasingly turned to a more sophisticated set of analytical approaches. For example, they might use regression models (Nelder and Wedderburn, 1972), matching estimators (Abadie and Imbens, 2006; Morgan and Harding, 2006), or selection models (Puhani, 2000). In each of these approaches, researchers use additional observed information about individuals in an attempt

to construct groups of individuals that differ only by some group-level characteristic. Returning to the running example, researchers might collect information about the race, age, education level, and household income of individuals charged with a crime. They might then create a dummy variable of whether charged individuals are convicted and regress this on a binary indicator of whether individuals are Black along with measures of socio-economic background and education. The general intuition here is that these additional, non-race measures capture any potential confounders between an individual's race and whether they are convicted. If the slope on the Black indicator is positive and statistically significant, then researchers might conclude that Blacks are, all else equal, more likely to be convicted than Whites, and thus experience discrimination in the judicial process.

There are several potential problems with this approach, though. First, to continue with the running example, there are often things such as criminal skill or education that differentiate between individuals but that are either unobserved or difficult to measure. Second, even if we could include all of the relevant variables, we often run into the problem of post-treatment bias (Montgomery, Nyhan and Torres, 2018; King and Zeng, 2006), which occurs when researchers control for the consequences of treatments. This bias can be in any direction. In the running example, researchers might try controlling for some characteristic that is affected by individual race. But most things that an individual does or experiences are affected by it. Indeed, many of the things that researchers might like to control for to isolate the effect of race on criminal justice outcomes, such as education or income, are likely the result of an individual's race. This means that it is extremely challenging to control for all potential confounders without inducing post-treatment bias. These issues highlight the fact that it is difficult, if not impossible, to identify discrimination through observational data.

It is partly in response to these research design issues that many people are increasingly

using experiments to measure and explain discrimination (Bertrand and Mullainathan, 2004a; Gaddis, 2018b). There are at least two experimental designs that researchers can use to study discrimination. The first design is typically referred to as an audit study, but is also known as a correspondence study or field experiment (Gaddis, 2018b). The phrase ‘audit study’ emphasizes the measurement aspect of these studies. In a financial audit, one looks at the books to see what is happening in a company. In an audit study, the researcher looks at the behavior of the responder to see if they are responding differently to certain messages (or senders) than they do to others. In this sense the researcher conducts an audit.<sup>3</sup> The key to this approach is for researchers to create identical messages, requests, or requesters, but randomize one essential attribute of them.<sup>4</sup> In the context of the running example, researchers could study this question by sending email requests for help to different judicial offices, randomizing whether the request comes from someone who is Black or White. They could then measure if offices are less responsive to requests from Blacks. By following these steps, researchers could measure levels of discrimination throughout the judicial system. While this approach provides researchers with a behavioral measure of discrimination, it also requires researchers to conduct research without participant consent and engage in deception.<sup>5</sup>

Despite the potential ethical issues, audit studies have a long history as a means of studying discrimination in housing and labor markets (Quillian et al., 2017; Wienk et al., 1979; Bertrand and Mullainathan, 2004b). Many of these studies looked at whether racial minorities were treated worse than their White counterparts (Wienk et al., 1979). While

---

<sup>3</sup>The phrase ‘correspondence study’ references the communication aspect of these studies. The researcher sends messages, or correspondence, and then measures how the receiver responds. Finally, the phrase ‘field experiment’ references the field nature of these studies and the corresponding gains in external validity. The researcher (should) design the study to reflect the type of messages that are normally sent so as to get a measure of how the people being studied normally respond.

<sup>4</sup>I provide guidance about how researchers can do this in Chapter 2.

<sup>5</sup>I address this trade-off in several parts of the dissertation and in several related working papers.

audit studies continue to study the treatment of racial minorities relative to Whites, the approach has also been applied to understand discrimination of many other groups. This includes studies of discrimination based on one's gender (Ayres and Siegelman, 1995), age (Ahmed, Andersson and Hammarstedt, 2013*a*), sexual orientation (Drydakis, 2014), religion (Adida, Laitin and Valfort, 2010; Pfaff et al., 2018), criminal record (Pager, Bonikowski and Western, 2009), and more (Gell-Redman et al., 2018*b*; Rivera and Tilcsik, 2016).

In comparison to their use in economics and sociology, audit studies have not been employed as widely in political science. As a result of this, researchers often do not have a clear sense of how to conduct them, or what best practices are. I shed some light on these issues in my dissertation.

One of the limitations of audit studies in political science is that they typically focus on identifying discrimination, rather than explaining it. While measuring discrimination is a necessary first step, we cannot ultimately reduce it until we better understand what factors are driving it. I address this gap in our understanding of discrimination by explicitly testing several mechanisms in the chapters that follow. I also take this line of research one step further and attempt to reduce discrimination in two important contexts.

The second type of design researchers are using to study discrimination is commonly called a survey experiment (Sniderman et al., 2011). This type of study has been used across the social sciences to understand a wide range of phenomena. The basic design involves presenting a short vignette to survey respondents. To identify the causal effect of some factor, researchers randomize one or more attributes of the vignette. Researchers could use a survey experiment to examine racial biases in the criminal justice system by asking a pool of potential jurors to evaluate a stylized version of a legal case, where the race of the accused is randomized. They could then ask respondents to evaluate the likelihood that the suspect committed the crime. While this approach minimizes some of the ethical

issues involved in audit studies, it typically provides an attitudinal rather than behavioral measure of discrimination. I use both types of experimental design to measure racial, gender, and religious discrimination in the ‘Application’ part of the dissertation.

In the first part of my dissertation, I describe audit studies in greater detail and provide methodological guidance on how to conduct them. In Chapter 2, I offer the first comprehensive guide to conducting audit studies in political science. In Chapter 3, I provide the first introduction to conducting audit studies via email. In Chapter 4, I address one of the most important methodological issues involved in audit studies — name selection. In the second part of my dissertation, I use audit and survey experiments to understand racial, gender, and religious discrimination in several important contexts. In Chapters 5 and 6, I build on past audit studies by not only measuring discrimination but also attempting to identify its causes. In Chapters 7 and 8, extend prior work by not only identifying discrimination in important contexts but also by attempting to reduce it. I describe each chapter in greater detail below.

In Chapter 2, I provide a general guide to conducting audit studies in political science research. I describe their basic design, and then provide advice on effectively carrying these studies out. My recommendations center on improving the external validity of audit studies.<sup>6</sup> In this vein, I offer suggestions about what sort of requests and names researchers should use in their studies, and how they can validate them.

In Chapter 3, I provide the first general introduction to conducting email audit studies. I describe the steps involved from experimental design to empirical analysis. I then offer 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

---

<sup>6</sup>This chapter is adapted from a working paper co-authored with Dan Butler.

want to conduct an email audit study. I conclude by suggesting several ways that email audit studies can be adapted to investigate a broader range of social phenomena.

In Chapter 4, I provide important guidance on how to select names for audit studies by illuminating a key variable plausibly related to racial perceptions of last names — geography.<sup>7</sup> I show that the probability that any individual belongs to a race is conditional not only on their last name but also on surrounding racial demographics. Specifically, I demonstrate that the probability of a name denoting a race varies considerably across contexts and that this is more a problem for some names than others. This result has two important implications for audit study research: it suggests limitations for (1) the generalizability of audit study findings and (2) for the interpretation of geography-based conditional effects. These implications mean that researchers should be careful to select names that consistently signal racial groups regardless of local demographics. I provide an R package that can help researchers do this.

In Chapter 5, I apply audit studies to measure discrimination by county election officials during the 2016 election cycle.<sup>8</sup> I demonstrate that the bias toward Latinos observed during the 2012 election has persisted. In addition to replicating previous results, I show that Arab/Muslim Americans face an even greater barrier to communicating with local election officials, but I find no evidence of bias toward Blacks. A design innovation allows me to measure whether emails were opened by recipients, which I argue provides a direct test of implicit discrimination. I find evidence of implicit bias toward Arab/Muslim senders only. This chapter extends existing research on racial biases in American politics, which has typically focused on legislators and other elected officials, by identifying these biases among a consequential group of bureaucrats. Perhaps more importantly, it moves beyond prior audit studies in political science by attempting to identify the mechanisms that drive

---

<sup>7</sup>This chapter is adapted from Crabtree and Chykina (2018).

<sup>8</sup>This chapter is adapted from Hughes et al. (2019).

observed biases.

In Chapter 6, I also look at the mechanisms driving biases by bureaucrats, but turn to examining them in the context of religious discrimination.<sup>9</sup> Despite growing descriptive evidence of discrimination against minority religious groups and atheists in the United States, little experimental work exists studying whether individuals face differential barriers to receiving public services depending on their religious affiliation. Here I report results from a large-scale audit study of street-level bureaucrats in the American public school system. I emailed the principals of more than 45,000 public schools and asked for a meeting, randomly assigning the religious affiliation/non-affiliation of the family. To get at potential mechanisms, I also randomly assigned family belief intensity. I find evidence of substantial discrimination against Muslims and atheists. These individuals are substantially less likely to receive a response, with discrimination growing when they signal that their beliefs are more intense. On the other hand, protestants and Catholics face no discrimination unless they signal that their religious beliefs are intense. These findings suggest that minority religious groups and atheists face important barriers to equal representation in the public arena. They also suggest that one of the mechanisms driving discrimination is belief intensity.

In Chapter 7, I provide the results from an adapted audit experiment designed to test whether making local officials aware of their possible biases could reduce discrimination.<sup>10</sup> I find no evidence that my informational treatment influences discriminatory behavior. While the limitations of the experiment design might make it difficult to determine whether information alone can reduce bias, this study makes two important contributions. First, I replicate prior studies by showing that White, local, elected officials are less responsive to

---

<sup>9</sup>This chapter is adapted from a working paper co-authored with Steve Pfaff, Holger L. Kern, and John L. Holbein.

<sup>10</sup>This chapter is adapted from Butler and Crabtree (2017*a*).

Black constituents. That local officials exhibit biased behavior is particularly worrisome, as local government is often the level that most directly affects citizens' daily lives. Second, I provide several suggestions for future audit studies that draw from the strengths and weaknesses of this study's design.

Finally, in Chapter 8, I turn to the important topic of racial discrimination by the police. Unlike existing studies, which focus on explicit or implicit biases among the police, I argue that racial discrimination depends in a conditional way on the extent of egalitarian views among the police *and* the public. To test the implications of my theory, I conduct an innovative survey experiment with American law enforcement administrators and elected officials who oversee the police. As predicted, elected politicians exhibit less racial discrimination in law enforcement oversight when informed that the public supports racial equality in policing. Contrary to my theory, though, police do not react to perceived public demand for egalitarianism. Overall, my results suggest that public attitudes toward racial equality influence police discrimination only indirectly, through the institutions that monitor and check their power. This chapter contributes to the growing inter-disciplinary literature on the politics of policing by illuminating how public opinion shapes law enforcement outcomes.



**Part II.**

**Methods**

# Chapter 2. Audit Studies in Political Science

## 2.1. Introduction

Audit studies are sometimes also referred to as ‘correspondence studies’ or simply ‘field experiments’. These studies typically involve sending a message or making a request and then measuring how the receiver responds. The phrase ‘correspondence study’ references the communication aspect of these studies. The researcher sends messages, or correspondence, and then measures how the receiver responds. The phrase ‘audit study’ emphasizes the measurement aspect of these studies. In a financial audit, one looks at the books to see what is happening in a company. In an audit study, the researcher looks at the behavior of the responder to see if they are responding differently to certain messages (or senders) than they do to others. In this sense the researcher conducts an audit. Finally, the phrase ‘field experiment’ references the field nature of these studies and the corresponding gains in external validity. The researcher (should) design the study to reflect the type of messages

---

This chapter is adapted from a working paper co-authored with Dan Butler.

that are normally sent so as to get a measure of how the people being studied normally respond.

Audit studies grew in popularity as a means to study discrimination in housing and labor markets (Quillian et al., 2017; Wienk et al., 1979; Bertrand and Mullainathan, 2004*b*). The passage of legislation barring discrimination, as a result of the civil rights movement, was accompanied by an interest in measuring whether discrimination persisted (see discussion in Gaddis (2017*ba*), sexual orientation (Drydakis, 2014), religion (Adida, Laitin and Valfort, 2010; Pfaff et al., 2018), criminal record (Pager, Bonikowski and Western, 2009), and more (Gell-Redman et al., 2018*b*; Rivera and Tilcsik, 2016).

Audit studies have also been widely used by governments as a way to test for discrimination. In the 1960s, the U.K. parliament created Race Relations Board, which commissioned several studies, including Audit studies aimed at measuring levels of racial discrimination (Daniel, 1968). The tests uncovered discrimination and led to the passage of laws barring racial discrimination in housing and employment (Smith, 2015*a*). In the United States, the U.S. Department of Housing and Urban Development (HUD) conducted several audit studies over the years in order to measure levels of discrimination in the housing market. In addition to several studies that focused on specific cities (Johnson, Porter and Mateljan,

1971), HUD commissioned several national audit studies over the years (Quillian et al., 2017). The federal government’s decision to use audit studies to measure discrimination influenced on academics by signaling that these studies were an acceptable, effective way at measuring discrimination (see discussion in Gaddis (2017b)).

Audit studies can be thought of as part of the larger literature that we refer to as field measurement studies. We use the term field measurement studies to refer to research aimed at measuring the behavior of subjects in the field. At their most basic design level, these studies provide some sort of stimuli to the subjects being studied (e.g., a request for help) and then measure how the subjects respond. In audit studies, the researcher sends a communication, while varying some aspect of the sender (e.g., their race) and then measures how the receiver responds. The purpose of these studies is primarily to measure whether one group is being treated more poorly than another.

Other field measurement studies follow a similar design, though not always for the purpose of measuring discrimination. For example, in *Making Democracy Work*, Putnam, Leonardi and Nanetti (1994) conduct a field measurement study to test whether government in northern Italy is more effective than government in southern Italy. For that study, the authors send three requests for help to bureaucrats in the different regions. They then tracked how the requests are treated as a way to measure bureaucratic efficiency. Similarly, Butler, Karpowitz and Pope (2012) sent requests to politicians, varying whether it was about a service or policy issue. The goal was to learn whether politicians put more effort into one type of activity over the other. The purpose was to not to change these politicians’ behavior, but rather to measure that behavior. The same is true of other field measurement studies.

In recent years, several studies have conducted similar studies to measure levels of censorship. King, Pan and Roberts (2014) used a field measurement study to learn about

social media censorship in China. For that research, the authors created numerous accounts on social media sites. The authors then randomly submitted different texts to see which messages would be censored and which would not. Through their approach, the authors were able to gain insights into what types of messages were being censored. Crabtree, Fariss and Kern (2015) similarly used a field measurement study to assess what messages private media firms in Russia censored. They did this by asking firms to publish an ad, randomizing the content of the ad.

Findley, Findley, Nielson and Sharman (2014) have conducted field measurement studies to make a significant contribution to understanding compliance with international law. The authors look at the formation of anonymous shell companies. While anonymous shell companies have been used for legitimate purposes, they also are a way that criminal and terrorist elements finance their operations. In response to concerns about shell companies being used to facilitate bad behavior, most countries in the world have signed on to international standards that require a notarized photo ID from the actual company owners. The authors conducted their study to learn whether the individuals who provide incorporation services are abiding by the international agreement. To do so, they posed as consultants seeking to form anonymous shell companies. They approached thousands of services that help clients form these companies and found that a large number of providers are willing to provide the service without requiring the required identification documentation. Their study also provided numerous insights into factors that lead to more or less compliance with the international standard. These studies highlight how audit studies are part of a large set of studies aimed at using similar methods to measure what is happening in the real world.

The advantage of audit studies, and all field measurement studies, is getting an externally valid measure of the behavior of subjects under study. One concern about surveys is that the responses can be cheap talk. This is especially true when the topic under investigation

involves behavior that is socially unacceptable. Findley, Nielson and Desposato (2016), for example, conducted follow up interviews with some of the same people who had been part of the field experiment they had done. In that survey they asked respondents what documentation they would require if someone asked for help in creating an anonymous shell company. Because these same individuals were part of the field measurement study they had conducted, they could compare the survey responses to the individuals' actual behavior.

The results of their study show that the survey results overstate the level of compliance. There are two reasons why the survey results understated the level of bad behavior. First, some of the worst offenders did not complete the survey. Second, the people responded to the survey, provided self-reports that were better than their actual behavior. In other words, respondents systematically underreported their own bad behavior.

These same two factors are likely to be a factor in any survey of discrimination. First, the people who are most discriminatory may be more likely to opt out because they know that their behavior is wrong. If we try to draw conclusions based on the people who opt in, we can get results that under estimate the level of bias. Field measurement studies avoid this issue by getting the full population or a random sample of the population of interest.

Second, social desirability is likely to be a big issue in these contexts. Discrimination, the focus of most audit studies, is the type of topic which is likely to suffer from social desirability effects. If we ask people about their own discriminatory behaviors and attitudes, they are likely to present themselves as better than they are. Audit studies avoid this potential pitfall by looking at their behavior when they do not realize they are being studied (and thus cannot artificially change their behavior to look less biased to the researcher than they actually are). The audit study, if well done, captures behavior in action, unaffected by social desirability bias.

Finally, audit studies have a relatively clear interpretation that make them an attractive tool for studying the how officials treat individuals from various groups. Many of the audit studies in political science have been used to compare how public officials treat different groups of individuals (see review in Costa (2017)). Political science is broadly interested in questions related to the equality of how groups are treated. Numerous studies on representation, for example, try to answer whether politicians give preference to one group over another. However, studying whether politicians represent a group's opinion on the issues better than they represent another group can be really difficult. However, in the context of an audit study, there is a straightforward way to measure if the politicians are being equally responsive: Do they put the same effort in responding? Everyone who sends a message wants a response. We can directly measure if they get a response. This is not to say that we shouldn't continue to look at other forms of responsiveness and representation. Rather the point is that this one feature that makes it easier to make a clear interpretation of how the officials respond.

## **2.2. Basics of Audit Studies**

Most audit studies involve testing for discrimination in how a group responds to some type of request (e.g., an email seeking help, a job application, a housing application, etc.). Audit study designs generally involve the following simple steps:

- Identify the question and sample.
- Develop the instrument(s).
- Send messages.
- Measure the outcome by looking at responses.

To illustrate these steps, consider studying whether bureaucratic offices exhibit racial discrimination against blacks relative to whites. Researchers could study this question by sending email requests for help to different bureaucratic offices, randomizing whether the request comes from someone who is black or white. They could then measure how the office responds to these requests to see if the offices are less responsive to requests from blacks. By following these steps, researchers can measure levels of discrimination. In the rest of this section we highlight a few of the major decisions that go into following these steps.

### **2.2.1. Be precise about the question and the sample**

Audit studies are well suited for studying discrimination. We follow Pager and Shepherd (2008) and define discrimination as the way a group is treated. Discrimination, which involves behavior, is distinct from holding racist attitudes, beliefs, or ideology. All of these other factors can motivate behavior. Studying discrimination does not presume what is causing the unequal behavior (see discussion in Pager and Shepherd 2008), though a promising direction of future work would be to examine those causes.

Most audit studies will focus on measuring whether a group is engaging in some form of discrimination. In our running example, the researchers are interested in testing whether bureaucratic offices are less responsive to blacks than whites. An audit study is appropriate for this question because it is focused on the behavior of the people in offices.<sup>11</sup>

It is also important to be precise about the sample. In many existing studies, researchers contact legislative offices (Butler and Broockman, 2011*a*, 2009; Butler, 2014; Gell-Redman et al., 2018*b*). Even if the researchers use the legislator’s email address, these requests are likely to be dealt with by staff. In other words, these studies are not about legislators, but

---

<sup>11</sup>If the theoretical question regards the reasons for the differential treatment, the audit study may not be the most appropriate design. The researchers should consider other designs.



rather about legislative offices. This is not to say that these studies are not informative; rather it is important to be precise about who or what we are studying. These studies speak to the behavior of legislative offices, which can be informative about how legislators represent their constituents (Salisbury and Shepsle, 1981). As a researcher it is important to be clear about the sample and ensure that it the correct sample for the question of interest. In our running example, the researchers are interested in learning about how the bureaucratic offices deal with requests.

### **2.2.2. Develop the instrument (or message) to maximize the likelihood that it reflects a commonly encountered communication**

Audit studies are typically used to measure the level of discrimination that citizens face in life. This is best done by creating an instrument (or message) that people are likely to send. Instrument here refers to the thing that the researcher is sending to the people being studied. In a study of job market discrimination, the instrument would be the resume used to apply for jobs. In our running example, the instrument would be the email message that the researchers send to the offices.

It is crucial that the researchers develop the instrument so that the sample being studied approach the communication in the same way they normally would. If the people in the sample suspect that they are being studied, they may behave differently leading the researcher to make incorrect conclusions. This point is so important, that the next section is devoted completely to advice on how to do this.

### **2.2.3. Hold confounding factors constant**

Early audit studies involved having actors from different racial groups, apply for jobs or housing programs. The actors would apply in person and the researchers would see if the racial minorities were treated differently. One concern about these studies is that the actor who were racial minorities would differ from the white actors in systematic ways. If these differences were also relevant to the hiring or housing decision, then it would be possible that these confounding factors might be responsible for the differential treatment.

In order to avoid this potential criticism, researchers would identify people who were similar to begin with and then they would train the actors to respond in similar ways (Gaddis, 2018*a*). The goal was to minimize any potential confounding characteristics. However, because it is nearly impossible to deal with all potential confounders, skeptics raised concerns that any measures of bias were inaccurate.

In response to these criticisms, researchers have transitioned to sending messages by mail or email. This allows them to send messages that are identical except for in ways that researchers intentionally manipulate. Returning to our running example, researchers might send email messages to bureaucratic offices that are the same in every way except in the name of the sender.

## **2.3. Maximizing External Validity**

As mentioned above, researchers need to maximize the realism of their instrument. There are at least two important reasons for this. One is because researchers should conduct the study in a way that does not cause study subjects to be suspicious. If the study population suspects that the message that they receive is not typical, they might doubt the identity of the sender, and come to think that the message was sent by a researcher. This could

change how they respond, potentially biasing the results away from the very thing you want from a study like this. For example, perhaps a group normally responds more to whites than blacks. However, imagine that they suspect that researchers are studying their potentially discriminating behavior. In this case, they might then be more careful at responding to communications from blacks to the point where they respond more to them than whites. This could lead researchers to actually reach the exact wrong conclusions that blacks receive better not worse treatment.<sup>12</sup>

Another reason why researchers should maximize the realism of their instrument related to what we want to learn from audit studies. Typically, researchers conduct audit studies because they want to learn about how the average member of a group is treated in some interaction. If the study population suspects that the message that they receive is not typical, then they might think that the individual sending it is also atypical in some unknown way. This could cause them to treat the sender differently than the average member of the sender's group. The result of this would be that the researcher's findings would be biased in that they would not describe the experience of an average group member but of an atypical one.

### **2.3.1. Include typical requests in the instrument**

There are several areas in which researchers might be particularly concerned about the realism of their instrument. One area is the type of requests included in the instrument. Before determining what request(s) to make of the study population, researchers should ensure that these requests are similar to the ones that their subjects usually receive. They

---

<sup>12</sup>Another way of thinking about this issue is in terms of social desirability bias. When people know they are being surveyed they underreport discriminatory attitudes. If people know they are being audited — they will similarly adjust their behavior on those specific communications to look less discriminatory. The whole advantage of an audit study is to avoid this type of social desirability bias. If subjects believe that the message they receive is not a genuine, then this advantage is lost.

can do this in several ways. One approach is to conduct qualitative interviews with members of the study population about the work they do and the type of interactions that they have with the public (Terechshenko et al., 2019).<sup>13</sup> Researchers often conduct audit studies to examine how offices (legislative, bureaucratic, etc.) behave. When this the case, another approach that researchers can use is to use requests that appear in the frequently asked question sections of office websites. When the study population consists of public offices or officials but neither of these two approaches is possible, researchers might want to consider issuing Freedom of Information Act requests for all messages received by the study population. After receiving these text corpora, researchers could use machine learning tools to summarize them and to construct typical requests.

### **2.3.2. Use different aliases**

In the majority of audit studies, researchers create a set of fictitious identities and use these to send messages. Sometimes researchers use these identities to detect discrimination, sometimes they use them to conceal the fact that the messages come from researchers. The names that researchers use with these identities should be carefully selected. This is because each name signals a number of things about the identity of the sender. In the interests of maximizing the believability of the messages, researchers should select names that are typical among members of the individual identity’s group. For example, this means that first names like ‘Apple’ and ‘Nefertiti’ should probably be avoided.

In many cases, researchers use the name to signal the race, ethnicity, or gender of their fictitious identities. When researchers do this, they should check how the names they use are perceived. A growing literature suggests that names commonly assumed to strongly signal a specific racial or gender identity might only signal that identity weakly (Gaddis,

---

<sup>13</sup>Researchers might then want to exclude the individuals they interview from their study, as they might be more likely to think that the instrument was sent from a researcher.

2017*e,b*; Crabtree and Chykina, 2018). One important finding from this literature is that that perceptions of names vary across study populations based on subject race, education, and geography. Researchers have two options to deal with this potential problem. One is that they can use names that have been tested by other researchers. For example, Hughes et al. (2019) provides in their appendix a list of the names they used along with the results of a survey they conduct about popular ethnic perceptions of these names. Another approach is that researchers can pre-test their own names through platforms such as Amazon’s Mechanical Turk. In the context of our running example, one could do this by selecting a large bundle of names and asking Mechanical Turk workers (MTurkers) to assess the likelihood that the name belongs to a black or white individual.<sup>14</sup> The results from this exercise would allow researchers to select the names that are most strongly associated with black or white individuals.

### **2.3.3. Use multiple requests and names**

A potential issue for researchers is that one of their subjects might receive multiple messages or that their subjects share received messages with each other. This is potentially problematic if those messages are identical are identical in nearly all aspects. That might lead subjects to doubt the authenticity of the messages or discern the intentions of the researchers, effectively spoiling the experiment. A way that researchers can potentially get around this issue is by randomly varying aspects of the instrument, such as included requests and sender names. In our running example, we might send several different types of requests to bureaucratic offices and use several different names to signal black and white

---

<sup>14</sup>One potentially useful source for names is birth certificate or United States census data. These data sources indicate the prevalence of names among racial groups. One limitation of these data, though, are that they indicate only how common the names are among groups, and not how common individuals perceive the names to be among groups. Because subject perceptions might not match objective reality, we care more about the former than the latter when selecting names.

identities. By doing this, we would make it less likely that the messages would seem related to each other, which would decrease the chances of potential discovery.

There is a second compelling reason to use a range of requests and names. Researchers often want to claim that the results of their study are indicative of more general social phenomena. By using different requests and different names, researchers can help ensure that their results are not specific to any one request or name. For example, if researchers in our running example only use one black name, then they cannot be sure that any discrimination that they observe generalizes beyond individuals with that name.

#### **2.3.4. Use reasonable email or postal services**

Once researchers have developed an instrument, they need to deliver it. To do that, researchers typically create email or mail addresses for each identity that they use in their study. To maximize the believability of their intervention, researchers should use addresses that do not raise subject suspicions. This can be done by using common email or mail services, such as gmail.com or a post-office box. If researchers are using email to deliver their instrument, they should consider creating unremarkable email addresses. For example, if the name for one identity is ‘Jane Smith’, they might want to create the email address ‘jsmith872998gmail.com’.<sup>15</sup> In some cases, researchers might want to have their identities associated with a real or fictional organization. This might lead them to partner with organizations and use their email domains to increase the believability of their messages. If researchers are pretending to have their emails come from a fictional organization, they could register a domain name of this organization and create a basic

---

<sup>15</sup>If the researchers include numbers in their email addresses, they should think carefully about using sequences that do not necessarily indicate a birthdate, area code, zip code, or some other attribute of the sender. This means that researchers might want to avoid using 3- and 5-digit sequences that might reveal sender place and 2-, 4-, or 6-digit sequences that might indicate sender age.

webpage for it as well.<sup>16</sup>

### 2.3.5. Check the final instrument

Once researchers have a draft of the final instrument, they should perform two additional checks. The first is that they should try and ensure that their instrument is not the same as one used in a prior study. As a corollary, they should definitely not borrow parts of their instrument from previously completed studies. This can cause significant problems for researchers. As an example, White et al (2015) used a set of names and email addresses to determine if local election officials exhibited bias against Latinos in September 2012. Approximately 4 years later, <https://www.thedenverchannel.com/news/investigations/fbi-probes-emails-sent-to-county-clerks-across-colorado-and-12-other-states>. Some local election officials detected the similarities and posted a notice on a public bulletin board that individuals should not respond to these emails, in effect contaminating Ayres' study.

The second thing that researchers should do is ask several individuals who are or have been part of the study population to read the instrument and provide input.<sup>17</sup> In our running example, we could ask individuals who used to work at bureaucratic offices what they thought about our instrument. These exchanges between researchers and subject population can help identify issues with the instrument or suggest new ways of improving it or the broader experimental design.

---

<sup>16</sup>If researchers take this approach, they also might want to consider adding using website analytics to determine the extent to which subjects the site. Many visits might indicate that subjects found the instrument atypical in some way.

<sup>17</sup>These individuals should be excluded from the study.

## 2.4. Additional Design Considerations

### 2.4.1. Internal validity

As we discuss above, the internal validity of any empirical claims made from audit studies depends on the subject pool not knowing that they are the participants in an experiment. Just as importantly, the internal validity of these claims also depends on the identities used by the auditors appearing identical (in both observed and unobserved ways) to participants, with the exception of whatever attributes researchers intentionally manipulate (Heckman, 1998*a*). When identities do not otherwise appear identical, then researchers cannot be sure that any discrimination that they measure is related to the characteristics that they manipulate or to some other characteristics that correlate with them and might vary across identities.<sup>18</sup> These two issues are part of the reason why most audit studies are conducted via correspondence now, rather than in person (Neumark, 2012; Gaddis, 2018*a*). For example, by sending messages to subjects, researchers have more control in how they construct identities, making it easier to create similar auditee profiles. In our running example, it would be more feasible to create two constituents who are the same except for their race than to find a white person who is interchangeable with a black person in every other way except for their race.

Thankfully, there are a variety of ways in which we can empirically assess the extent to which identities might appear the same. One approach is to pretest the different identities with some survey population, such as MTurkers (Gaddis, 2018*a*). The idea here is to ask respondents a series of questions about each identity's observed and unobserved characteristics. If the responses indicate that the only differences relate to the manipulated

---

<sup>18</sup>Another way of thinking about this is that the results from audit studies depend on the excludability assumption that the manipulated characteristic drives differences in how subjects respond and not some other characteristic (Butler and Homola, 2017*a*; Gerber and Green, 2008).



characteristics, then researchers can be more confident that they have not failed to set some attributes constant.

### **2.4.2. SPAM Concerns**

One potential concern with conducting audit studies via email is that messages might be automatically marked as SPAM. This would mean that subjects might not receive their assigned treatments. This would potentially be very problematic if certain experimental treatments or treatment combinations were more likely to be identified as SPAM. The issue here is that this would decrease the probability that messages with those treatments would receive a reply, potentially leading researchers to believe that bias exists where it does not.

To help guard against that possibility, researchers should test their messages with SPAM classifier software. This type of software estimates the probability that a message would be marked as SPAM. While there are free online tools that can do this, they are often slow and require manual input, which could lead to errors. We have developed an R software package, `spamcheck`, that allows users to query a number of SPAM classifiers and returns a set of probabilities that they can use to evaluate their messages. The package is available on <https://github.com/cdcrabtree>.

### **2.4.3. Email tracking**

## **2.5. Ethical and Other Considerations**

In addition to the design considerations we have discussed, authors should also consider the interests of the research subjects. We should adjust the design when needed to minimize any harm to subjects. And if there is sufficient cause for concern (and no way to mitigate those concerns) we should not conduct the research. This is true even if the research

receives approval from the institutional review board (IRB).<sup>19</sup> Taking steps to mitigate potential concerns is in the interest of research subjects and our research community.

The ethical considerations of all projects need to be evaluated on their own merit. However, there are some concerns that frequently arise with audit studies. First, research subjects may have to waste significant resources to respond. Even if this is not true for any one individual — some replies might take only a minute — it might still be true in the aggregate. Second, there is a concern about embarrassing study subjects by releasing their information. Third, there is a concern about the fact that researchers who conduct audit studies do not ask subjects for their consent.

In trying to deal with these and other issues, our own advice is that researchers think about what the worst thing is that could happen if they implement their study. What is the maximum level of regret they could have? With this in mind, they should then make design changes that minimize that potential maximum level of regret. For instance, a researcher might want to conduct a study of election officials in advance of an election. The worst thing that might happen here is that the intervention causes election officials to have less time for registering voters, resulting in lower voter turnout. To minimize the maximum level of regret, the researcher might conduct the study during a non-election period.<sup>20</sup> We can think about this rule of thumb as we discuss the three issues raised above. Researchers can also apply the same rule of thumb as other project-specific issues arise in their own research.

One common concern is that researchers are wasting public officials' time. This concern is one reason why many audit studies include really simple requests in their instruments.

---

<sup>19</sup>As Driscoll (2015) points out, IRBs exist principally to minimize the legal liability of colleges, and not to carefully evaluate the ethics of all studies. Since researchers know comparatively more about their subjects and their study design, they must actively police their own work.

<sup>20</sup>It might be less interesting, of course, or less relevant theoretically to examine biases in election official behavior at non-election times. This example highlights the trade-off that researchers must face in minimizing the maximum level of possible regret.

We think that researchers should continue this general practice and keep time-consuming requests to a minimum, even if this imposes constraints on what can be studied.

Another thing we can do more often is to recruit real people to send the messages. Butler, Karpowitz and Pope (2012) conducted an experiment where they asked students to write to their members of Congress and their state legislator. Not only did the students write the letters themselves, they were given the responses that the researchers received. It is true that these students would probably not have written their letters without encouragement from the researchers; however, we think that encouraging people to communicate with their elected officials is generally a good thing. One could easily imagine that being part of a homework assignment in a class. Indeed, this is one way that we as researchers might implement this suggestion; by asking students in a class to participate by writing a letter where something about the communication is randomized.

While we think that researchers should use real people when possible, the reality is that researchers will not use this approach unless reviewers reward it. This should not be a decisive factor for a paper's fate. We should not simply accept an audit study because it uses real people (nor should we reject an audit study because it does not). Rather, we are advocating that using real people should be a positive consideration when evaluating a paper for publication.

Embarrassing the research subjects is one of the other concerns that is common to most audit studies. We must guard against this. The goal of the people is not to embarrass specific people. The goal is to identify problems so they can be improved. Perhaps the single most important way we can mitigate this concern is by maintaining confidentiality. This is a simple solution, but it is vital. Messing this up, even once, could have very negative effects not only for individual researchers, but for the study of political elites, and perhaps even the field as a whole.

The third factor we raised above — consent — is the hardest issue. For all of the reasons we have laid out above, an audit study is most useful when people do not know they are being studied. If we have to get consent, the worst offenders could opt out and/or people might change their behavior to act better than they normally do when interacting with others. Requiring consent would ruin the usefulness of most audit studies. The concerns about not getting consent, along with other potential issues, has to be weighed alongside the benefits of the study. We think that if other concerns are sufficiently minimized, and the articulated benefit is clear, then many audit studies are still worth pursuing. Identifying bias and studying ways to minimize it is an important societal benefit. Indeed, the recent rise of exclusionary political rhetoric and the concurrent increase in hate crimes throughout the United States and Europe, suggest a profound need for social science research that helps diagnose and eliminate discrimination.

Because of these concerns, all researchers should ask whether they need to conduct an audit study to answer their question of interest. In some cases, researchers can answer their question by reanalyzing the results of a previous audit study. If so, then the researchers should approach the question in that way. In practice this can be done by researchers sharing their data in ways that maintains the research subjects' confidentiality. Ideally, these deidentified data would be shared on the Dataverse or a GitHub repo, so that others might have wide access to them and be able to easily build on prior work.

# Chapter 3. Email Audit Studies

## 3.1. Introduction

What is an audit study? An audit study (or correspondence study) is one way of assessing hard to observe behaviors, such as discrimination (Heckman, 1998*a*). 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 the identities, such as race or gender. Next they use these identities to accomplish some task, like applying for jobs, renting housing, or contacting legislators. These tasks can be 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 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 against Blacks.<sup>21</sup> 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*a*; Grose, 2014; Costa, N.d.), in housing transactions (Gaddis and Ghoshal, 2015; Turner et al., 2002; Hogan and Berry, 2011; Oh and Yinger, 2015), in

---

<sup>21</sup>The other chapters in this volume describe these studies in greater detail.

economic exchanges (Riach and Rich, 2002), in employment decisions (Neumark, Bank and Van Nort, 1995; Bertrand and Mullainathan, 2004*a*), 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.<sup>22</sup> 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.

Another 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, 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

---

<sup>22</sup>Some recent examples of this include Gaddis (2014); Gaddis and Ghoshal (2015); Sharman (2010); Radicati and Hoang (2011); Oh and Yinger (2015); Milkman, Akinola and Chugh (2015, 2012); Lahey and Beasley (2009); Hogan and Berry (2011); Giulietti, Tonin and Vlassopoulos (2015); Findley, Nielson and Sharman (2015); Bushman and Bonacci (2004); Butler (2014).

correspondence studies.

Despite these advantages, email audit studies are perhaps underused. One reason for this could be that they are often difficult to implement, particularly for scholars who are inexperienced with conducting audit studies. Surprisingly, there are no general introductions to the approach. Another reason why email audit studies might be underused is because scholars might think that they can only examine a narrow range of social phenomena. While the vast majority of email audit studies have focused on unearthing evidence of racial, ethnic, or gender 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.<sup>23</sup> 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](http://charlescrabtree.com/email_audit) and [auditstudies.com](http://auditstudies.com).<sup>24</sup>

---

<sup>23</sup>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 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.

<sup>24</sup>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.

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.

## 3.2. 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 Figure 3.2. They include (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.<sup>25</sup>

### 3.2.1. Experimental Design and Sample Selection

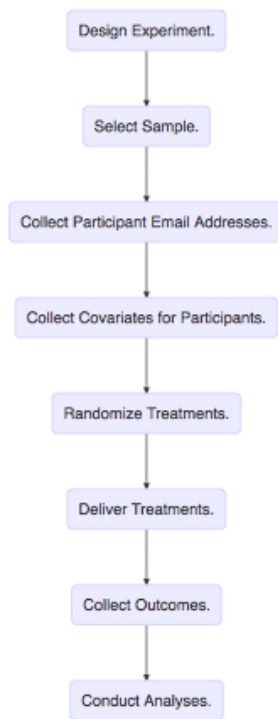
While each of these steps is extremely important, I do not discuss the first two. Many excellent texts deal with issues related to design and sampling (e.g., Gerber and Green (2012); Lohr (2009)). I refer interested readers to them.

Regardless of what researchers decide regarding experimental design and sample selection, they should consider pre-registering these choices, along with their theoretical ex-

---

<sup>25</sup>One additional stage not discussed here is getting institutional approval, typically provided by an institutional review board (IRB), for conducting the intended study (Driscoll, 2015). Gaddis (2017a), Riach and Rich (2004), and Yanow and Schwartz-Shea (2016) provide excellent guides that can help scholars navigate IRB concerns and ethical issues.





**Figure 3.1:** The eight stages of a typical email audit study.

pectations and analytic strategy (Olken, 2015; Franco, Malhotra and Simonovits, 2014).<sup>26</sup> There are many possible reasons to write a pre-analysis plan.<sup>27</sup> 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 American Economic Association’s RCT Registry, Evidence in Governance and Politics, or on personal academic webpages.

### **3.2.2. 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. 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 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

---

<sup>26</sup>Lin and Green (2015) offer detailed advice on some of these decisions.

<sup>27</sup>Coffman and Niederle (2015) discusses some of the limitations of pre-analysis plans.

list. This is particularly true in the case of public figures. Sites like [everypolitician.org](http://everypolitician.org) and [sunlightfoundation.com](http://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 previously been 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](http://github.com) or [stackoverflow.com](http://stackoverflow.com), have made it so that even individuals inexperienced with programming can adapt existing scrapers to their own purposes. The online appendix to this chapter contains code to build a minimal scraper, as well as instructions on how it might be modified to scrape other websites.

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](http://elance.com) and [guru.com](http://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 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.

### **3.2.3. Covariate Collection**

Researchers might gather covariates 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. Another is to include it in the randomization scheme. In many cases, scholars can use the same techniques to collect covariates as they do to collect email addresses.

### **3.2.4. 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. I typically use the R package `blockTools` for this (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.

### **3.2.5. 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, 2017a). 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.

### 3.2.6. 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 untenable solutions for researchers who want to send their emails through programming scripts.

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 to \$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., `thefastestmailever.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). Unfortunately, there is not a clear solution to this problem, and researchers simply have to evaluate the advantages and disadvantages of each approach within the context of their experiment.

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 text. For example, scholars might need to assign all observations with the value 1 in the `treatment` column to the text ‘string’. Scholars should also create the strings of text that comprise the non-random email components, such as email valedictions or salutations.<sup>28</sup> After that, scholars will need to write the code that combines the random and non-random strings of code into a complete email. The online appendix for this chapter includes R code for both steps.

Finally, 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

---

<sup>28</sup>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, 2017a).

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 an a ‘time to reply’ outcome measure, as I discuss later. Finally, 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.

### **3.2.7. 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 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.<sup>29</sup>

---

<sup>29</sup>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 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.

### **3.2.8. Outcome Collection**

Having sent emails, scholars can begin collecting outcomes measures. The primary outcome of interest in email audit studies is typically a binary indicator that is coded 1 if participants replied and 0 otherwise (e.g., Butler (2014), Bertrand and Mullainathan (2004a), 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 problem with this approach, 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.

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 most 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 email responses to observations in the dataset. 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. They might, for instance, 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).

### **3.2.9. 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).

## **3.3. Extending Audit Studies**

As noted above, scholars across the social sciences are increasingly using email audit studies to test for discrimination against ethnic, racial, political, and religious groups. Their efforts have resulted in the accumulation of new empirical findings that have both encouraged additional theoretical development and stimulated additional empirical work. While email audit studies have been used productively to examine some questions, I want to suggest several ways that researchers can adapt the approach to examine different social phenomena.

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 (2017a) 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*a*) 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.

### **3.4. Discussion**

In this chapter, I have discussed how researchers can conduct email audit studies and how they can be adapted to address a broader range of possible questions. Specifically, the first section of the chapter provided both a general outline on how to implement an email audit study as well as detailed directions about collecting, sending, and coding emails; the second section discussed several recent non-traditional email audit studies and examines how the general study design might be adapted to answer a broader range of questions. 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.

# Chapter 4. Name Selection in Audit Studies

Since Bertrand and Mullainathan (2004*a*)’s pioneering study, hundreds of researchers have conducted audit studies to investigate the extent of racial (or ethnic) discrimination in America across myriad contexts (Crabtree, 2017).<sup>30</sup> The results from these studies have done much to advance scholarly research on discrimination across the social sciences (Pager and Shepherd, 2008; Crabtree and Fariss, 2016). For this reason, papers that center on audit studies often garner tremendous attention within the research community and even in the broader public, where they have helped deepen public understanding about the serious barriers that racial (or ethnic) minorities face in nearly every aspect of common life.

There are reasons, however, to be at least somewhat concerned about the findings from these studies.<sup>31</sup> One of the largest concerns relates to the first and last names that researchers use to signal racial (or ethnic) identities in America (Butler and Homola, 2017*a*).

---

This chapter is adapted from Crabtree and Chykina (2018).

<sup>30</sup>Gaddis (2017*d*) provides a brief history of this growing literature. While we focus on studies on racial (or ethnic) discrimination conducted in America here (e.g., Butler (2014), Gell-Redman et al. (2018*b*)), scholars are increasingly conducting audit studies in other countries and to detect other types of biases (e.g., Adida, Laitin and Valfort (2014); Ahmed, Andersson and Hammarstedt (2012, 2013*b*); Neumark, Bank and Van Nort (1995); Baert (2016)).

<sup>31</sup>Pager (2007*a*) reviews and addresses the major criticisms of audit studies.

Researchers typically select these names based on either (a) those used in prior audit studies or (b) government-provided lists that contain population-level statistics of use across races Gaddis (2017c). The potential problem here is that scholars often ignore the extent to which these choices accurately map onto how individuals perceive names. In his welcome and long-overdue study on that subject, Gaddis (2017c) demonstrates that current practices do not acknowledge or take into account the many different factors that shape public perceptions of given names. His findings have obvious implications for how scholars choose *first names* to represent racial identities.

Perhaps one of the most important findings from Gaddis (2017c) is that individual perceptions of first names can change considerably depending on the last names with which they are paired. Specifically, he shows that the last names researchers use in their audit studies can substantially strengthen (or weaken) the extent to which individuals perceive first names as belonging to specific racial (or ethnic) identities. In a separate, related paper, Gaddis presents additional evidence that this is the case (Gaddis, 2017e). This means that researchers should think carefully about the *last names* they pair with first names.

Unfortunately, as Gaddis (2017c) acknowledges, researchers often don't. He and others typically classify last names by race based on population-level usage statistics (Butler, 2014; Butler and Crabtree, 2017a; Gaddis, 2017c,e). This practice, however, relies on the assumption that individual racial perceptions are in line with the country-wide popularity of last names among racial groups.

In this paper, we contend that this assumption ignores the crucial importance of geography and local demographics. We show that the probability that any individual belongs to a race is conditional not only on their last name but also on surrounding racial demographics. This result has two important implications for audit study research: it suggests important

limitations for (1) the generalizability of audit study findings and (2) for the interpretation of geography-based conditional effects. This means that researchers should be careful to select names that consistently signal racial groups regardless of local demographics.

This paper proceeds as follows. First, we describe how researchers typically select last names for audit studies, laying bare a key assumption behind these choices. Second, we explain why this assumption is too restrictive in many cases, as it requires individuals to ignore information about local demographics when inferring the race of an individual based on their name. Third, we present and justify an alternative assumption that allows individuals to combine information about name popularity and racial context. We contend that this assumption more closely mirrors how individuals infer the race of others. Fourth, we illustrate why these assumptions matter by showing that the probability of a name signaling a racial group varies across geographic contexts. Finally, we close by outlining several implications of this finding and introducing an open-source software solution that researchers can use to assess how the racial meaning of names varies across geographic locations.

## 4.1. Last Name Selection

Researchers generally select names for audit studies using one of two strategies. The first is to use names reported in other audit studies. This is a very common approach, made easier by evolving norms of transparency in social science research, which increasingly require that researchers make available details like this. The second strategy is to use population-level name lists, such as those provided by the United States Census. These lists contain a series of last names and report their frequency of use across racial groups.<sup>32</sup>

---

<sup>32</sup>It is probably the case that scholars who adopt the first strategy are often indirectly adopting the second strategy. This is because some of the earliest audit studies relied on these population-level lists (e.g., Bertrand and Mullainathan (2004a)).

The problem with using population-level data, though, is that it requires researchers to assume that people evaluate names based on the national popularity of those names within racial groups. In other words, individuals assign the probability,  $P$ , that an individual belongs to a race,  $R$ , based on the frequency with which members of that race use a last name,  $S$ , in the population.<sup>33</sup> Formally, this means that  $P(R|S)$ .<sup>34</sup>

In some contexts, this might be a reasonable assumption. We can think, however, of many cases in which individuals likely make inferences about a person’s race based not only on their surname but also on local demographics. This is because last names can signal different races in different places. Elliott et al. (2009) provide a couple powerful examples where this is the case.

“Persons with the common surname ‘Lee,’ for example, are likely to be Korean or Chinese if they reside in a predominantly Asian neighborhood but not if they live in, say, Williamsburg, Virginia. Likewise, the Asian surname ‘Ohara’ could easily misidentify persons living in predominantly Irish neighborhoods.” (4)

The obvious implication of this is that individuals in a predominantly Asian neighborhood are more likely to think that someone with the last name ‘Lee’ is Asian than White. Likewise, individuals who live in neighborhoods with a large proportion of Irish are more likely to think that a person with the name ‘Ohara’ is Irish than Asian.

While these are perhaps exaggerated examples, they highlight the fact that in some contexts individual associations between names and race might be conditional on local racial demographics, or location, which we denote as  $L$ .<sup>35</sup> This could happen when the

---

<sup>33</sup>We adapt our notation from Imai and Khanna (2016).

<sup>34</sup>Researchers typically assume that  $P$  equals either 0 or 1. We assume instead that  $P$  is bound from  $0 - 1$ . This means that individuals think that some names are more likely to be held by members of different races, but does not require that they have perfect confidence about the correspondence between names and races. Our general argument, however, does not depend on this assumption.

<sup>35</sup>This seems particularly likely in the context of audit studies, which often focus on local interactions, such as those between putative constituents and their elected representatives.



subject of these studies might assume that the fictitious individual contacting them lives within their community, broadly defined. In light of this, we think that assuming  $P(R|S, L)$  is more reasonable than assuming  $P(R|S)$ . In the next section, we empirically illustrate what happens when we incorporate information on  $L$  when classifying last names.

## 4.2. Data and Results

To calculate  $P(R|S, L)$ , we follow a well-developed healthcare literature and use Bayes rule (Elliott et al., 2008, 2009; Fiscella and Fremont, 2006; Imai and Khanna, 2016).<sup>36</sup> This approach provides us a “probabilistic prediction of individual [race or] ethnicity” (Imai and Khanna, 2016, 265) for a given surname in a geographic area. In the context of this paper, we interpret this quantity as representing the extent to which an individual believes a person belongs to a racial (or ethnic) group.

Before applying Bayes rule, we need several inputs. First, we need a selection of last names. We take our list of 20 surnames from (Gaddis, 2017c). Second, we need a set of spatial units. For this application, we use the population of 3,142 US counties.<sup>37</sup> Data for this is supplied by the Census API. Third, we need population-level data on surname use across racial groups as well as the racial demographics of counties. We use the R package `wru` to dynamically call these data.

We also use `wru` to apply Bayes rule and generate the predicted probability of a name signaling a race (or ethnicity) (Imai and Khanna, 2016) for all 3,142 US counties. We generate these probabilities for three racial (or ethnic) groups — Blacks, Hispanics, and Whites. We then plot the results of this exercise in Figure 4.1. The vertical axis lists the 20 names we used. The horizontal axis displays the distribution of the probabilities

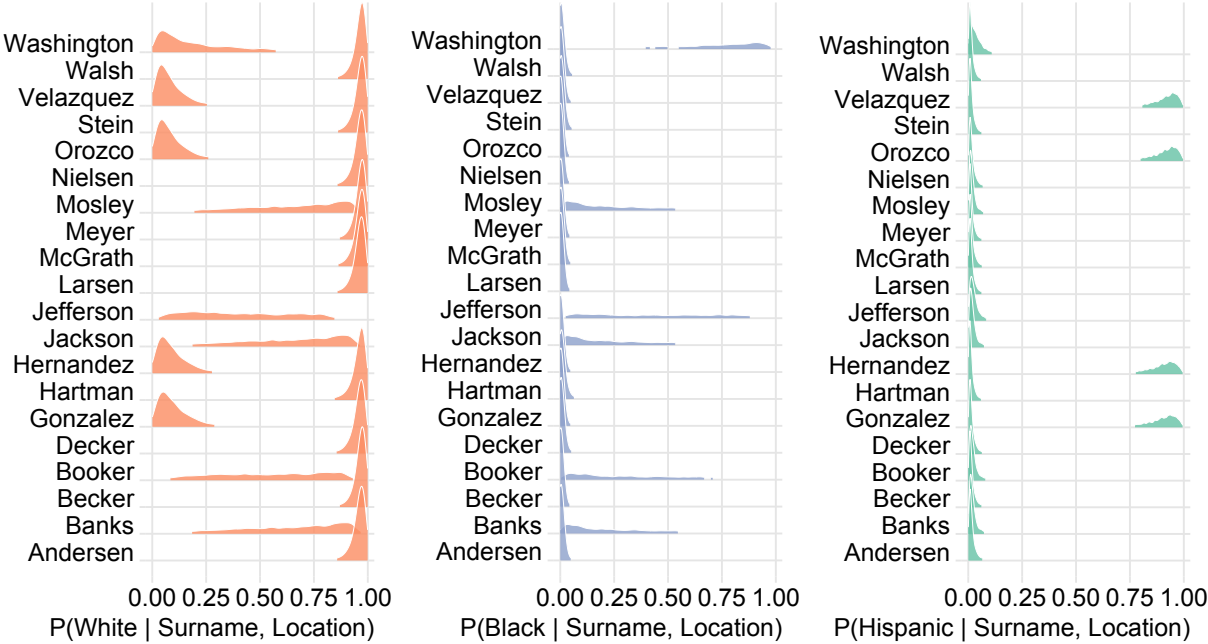
---

<sup>36</sup>Imai and Khanna (2016) provide an accessible introduction to the math involved.

<sup>37</sup>Researchers, however, could adapt our approach to other geographical areas, such as Census blocks, tracts, or voting precincts (Imai and Khanna, 2016).

generated using Bayes rule across US counties. The probabilities of a name belonging to a race sum to 1. The left panel plots  $P(White|S, L)$  for each name, while the middle panel plots  $P(Black|S, L)$ , and the right panel plots  $P(Hispanic|S, L)$ .

**Figure 4.1:  $P(R|S, L)$  Plots**



*Note:* Figure 4.1 shows the predicted probabilities of a name signaling a race (or ethnicity) across geographic locations. The vertical axis lists 20 last names commonly used in audit studies. The horizontal axis displays the distribution of probabilities generated using Bayes rule. The probabilities of a name belonging to a race sum to 1. The left panel plots  $P(White|S, L)$  for each name, while the middle panel plots  $P(Black|S, L)$ , and the right panel plots  $P(Hispanic|S, L)$ .

If a last name consistently signals a race (or ethnicity) across geographic contexts, we would expect the distribution of probabilities for that name to be closely grouped around 1. Similarly, if a name consistently *does not* signal a race (or ethnicity) across geographic contexts, we would expect the distribution of probabilities to tightly center on 0. We see, however, that either of these cases are rare, characterizing only about half the last names in our sample. Examples of this include Walsh, Nielsen, and McGrath. Regardless of where these names appear, individuals are likely to think that they belong to White individuals.

Importantly for audit study work, many names do not appear to consistently signal race across geographic contexts. Indeed, several of the names would seem to be particularly bad at signaling race no matter where they might be used. This can be seen in the low flat distributions of probabilities seen for some of the names, such as Mosley, Jefferson, and Jackson. Last names like these appear not to provide any additional information about the race of their bearer.

One could argue, though, that individuals do not form their racial perceptions based on last names alone but on first names as well. This is objectively true. For instance, individuals are more likely to think that a man with the last name ‘Washington’ is Black if his first name is ‘Jamal’ than if it is ‘Chad’. Yet, the racial signal sent by first names is rarely so clear, as Gaddis (2017*c*) and Gaddis (2017*e*) show. When first names do not clearly signal a race (or ethnicity), they are unlikely to help boost the signal sent through a last name, making it necessary for last names to convey a racial signal on their own.

### **4.3. Discussion**

What are the implications of the empirical finding introduced above? We can think of at least two noteworthy consequences. The first has to do with generalizability. As a reminder, Gaddis (2017*c*) and Gaddis (2017*e*) show that first names alone often provide an insufficient signal about racial identity, causing people to rely on the information provided in last names. We show, though, that the racial (or ethnic) information provided by surnames varies across geographic context. Taken together, this suggests that the racial perception of some full names (first and last names) likely changes across space, leading to geographic variation in treatment intensity. If this is true, then it means that the effect of racial cues based on these names likely varies geographically as well. This calls into question the extent to which previous audit study findings travel across spatial contexts,

particularly when those studies are conducted at a local or regional level (e.g., Wallace, Wright and Hyde (2014)).

The second implication has to do with treatment effect heterogeneity. Researchers often want to investigate the extent to which their impact estimates vary across spatial domains. For example, political scientists might want to know whether the effect of being Black is different in the South. If individual perceptions of names are influenced by place, though, then it is not entirely clear how researchers should interpret statistically significant interactions between location and racial (or ethnic) treatments. The problem here is that scholars cannot know if these interactions are the result of some contextual-level process, the fact that the treatment varies across space, or both.

## 4.4. Conclusion

So what can scholars do about these issues? We recommend that prior to using a surname in an audit study, researchers should first examine the extent to which the probability that the name denotes a racial group varies across geographic contexts. When possible, scholars should select names where the probability varies across a limited range. This will help ensure that individuals perceive the race of the putative individual as intended.

We provide a slim R package, `auditR`, available at <https://github.com/cdcrabtree/auditR> to help researchers accomplish this. Relying heavily upon the `wru` package provided by Imai and Khanna (2016), the software takes a vector of last names, generates a matrix of name and county pairs, uses this matrix to return the probability that a name denotes one of four racial (or ethnic) groups (i.e. Asian, Black, Hispanic, and White) for all spatial units, and then plots these values. This allows individuals to visually identify the extent to which the racial information provided by surnames varies across geographic contexts and to identify potentially problematic surnames.

While this package does not solve all the potential problems that researchers might face when selecting appropriate last names for their audit studies, we think that it helps address a potentially important problem with current practices. Scholars can build on this work — and on Gaddis (2017*c*) — by examining how other factors condition individual perceptions of last names.

## **Part III.**

# **Applications**

# Chapter 5. Persistent Bias Among Local Election Officials

Racial bias that limits access to the ballot threatens basic principles of democratic equality. One potential source of bias that has received little attention are the street level bureaucrats who administer elections in the U.S. (Lipsky, 1980). An audit study conducted during the 2012 U.S. election cycle showed these local election officials responded at significantly lower rates to inquiries from voters with putatively Latino, as opposed to white, surnames (White, Nathan and Faller, 2015*a*). In this paper we report the results of a similar audit study performed during the 2016 election cycle. We find that the previously observed bias against Latinos is persistent. We also extend the previous study by testing the effects of two racial primes other than Latino. Voters with Arab/Muslim names received responses at significantly lower rates (11 percentage points) than whites, while black voters did not.

The two primary motivations for this study are to determine whether the previous finding of bias toward Latinos stands up to replication, and to examine whether this bias extends to blacks and Arab/Muslim Americans. In spite of the ample evidence of racial disparities

---

This chapter is adapted from Hughes et al. (2019).

in political participation (Hajnal and Lee, 2011; Abrajano and Alvarez, 2010; Hajnal and Abrajano, 2015; García-Bedolla and Michelson, 2012) and in every-day life (Bertrand and Mullainathan, 2004*a*), relatively little empirical work demonstrates the role of race in limiting access to the ballot in contemporary America (McNulty, Dowling and Ariotti, 2009), and some claims in this area have aroused skepticism (Hajnal, Lajevardi and Nielson, 2017; Grimmer et al., 2018). The pervasive discrimination that blacks face in various arenas of American politics (Butler, 2014) suggests that this group could be at risk of bias in interacting with local election officials. While there is also ample evidence of discrimination toward Arab and Muslim Americans (Gaddis and Ghoshal, 2015), this group has received comparatively less attention from scholars (Jamal and Naber, 2007; Panagopoulos, 2006). In an era of political rhetoric increasingly characterized by appeals to group identity, it is particularly important to understand how racially-motivated bias impacts the day-to-day mechanics of elections for a range of racial/ethnic groups.

To seek evidence of bias, we focus on the thousands of local-level administrators charged with conducting elections in the United States. These bureaucrats are generally capable of exercising discretion in carrying out their job duties, which include responding to inquiries about the mechanics of voting and eligibility to participate in elections. Our core contention is that in exercising such discretion, street-level bureaucrats may be consciously or unconsciously influenced by the characteristics (e.g., race or partisanship) of individuals seeking public services (Lipsky, 1980; White, Nathan and Faller, 2015*a*).

## **5.1. Experiment Design**

To determine the extent to which previously documented bias is persistent and extends to other racial groups, we conduct an email audit study of local election officials (Pager,



2003).<sup>38</sup> Our intended sample comprises all such officials with publicly available email addresses and the analytic sample includes 6,439 local election officials from 44 states (Figure A.1).

The experimental stimulus consists of a single email sent to each local election official. All emails follow the same structure, greeting the official by name, referencing voter identification laws, and asking about the requirements to vote in the state corresponding to the official. Our design closely parallels White, Nathan and Faller (2015*a*), but differs in that we send only messages that mention voter ID laws. Additionally, to minimize possible spillover issues, we create 27 variants of this request (SI section A.4 and section A.6).

Our experimental treatment is the putative identity of the email sender. In line with convention we expose officials to four distinct group identities by manipulating senders' names (Bertrand and Mullainathan, 2004*a*; Bertrand and Duflo, 2017*a*; Butler and Homola, 2017*b*). Because the identities signaled in our treatments have elements which could be described as racial, ethnic, or religious, we refer to these generically as group identity treatments. To mitigate possible name effects, each group identity condition is signaled by 100 unique names. We check that the chosen names reliably prime ethnicity by conducting a manipulation check on Amazon's Mechanical Turk service in which workers read sets of names and ascribe probabilities that a name belongs to a particular racial or ethnic group.<sup>39</sup> In total, we send 4,900 unique experimental conditions which combine variants of the contact language with treatment identities.

---

<sup>38</sup>We received Human Subjects approval from the University of California, Berkeley and University Michigan Human Subjects Committees. Both committees waived the requirement of informed consent. Additional implementation details are made available in the Supplemental Information. The study design, and pre-analysis plan were registered at Evidence in Governance and Politics. Data, code, and computing environments are available at ALEX INSERT.

<sup>39</sup>SI section A.7 describes the procedure for choosing names, and section A.17 provides the complete list of names.

### 5.1.1. Treatment assignment and implementation

We block treatment assignment on logged population density, two-party vote share in the 2012 presidential election, percent African American, percent Latino, percent of households with incomes below 150 percent of the federal poverty level, and a dummy variable indicating whether a county was previously covered by Section 5 of the Voting Rights Act. Further details are provided in SI section A.8. Within each block we assign local election officials a racial condition and message version at random.

We sent 6,235 emails the morning of October 31, 2016, one email to each election official that was a part of the study.<sup>40</sup> Emails were sent from a purpose-built domain, `ez-webmail.com`. Sending addresses took the form of the senders' first initial, last name, and a two-digit string between twenty and forty. To mitigate the possibility that elections officials would be suspicious of our contact, we structured the email headers so that inboxes displayed the full name of the purported voter (see Figure A.2). The variety in our treatments was intended to reduce the likelihood that different offices would receive emails from identical senders. In twenty-nine of the forty-three states in our analytic sample every official received a contact from a distinct name.

One key innovation in this experiment permits the identification of whether emails were received and opened by election officials. We include a 1x1 pixel image with a unique link – commonly referred to as a tracking pixel – in the email body so that upon opening the email, most email clients loaded the image from our server and provided a positive record that the email had been opened by a particular official. This measurement permits inference about differential open-rates, a test of implicit bias we examine in subsection 5.2.1.

An open question in correspondence studies concerns whether observed effects are merely an artifact of differential treatment of stimulus by the internet and email infrastructure,

---

<sup>40</sup>We also sent two waves of pilot email, 54 on October 26, 2016; and, 146 on October 28, 2016. For details, see SI section A.12.

i.e., spam filters. Through pilot testing we are able to comment on this question. Before taking steps to develop positive server reputation, no messages reached any test inboxes. However, by carefully managing our digital authentication and consulting with individuals at a digital marketing company, in pilot testing we were able to place every message, from every attempted sender, into test inboxes (see SI section A.2).

The choice to contact election officials eight days before the election is designed to make our study reflective of the real constraints on individuals seeking and providing information about voting requirements. To minimize the impact of our intervention on election officials' time, the specific request contained in the email is one that would require little effort to fulfill. Using data gathered via our mailing system, we estimate that the median time to compose and send a response to our email is three minutes, six seconds. We contend that any costs borne by public officials as a result of our intervention are counterbalanced by the benefits of uncovering persistent bias in electronic communications between constituents and local election officials.

Our pre-registered analysis uses a single outcome measure, GOTRESPONSE, coded 1 if an election official replied to our email prior to election day, and 0 otherwise.<sup>41</sup> We do not count auto-replies, away messages, or bounces as valid replies. We further report an exploratory analyses of a novel outcome measure made possible through our engineering: whether a local election official opened the message.

## 5.2. Results

Overall, 57.8 percent of the emails we sent received at least one reply from local election officials. While lower than the 67.7 percent response rate previously obtained from a similar sample (White, Nathan and Faller, 2015a), this rate compares favorably with experiments

---

<sup>41</sup>Pre-analysis Plan Registered at EGAP (Hughes, Gell-Redman and Crabtree, 2016).

**Table 5.1: Response Rates by Experimental Condition**

<b>Ethnic Cue</b>	<b>White</b>	<b>Minority</b>	<b>Latino</b>	<b>Black</b>	<b>Arab</b>
Response Rate (%)	61.3	56.6	58.4	61.4	50.1
Standard Error	1.21	0.71	1.23	1.21	1.25
N	1,611	4,828	1,609	1,613	1,606

*Notes:* The *Minority* column includes all data from the *Latino*, *Black*, and *Arab* columns. Response rates and standard errors are reported in percentage terms.

on elected officials in the U.S., suggesting that our requests were taken at face value (Butler and Broockman, 2011*b*).

Election officials respond at considerably lower rates when queries come from minority as opposed to white senders (difference in mean,  $\Delta\mu = -4.70$  percentage points, *Wilcox Rank-Sum*  $P < 2 \times 10^{-16}$ ). However, as we report in Table A.5 responsiveness to minority senders is not uniformly lower. Nonparametric tests using white senders as the baseline find that a Latino name is sufficient to suppress the likelihood of a response by nearly 3 percentage points ( $\Delta\mu = -2.97$ ,  $P = 0.07$ ). Strikingly, an Arab/Muslim name lowers the likelihood of a response by greater than 11 percentage points ( $\Delta\mu = -11.3$ ,  $P < 1 \times 10^{-10}$ ). In contrast, black senders receive responses at a rate indistinguishable from white senders ( $\Delta\mu = 0.11$ ,  $P = 0.90$ ). Figure 5.1 (a) plots the Intent to Treat (ITT) causal effects of our treatments. Regression estimates with robust standard errors are reported in columns 1 and 2 of Table A.6, and produce similar results.

Figure 5.1 (b) plots a precision weighted meta-analysis estimate (Gerber and Green, 2012, p. 361) that combines the results of our intervention with those previously reported (White, Nathan and Faller, 2015*a*). These data, gathered in independent audits conducted over two election cycles, show that Latinos receive replies from local election officials at a rate 4.4 percentage points lower than whites ( $\Delta\mu = 4.4$ , precision weighted  $SE = 1.18$ ,  $P < 0.0001$ ).

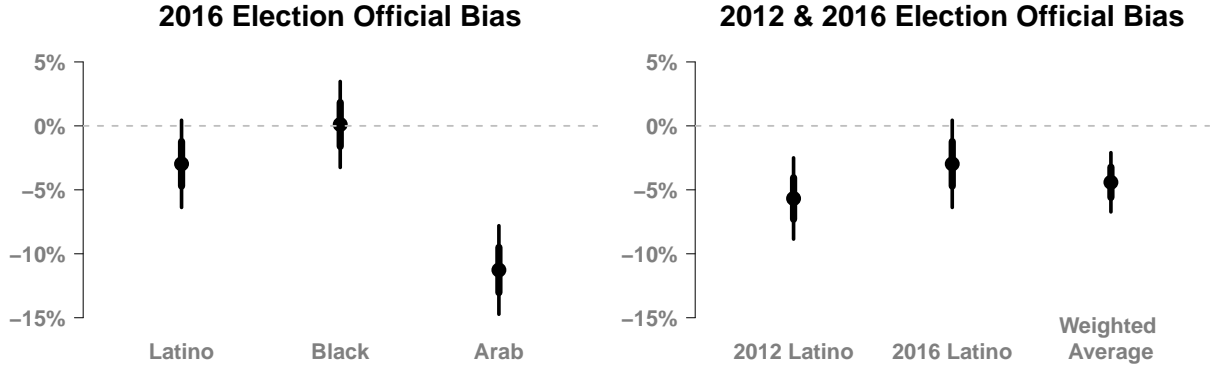
While the persistence of the treatment of Latino senders in the 2012 and 2016 elections is remarkable, perhaps more striking is the finding that Arab/Muslim names suffer a penalty more than two times greater than the one produced by a Latino stimulus. One potential concern is that the observed effect could be driven by the implausibility of the treatment, since many parts of the country do not have any appreciable population of Arab-Americans. To examine this possibility, we investigate whether treatment effects are smaller in the jurisdictions where Arab-Americans are more numerous. If treatment effects are driven by implausibility then they should be smaller in places where the presence of citizens with Arab names are more plausible. We do not find clear evidence that the proportion of Arab Americans moderates the treatment effect (Table A.13, Model 3; Table A.14; Table A.15). Our most credible estimates find a 10.6 percentage point bias against Arab senders in counties with no Arab population ( $\Delta\mu = -10.6$  percentage points,  $SE = 2.5$ ,  $P < 0.001$ ), but only a 2.6 percentage point improvement in the highest Arab population quartile of counties ( $\delta\Delta\mu = +2.6$  percentage points,  $SE = 4.4$ ,  $P = 0.55$ ), although the distribution of Arab American settlement limits the strength of this robustness check.<sup>42</sup>

### 5.2.1. Evidence of implicit discrimination

Local election officials who receive our intervention demonstrate bias insofar as they respond differentially based only on the signal of group identity delivered through our treatments. This observed response behavior is part of a chain of actions: the official must open, read, and then respond to the email. Standard analyses of audit experiments, which report an indicator of response or non-response as the dependent variable, focus only on the final result of this compound process. Innovations of our design allow us to consider the outcome at a prior step, the decision by the official to open the received email, conditional

---

<sup>42</sup>In the highest Arab quartile, the mean Arab population is 1%.



**Figure 5.1:** Points represent the ITT, the estimated difference in response rates to emails from the named identity, compared to the white response rate baseline. Thick bars report  $ITT \pm SE$ , thin bars report  $ITT \pm 1.96 \times SE$ . All estimates are difference in means, except the *Weighted Average* which estimates a precision weighted difference (Gerber and Green, 2012) utilizing 2012 (White, Nathan and Faller, 2015a) and 2016 Latino evidence.

on the treatment delivered.

To respond to our experimental stimulus, an election official must identify our request from among the large number of other requests, categorize it mentally, and then open it. We argue that opening an email is a high-volume, low-attention task of the type scholars have associated with implicit, rather than explicit bias (Devine, 1989; Bertrand, Chugh and Mullainathan, 2005, p.96). The pattern of email opens suggests that, indeed, elections officials may be unintentionally or automatically screening requests from Arab/Muslim senders. There is no difference in open rates between white and latino names ( $\Delta\mu = -0.74$  percentage points,  $SE = 1.7$ ,  $P = 0.68$ ) or white and black names ( $\Delta\mu = -0.24$ ,  $SE = 1.7$ ,  $P = 0.90$ ). However, there is a pronounced gap in open rates for emails sent by senders with Arab/Muslim names, who have their emails opened at a rate 6.8 percentage points lower than white senders ( $\Delta\mu = -6.8$ ,  $SE = 1.8$ ,  $P = 0.00013$ ).

### 5.2.2. Awareness of experiment

During the analysis phase of this project, it came to the researchers' attention another entity was pursuing a similar line of research using the same sending domain as White, Nathan and Faller (2015*a*). As a result, some public officials became concerned that an audit study might be underway. News reports claim that these concerns prompted the National Association of Secretaries of State (NASS) to alert its state branches, who in turn had the opportunity to alert individual officials. In sum, some of our experimental subjects may have become aware of the presence of interventions.

Subjects' awareness of the intervention poses a general threat to audit studies, either by compromising independence between units, or by violating the exclusion restriction if minority names are more likely to raise suspicion than white names. Because subjects' awareness might prevent identification of causal effects, researchers should mitigate this risk by using many identities and a well-tuned sending architecture whenever feasible. When there is any observable information about the possibility of discovery, researchers can use this information to evaluate whether apparent differences are likely the result of discovery.

Analysis of the timing of responses in this experiment does not suggest that discovery is leading to the observed results. First, as we present in Figure 5.2, the systematic pattern of unresponsiveness to minority names appears rapidly and well before the reported NASS broadcast. Second, as we report in Table A.11 and Table A.12, models that censor response data at the time of the NASS broadcast, and models that exclude states that witnessed interference between units both produce estimates very similar to our main results.

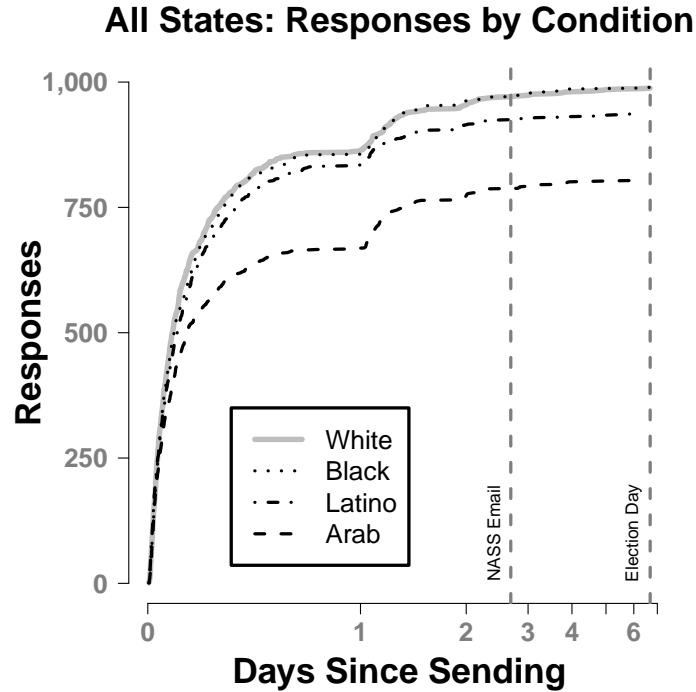


Figure 5.2: Rapidly slowing rates of response. The vertical axis plots the cumulative number of responses, split by group identity of sender; the horizontal axis plots time since sending. Election Day and NASS emails are noted with vertical dashed lines. Responses follow a clear diurnal rhythm, and patterns of bias appear rapidly.

### 5.3. Conclusion

Previous experimental evidence showed local election officials were less responsive to inquiries from Latinos, raising concerns about bias in the electoral process. Using a similar experimental design, we demonstrate the firm basis for these concerns by replicating the initial finding. We also extend the results by testing for bias against other groups.

Our intervention showed Arab/Muslim Americans to be markedly disadvantaged in their interactions with local election officials. This finding is particularly salient given that it is not simply an artifact of Arab/Muslims being a relatively less numerous part of the electorate. We encountered no evidence of bias from local election officials toward African



Americans, making ours at least the third recent study to produce a similarly unexpected null finding (Einstein and Glick, 2017*a*; Gell-Redman et al., 2018*a*). Rather than evidence of a lack of bias against African Americans, these null findings may be an artifact of the correspondence study method in which name alone, rather than other cues such as appearance, is used to signal identity.

Through this design, we also engage a challenge inherent to all audit studies, the risk that subjects become aware of the experiment. The relatively low technical sophistication required to conduct some forms of audit studies, mated with the potentially large sample size that is possible through email-based audits make these designs a potentially attractive way to identify discriminatory behavior. However, in an increasingly crowded field, researchers must face the possibility that experimental subjects become aware of the study, thereby damaging the inference. We determined that sending 4,900 distinct treatments on a custom-built server provided the best balance of a low possibility of discovery with the ability to identify a novel open rate outcome measure, and we would encourage future researchers to make a similar assessment.

# **Chapter 6. Does Religious Bias Shape Access to Public Services? A Large-Scale Audit Experiment Among Street-Level Bureaucrats**

In recent years, minority religious groups in the United States have faced heightened opposition and scrutiny. With the rise of Donald Trump, and the broader social forces that gave rise to his election, religious minorities like Muslims have faced harsh rhetoric, violence, and an unfavorable policy environment at the federal level.<sup>43</sup> There is qualitative/survey-based evidence that minority religious groups that depart from the religious mainstream as well as atheists face substantial hurdles.<sup>44</sup> While scholars have long documented the

---

This chapter is adapted from a working paper co-authored with Steve Pfaff, Holger L. Kern, and John L. Holbein.

<sup>43</sup>See “U.S. Muslims Concerned About Their Place in Society, but Continue to Believe in the American Dream,” *Pew Report*, July 26, 2017.

<sup>44</sup>See “Americans Express Increasingly Warm Feelings Toward Religious Groups,” *Pew Report*, February

importance of religion in the political realm (Inglehart and Norris, 2004; Putnam and Campbell, 2010), little work with a compelling identification strategy exists quantifying the extent, scope, and mechanisms behind potential discrimination of minority religious groups and atheists in the public domain. Although the Constitution of the United States prohibits the state from discriminating against individuals or groups based on their religious identification, it is by no means clear that equal treatment is afforded in practice as discrimination has been observed for other protected subgroups (Butler and Broockman, 2009).<sup>45</sup> The extent and nature of this bias are of vital import given the foundational principles of religious freedom and neutrality in the United States. To what extent do minority religious groups and atheists face discrimination in the public domain in the U.S.?<sup>46</sup> What forces drive any such biases?

In this paper, we begin to address these questions. To do so, we conducted a large-scale audit experiment of prospective school enrollment in which we emailed the principals of more than 45,000 public PK-12 schools in 33 U.S. states. We focus on local public schools as these core institutions, and the street-level bureaucrats they employ, shape the dynamics of local communities and serve as one of the most common touch points between citizens and their government (Holbein, 2016; McDonnell, 2013; Soss, 1999; Soss and Schram, 2007). Indeed, work in political science on the behavior and performance of street-level bureaucrats has often looked to public school officials as key actors (Lipsky, 1971; Prottas, 1979). Given that street-level bureaucrats have a great deal of discretion in

---

15, 2017.

<sup>45</sup>Beyond the Constitution, Title IV and Title IX of the Civil Rights Act of 1964 prohibit discrimination against students (our subgroup of interest) based on their faith. Case law clarifies that students can religiously identify at school and can take part in religious activities of their own devising. State constitutions also generally either make public education a fundamental right or contain protection clauses or their equivalents that prohibit religious discrimination (Alexander and Alexander, 2011, 46).

<sup>46</sup>By religious discrimination we refer to differential treatment based on religious affiliation or non-affiliation. In line with the literature (Butler, 2014), we use the terms “bias” and “discrimination” interchangeably.

how they enforce the rules, laws, and policies made by other government institutions, they represent a vital group to study as to whether they exhibit systematic bias. Moreover, as we describe below, public schools are at the center of fundamental debates about how the state and religion can, do, and should interact.

In our emails, we claimed to be a parent who was considering enrolling his or her child in that school and asked for a meeting with the principal. We randomly assigned the religious affiliation/non-affiliation of the family (no information given, Protestant, Catholic, Muslim, or atheist). To go one step further, we explored the potential mechanisms by which any discrimination occurs by also randomizing the intensity of the belief signal (low [identification], medium [compatibility inquiry], high [accommodation request]). This allows us to experimentally explore a key mechanism that might be driving any discriminatory effect: the perceived costs attached to the enrollment of religious adherents or atheists. We then observed whether principals replied to our email.

Compared to baseline emails, which provide no information about religious background, we found high levels of discrimination against Muslims and atheists. We found that Muslim and atheist parents are discriminated against for merely revealing their beliefs in the signature part of their emails. Signaling membership in these groups decreased the probability of a reply by 4.6 and 4.7 percentage points, respectively. This difference is statistically and substantively meaningful; it is only slightly smaller than (but not statistically distinct from) the discriminatory effects of race/ethnicity shown in previous audit studies (Butler and Broockman, 2011*a*). For these marginalized groups, discriminatory effects are present *regardless* of whether enrollment costs are explicitly signaled. Moreover, discrimination against Muslim and atheist parents increases *dramatically* if they inquire about the compatibility of the school with their beliefs or ask for religious accommodations, with such signals reducing response rates by 8.7 and 13.8 percentage points, respectively. These strik-

ingly lower response rates for (randomly assigned) higher levels of belief intensity suggest that an important mechanism behind the discriminatory effects we observe is the degree to which these individuals are seen to be imposing a cost on public officials. Response rates for Protestant and Catholic parents are indistinguishable from the no information baseline; discrimination only appears when parents inquire about the compatibility of the school with their beliefs or ask for accommodation of their beliefs. Finally, we show that discriminatory effects are systemic in the public education system. Given our large sample size, we can show with a great degree of precision that such discriminatory effects are remarkably consistent across the racial/ethnic composition of the school, the school type (primary, middle, or high), the median household income/poverty rates, the share of adults holding bachelor degrees, Republican vote shares in the 2012 presidential elections, and religious adherence rates of the surrounding community. Discrimination against citizens with non-mainstream beliefs about religion seems to be widespread in the American public school system — the venue in which many citizens most commonly interact with government.

Our work makes important contributions to the study of democratic representation, local politics, and experiments designed to detect bias among public officials. Our research suggests that not only do public officials discriminate on the basis of race/ethnicity (Adida, Laitin and Valfort, 2010; Butler and Broockman, 2011*a*; Einstein and Glick, 2017*b*) — the most commonly used treatment in audit studies of public officials (Costa, 2017) — they also do so against minority religious groups such as Muslims and atheists. *Even though* religion is a legally protected category, these groups still face substantial hurdles as they seek to gain access to basic public services. Our work is unique in its size and design. To our knowledge, ours is one of the (if not the) largest audit studies of public officials to date.<sup>47</sup> This feature

---

<sup>47</sup>For comparison, Butler and Broockman (2011*a*):  $n = 4,859$ ; Hemker and Rink (2017):  $n = 408$ ; Teele, Kalla and Rosenbluth (2017):  $n = 8,189$ ; and Carnes and Holbein (2015):  $n = 4,492$ .

allows us to make very precise inferences as well as use an experimental design that enables us to unpack the treatment effects that we observe. In broad strokes, whereas many audit studies focus on elected officials, ours focuses on street level bureaucrats, a group that has received much less attention in the literature (for notable exceptions, however, see Einstein and Glick (2017*b*) and Hemker and Rink (2017)) and that may behave differently than elected officials given the relative lack of electoral pressures (Dropp and Peskowitz, 2012). Our results suggest that Muslims and atheists face substantially higher hurdles in obtaining public services and that these effects are at least partially driven by a perception that intense beliefs about religion impose a cost or burden on public officials that they would rather avoid.

## **6.1. Background and conceptual framework**

The extent of religious discrimination in American public institutions is an important question not only because such discrimination is illegal but also because the role of religion in American society is changing. The United States stands out among advanced democracies not only for its relatively high level of religiosity but also for its religious diversity. In striking contrast to the citizens of many other wealthy democracies, an overwhelming majority of Americans continue to profess religious belief. At the same time, the religious landscape has been reshaped in recent decades as mainline groups have declined, religious diversity has increased, and a growing share of Americans identify themselves as non-believers (Baker, 2015; Putnam and Campbell, 2010; Sherkat, 2014). These changes raise questions about the ability of public officials — particularly in the education domain — to observe the civil rights of religiously diverse American families.

The most prominent and controversial markers of the changing American religious landscape are the growing shares of Americans identifying as Muslims or non-believers. The

Pew Research Center (Center, 2015) found that the share of the U.S. adult population identifying as Muslim doubled between 2007 and 2014 (albeit from only 0.4% to 0.9%), mostly as a result of immigration. Although the share of Americans that profess Islam as their religion is still small, this group has become a highly visible and controversial minority (Peek, 2011). As the median age of adult American Muslims is only 33 years (the U.S. median is 46 years) (Center, 2015), the enrollment of Muslim students in public schools can be expected to increase substantially in coming years.

In a reversal of the long-term trend which had previously “churched” American society (Finke and Stark, 1992), the share of American “nones” has increased steadily over the last two decades (Sherkat, 2014). In 2014, Center (2015) found that the unaffiliated, atheists, and agnostics comprise about 23% of the adult U.S. population, a share that has increased from 16% since 2007. The increase in “nones” is the result not only of disaffiliation from many Christian denominations but also the avoidance of religion by many young people — fully 35% of Millennials report no religious preference. Of course, not all “nones” are non-believers (Hout and Fischer, 2002; Marler and Hadaway, 2002), but the share of American adults that identify as atheist (the corresponding share of agnostics is in parentheses) has likewise nearly doubled from 1.6% (2.4%) to 3.1% (4.0%) during the 2007 to 2014 period. Because the median age of adult atheists is just 34 years (Center, 2015), the public school enrollment of students from households of committed non-believers can also be expected to increase in the future.

Citizens of all religious creeds as well as committed non-believers enjoy formal protection from harassment and equal rights in the public domain. In practice, however, the religious liberties of citizens appear to be frequently violated (Lippy, 2006; Peek, 2011). There are many qualitative examples of religious discrimination. Many of these have occurred in public schools — the setting of our audit study. In 2004, for example, the Department of Justice

sued a school for prohibiting a Muslim girl from wearing a headscarf (Hearn and *United States v. Muskogee Public School District*). In 2007, a high school student was kicked off the women’s basketball team for refusing to take part in the Lord’s Prayer (*Smalkowski v. Hardesty Public School District*). In 2012, a high school student was subject to harassment after asking that a prayer banner be removed from a place of prominent display within a public school (*Ahlquist v. Cranston*).

Despite these qualitative examples, little systematic or experimental research exists detailing whether, how often, and why religious discrimination occurs. Most of the existing research on diversity and discrimination in American public education focuses on ethnic and racial disparities in performance and enrollment (Johnson, Crosnoe and Elder Jr, 2001; Kao and Thompson, 2003; Roscigno, 1998). This is also true for research on discrimination by public officials more generally, which focuses predominantly on issues of race and ethnicity. While this social dimension is vitally important, what is missing from the literature is research on religious discrimination in public schooling. We simply do not know whether reported cases of discrimination represent exceptional incidents or merely the tip of the iceberg. Given the lack of systematic, rigorous evidence it is impossible to gauge the extent of religious discrimination in American public institutions in general and public schools in particular.

## **6.2. Previous experimental research on discrimination**

Over the last decade, social scientists have advanced the study of discrimination by using field experiments to address the well-known limitations of surveys and observational studies in demonstrating bias (Bertrand and Duflo, 2017*b*; Costa, 2017). Audit experiments have examined if bias occurs in response to group-based identification on the basis of race, ethnicity, gender, or sexual orientation (Blommaert, Coenders and van Tubergen, 2014;



Butler, 2014; Butler and Broockman, 2011*a*; Gaddis, 2014; Teele, Kalla and Rosenbluth, 2017; Neumark, 2012; Pager and Shepherd, 2008; Pedulla, 2016). A few experimental studies have also examined the potential for religious discrimination in the workplace. In an influential set of papers, Wright et al. (2013) and Wallace, Wright and Hyde (2014) found that U.S. job applicants expressing a religious identity were less likely than those who did not to receive a response from a potential employer, with minorities such as Muslims and atheists suffering the greatest bias and evangelical Christians and Jews suffering little or no discernible bias. The U.S. does not appear to be unique in this regard. In France, not only do Muslims have lower incomes than matched Christian households but a Muslim job candidate is about 2.5 times less likely to receive a job interview callback than a racially similar Christian counterpart (Adida, Laitin and Valfort, 2010).

A small, but rapidly growing, literature examines various biases among elected public officials (Butler and Broockman, 2009, 2011*a*; Butler, 2014; Costa, 2017).<sup>48</sup> This strand of research, however, has paid less attention to the question of whether appointed public officials exhibit bias. Moreover, studies involving both elected and appointed public officials primarily focus on partisan and racial discrimination (Einstein and Glick, 2017*b*; White, Nathan and Faller, 2015*b*). To our knowledge, no experimental research exists on religious biases in either elected or appointed public officials such as American public school principals (our population of interest). Moreover, because of the challenges involved in conducting audit studies on political elites, such studies rarely have the statistical leverage to (experimentally) explore a key, theoretically-driven potential mechanism behind discrimination.<sup>49</sup>

---

<sup>48</sup>See Broockman and Soltas (2017) for an innovative example of research on racial discrimination against elected officials (i.e., delegates).

<sup>49</sup>There are, of course, exceptions. For example, Butler and Broockman (2011*a*) explore whether party membership is a mechanism behind racial discrimination by state legislators.

## **6.3. Potential sources of religious discrimination in American public education**

Broadly speaking, over the past sixty years there have been two opposing trends affecting the relationship between religion and public education. Secularizing activists and federal courts have drawn a sharper line between religion and public education. At the same time, such efforts have been met by push-back, especially from conservative religious groups and politicians. As local public officials entrusted with educational management, PK-12 principals are pulled into different directions by these opposing trends (Justice and Macleod, 2016; Reese, 2011). We propose three theoretical perspectives on the sources of religious discrimination by principals, focusing on the influence of secularism, Judeo-Christian nationalism, and civil religion. These perspectives draw heavily from the field of sociology, which traditionally has devoted much attention to the effects of religion on social interactions (Weber, 2013). They allow us to make explicit predictions about the effects of signaling religious/non-religious views, relative to the no information baseline condition, and the role that a key mechanism — the perceived costs attached to the enrollment of students from households with intense beliefs — plays in generating discrimination. Note that we rely on these literatures to set expectations about the possible extent and shape of religious discrimination in American public schools. Our goal is not to conduct a “horse race” between these explanations. They are not mutually exclusive.

### **6.3.1. Secularism as a basis for anti-religious bias**

In the study of religion, secularization theory proposes that modernization propels the decline of religion at the level of institutions, attitudes, and beliefs (Swatos and Christiano, 1999; Wald and Wilcox, 2006). Modernization provides individuals with moral autonomy,

opportunity, and personal security, initiating a “culture shift” away from religious traditions to post-traditional values and lifestyles (Inglehart and Norris, 2004). At the same time, education and social diversity erode religion’s plausibility, intensity, and authority, leading to a widely shared preference that religion be left to the private realm (Berger, 1967; Bruce, 2002).

Although the U.S. was long considered an exception to the secularization thesis, recent developments suggest that secularization processes are unfolding and may be accelerating (Putnam and Campbell, 2010; Sherkat, 2014; Voas and Chaves, 2016). An important factor driving contemporary secularization is political conflict over values and religious issues, which is leading increasing shares of moderates and liberals to eschew organized religion altogether (Baker, 2015; Hout and Fischer, 2002; Manning, 2015).

Clashes over the secular nature of the public school system, the limits of religious accommodation, and state support for religious activities are nothing new (Alexander and Alexander, 2011; Matzke, 2016; Justice and Macleod, 2016). Historically, Protestant churches exerted substantial influence on public schooling. Religious discrimination toward minority faith communities and newcomer religions was commonplace (Reese, 2011). Even though campaigns to extend church-state separation faced substantial resistance from religious conservatives, they were remarkably successful in the domain of public schooling. Despite popular religiosity, American public education has secularized more rapidly and thoroughly than in many other advanced democracies. Secular values deeply influence the institutional culture of American public schooling (Mayrl, 2016).

The result is a tension between a procedural secularism, which guarantees that public institutions hold no religious preference, and a programmatic secularism, which insists that the public sphere admit no religion (Williams, 2012). This tension generates unease about addressing religion in school (Justice and Macleod, 2016). School officials are trained to

be zealous guardians of church-state boundaries and to embrace secular norms in public education. Essex's widely-used *School law and the public schools: A practical guide for educational leaders* is very clear in this regard. It instructs principals and administrators that the law "compels public schools as state agencies to maintain a neutral position in their daily operations regarding religious matters" (Essex, 2002, 17) and insists that they are legally obligated to refrain from endorsing religious symbols, devotions, and expressions at school and to avoid supporting students' religious activities (Essex, 2002, 16–47).

Discrimination could arise from a desire to avoid controversies and the anticipated costs of religious accommodation. In practice, school administrators wish to avoid issues which create friction among students, between students and staff, with parents, and with the broader public (Bess and Goldman, 2001). We posit that a theoretical reason for bias could involve a moral judgment that religious families are prone to make illegitimate, disruptive, and/or costly requests for special treatment. The result would be discrimination against parents who are affiliated with any religious group, especially if the parents explicitly inquire about the compatibility of the school with their beliefs or request religious accommodation. These predictions are captured in **H1** below.

**[H1] Bias against the outwardly religious hypothesis:** Parents who reveal a religious affiliation when requesting a meeting with public school principals will experience discrimination from principals relative to parents who are silent about their beliefs. Moreover, the strength of the discrimination against parents revealing a religious affiliation will increase if they inquire about the compatibility of the school with their beliefs or request religious accommodation.

### 6.3.2. Judeo-Christian nationalism and religious bias

Religion continues to play a central role in establishing the boundaries of American national identity (Bonikowski and DiMaggio, 2016; Hartmann et al., 2011). Christianity is especially important in subjectively defining “legitimate” membership in the American nation (Gerteis, 2011). In a representative sample of Americans, 65% of respondents reported that Christianity was a “fairly” or “very important” criterion for being considered “truly American” and nearly half (48%) said it was “very important” (Bonikowski and DiMaggio, 2016, 955).

Christian identity has expanded since the mid-twentieth century, with intellectuals and politicians drawing heavily on explicitly “Judeo-Christian” religious discourses to construct the moral boundaries of America (Neuhaus, 1984). Historians have shown how denominationalism and immigrant-driven diversity suggested to many public elites a new conception of the political role of religion. The struggle against Fascism gave this project a special urgency, leading to a vision of national identity which included Protestants, Catholics, and Jews — three groups purportedly united by Judeo-Christian values into a new kind of multicultural nation (Carenen, 2012; Sarna, 2004). Over time, national political discourse reflected this broadened notion of religious values just as denominational groupings replaced sectarianism and ethnicity in discussions of collective identity. Explicit anti-semitism went from mainstream politics to the margins of American life (Hartmann, Zhang and Wischstadt, 2005; Herberg, 1983).

Judeo-Christian political ideology was particularly useful in helping to integrate generations of white immigrants from Southern and Eastern Europe. While compared with the sectarianism of the past, the Judeo-Christian formulation was inclusive, it was inclusive only up to a point (Douthat, 2013). With the acceleration of immigration from non-European countries that began in the late 1960s, the American religious landscape be-

came far more diverse. In practice, newcomer religions have quickly adapted to American denominationalism (Berger, 2017; Hirschman, 2004). Nevertheless, the question for some Americans has become *which* religious groups belong and *which* religious voices should be heard in public life. From the 1990s onward, conservatives prominently reasserted claims about America as a Judeo-Christian nation (Wilcox, 2018). Conservative Protestantism thrived, at least in part, because its leaders portrayed it as the embattled defender of “true” American values (Lindsay, 2007; Sutton, 2014).

The limits of inclusion are apparent in the unease of many Americans toward members of unusual religious groups and religious newcomers (Bonikowski and DiMaggio, 2016; Edgell, Gerteis and Hartmann, 2006). Anti-Muslim discourse resonates with many Americans, particularly the more than 30% who identify with conservative Protestantism (Pew 2015) and the quarter of Americans who can be classified as ardent nationalists (Bonikowski and DiMaggio, 2016). In the context of a Judeo-Christian understanding of national identity and moral belonging, Muslim Americans pose a special problem, particularly in the wake of the 9/11 attacks and the War on Terrorism, the ongoing Arab-Israeli conflict, and the rise of the Islamic State.

In their analysis of American nationalism, Bonikowski and DiMaggio (2016) found that those Americans they classify as “ardent” nationalists are overwhelmingly white conservative Protestants who are prone to exclude religious minorities from those they consider truly American. In 2016, Donald Trump, who received about 80% of the white conservative Protestant vote, made the depiction of Muslims as outsiders a prominent theme in his campaign (Braunstein, 2019). Hence, a theoretical reason for bias against Muslims may be moral judgment. In the context of the politicization of Islam as a supposed threat to American society and values, principals might be reluctant to assist Muslim families in enrolling in their schools either because of their own moral bias or because of what

they consider to be prevailing community standards. These predictions are outlined in **H2** below.

**[H2] Bias against Muslims hypothesis:** Parents who reveal an affiliation with Islam when requesting a meeting with public school principals will experience more discrimination from principals than parents who reveal an affiliation with Protestantism or Catholicism, regardless of the intensity with which the parents' religious beliefs are communicated. Moreover, the strength of the discrimination against parents revealing an affiliation with Islam will increase if they inquire about the compatibility of the school with their beliefs or request religious accommodation.

### **6.3.3. Civil religion and bias against non-believers**

In American society, religion has long been an important source of conceptions about political community and social belonging. Opinion research has shown that, even as they have become more accepting of ethnic, racial, gender, and cultural diversity generally, Americans across the political and racial spectrum remain notably hostile in their attitudes toward non-believers in roles of political leadership or as appropriate marriage partners for their children (Cragun et al., 2012; Edgell, Gerteis and Hartmann, 2006; Edgell et al., 2016).

Popular suspicion and hostility toward atheists is rooted in the historical evolution of American political culture. Religion has been important for a restless American civil society in which faith-based groups provide social attachments and serve secular needs such as charitable assistance, opportunities for dating and marriage, daycare, connections to potential business partners and employers, and the integration of immigrants into community life (Sherkat and Ellison, 1999; Hirschman, 2004). In spite of the growing disaffiliation of Americans, congregations of all kinds remain the most common form of civic membership in America (Putnam and Campbell, 2010). The view that religion is important for society

has persisted even as the salience of denominational boundaries has declined (Hout and Fischer, 2002; Sherkat, 2014).

In part because of religious organizations' practical social importance and role in fostering American democracy, many Americans have come to understand the U.S. as a religious country (Noll, 2002). Scholars of civil religion argue that this is envisioned not in terms of an established church or favoritism toward a particular denomination, but rather as a consensus about the importance of religion for society. American civil religion endorses religious pluralism and a broadly spiritual notion of national community, resulting in "a public religious dimension [that] is expressed in a set of beliefs, symbols, and rituals" (Bellah, 1967, 4). The legacy of American civil religion is evident in comparative perspective. The U.S. is noteworthy among advanced democracies for the effectiveness with which religious interest groups influence public policy and the high share of the population (about half) which believes that religious leaders should influence public policy (Grzymała-Busse, 2015; Pfaff, 2008).

Historically, civil religion has bolstered a bi-partisan strategy to seek political consensus around a vision of a diverse society united against "godless" Communism (Sutton, 2014). Many influential religious leaders endorsed this vision of civil religion, ranging from the liberal theologian Reinhold Niebuhr to the evangelical preacher Reverend Billy Graham. America became for many the "new Israel," with Americans being the "chosen people" elected by Providence to safeguard freedom. Although Ronald Reagan was one of the most persuasive prophets of this vision, repeatedly evoking images of America as a God-given "city on a hill," he was hardly alone in his willingness to draw upon religious symbols and discourse to foster national unity. Religion has been a remarkably persistent feature of American political speech across both major political parties (Coe and Domke, 2006).

While they tolerate religious diversity, the explicit rejection of religion is intolerable to a



majority of Americans (Edgell, Gerteis and Hartmann, 2006; Edgell et al., 2016). Edgell, Gerteis and Hartmann (2006) report that “[a]theists are at the top of the list of groups that Americans find problematic in both public and private life” (230) (see also Cragun et al. (2012)). Survey research finds that anti-atheist bias in the United States is “persistent, durable, and anchored in moral concern” (Edgell et al., 2016, 629). Recent experimental research reveals that people intuitively judge atheists as immoral (Gervais, 2014) and regard them as lacking pro-social values (Simpson and Rios, 2017). Consequently, atheists are strongly associated with immorality and contempt for common values and seen as “moral outsiders” in American society (Edgell, Gerteis and Hartmann, 2006, 227).

Distrust toward non-believers extends to attitudes about schools. More than one third of Americans in recent General Social Surveys say that atheist teachers should be fired (Sherkat, 2014, 159). Many Americans appear to believe that, by openly rejecting religion, atheists are rejecting the normative foundations of community and the broader civic good. We posit that the mechanism producing bias against atheists is moral judgment. If parents identify themselves as atheists they may invite suspicion from school principals who fear that atheists and their children would be immoral, ideologically strident, and likely to opt out of civil rituals (such as the Pledge of Allegiance). Principals might be reluctant to assist atheist families in enrolling in their schools either because of their own moral bias or because of what they consider to be prevailing community standards. These predictions are outlined in **H3** below.

**[H3] Bias against atheists hypothesis:** Parents who reveal their atheist beliefs when requesting a meeting with public school principals will experience more discrimination from principals than parents who reveal an affiliation with any religion, regardless of the intensity with which the parents’ beliefs are communicated. Moreover, the strength of the discrimination against parents revealing atheist beliefs will increase if they inquire about the compatibility of the school with their atheist beliefs or request the accommodation of their atheist beliefs.

## 6.4. Research design and data

We use a large-scale audit experiment to investigate religious discrimination by PK-12 principals.<sup>50</sup> Our experimental sample consists of regular, operational, non-charter public PK-12 schools in 33 U.S. states. We included all states for which we were able to acquire principals' email addresses either by contacting state Departments of Education or by downloading contact information from the websites of those institutions. Within these 33 states, we dropped all schools with missing principal contact information. We also excluded schools that could not be uniquely matched to NCES (National Center for Education Statistics) data and schools with missing covariate data in the NCES or American Community Survey (ACS). Based on state and NCES data, we dropped inactive, private, charter, non-traditional, adult, and virtual schools as well as schools serving restricted populations such as schools for the blind and deaf and schools located on military bases. We also excluded schools with less than 100 students, schools that are majority American-Indian, and schools that offer pre-Kindergarten or Kindergarten as the highest grade. If several schools shared a principal we only kept one of the schools, chosen randomly. If several schools were located at the same physical address we only kept one of the schools, also chosen randomly.

Based on these selection criteria our sample size equaled 47,550 schools. When we conducted our audit experiment some of our emails could not be delivered due to misspelled or outdated principal email addresses.<sup>51</sup> Hence, our final sample size equals 45,710 schools.<sup>52</sup>

---

<sup>50</sup>Our experiment has been approved by our Institutional Review Board. We discuss the ethics of our experiment in the SI (supplementary information).

<sup>51</sup>We dropped all schools with bounced emails from the experimental sample. This is unproblematic because invalid or outdated email addresses are orthogonal to treatment assignment by virtue of randomization. Our results are unchanged if we treat principals with bounced emails as non-replies (results available upon request).

<sup>52</sup>The following 33 states make up the experimental sample (with number of schools in parentheses): AL (851), AR (761), CA (5892), CO (832), DE (135), FL (1902), GA (1636), IA (864), ID (389), IL (2519), IN (1486), LA (821), MA (1314), MI (1979), MN (813), MO (1412), MS (702), NC

The reason for our large sample size relative to previous audit studies is that we desire to (1) precisely estimate the effect of multiple main treatment conditions (i.e., religious affiliation/non-affiliation), (2) precisely estimate experimentally assigned second-order conditions (i.e., intensity of beliefs) to evaluate a theoretically driven potential mechanism, and (3) precisely estimate heterogeneous treatment effects across a host of contextual variables. Each of these — especially (2) and (3) — require higher statistical power than previous audit studies. Indeed, recent research shows that tests for heterogeneous treatment effects are often woefully underpowered (Blair et al., 2016). As can be seen in our results, our estimates are precise, but not so precise as to suggest that our study’s sample size is extravagant.

As in all audit studies, our outcome of interest is whether or not an individual responds to our inquiry. While this measure is not perfect, the justification for using it is that it gives us a glimpse into a real-world behavior, thus offering an improvement over other measures of bias such as implicit association tests or list experiments. In seeking to get more information out of an audit study, some may feel compelled to argue that in addition to exploring whether principals responded, we should measure characteristics of their responses such as how “friendly” they were. The problem with doing so is that such an approach implicitly conditions on a post-treatment variable (getting a response in the first place), thus risking post-treatment bias (Montgomery, Nyhan and Torres, 2018). Using measures that are only defined for the subset of subjects who responded “de-randomizes” an experiment in the sense that the resulting treatment and control groups no longer have potential outcomes

---

(2027), ND (161), NE (571), NH (326), NJ (1759), NM (412), NY (2904), OH (1974), RI (189), SC (930), TN (1067), TX (4723), VA (1427), VT (179), WA (1336), WI (1417). One complication arose during our experiment. In Massachusetts, our emails coincided with a malware attack targeting public schools. At least one principal thus forwarded our email to Massachusetts state police, which contacted all Massachusetts public school principals warning them that our emails were probably spam. We chose to keep Massachusetts in our sample since this warning only occurred one week after we had emailed principals; many Massachusetts principals had already replied by this point. Our results are entirely unchanged if we drop all Massachusetts schools. We have included this set of results in the SI.

that are in expectation equivalent” (Coppock, 2019), turning an experiment into a poorly designed observational study.

We observe a number of covariates drawn from the NCES (2013), ACS (2012), and the Religious Congregations & Membership Study (RCMS) (2010). From the NCES, we observe the share of Asian, Hispanic, Black, and White students at the school level. We also observe the share of students eligible for free or reduced price lunches, the share of male students, the school size, and the pupil/teacher ratio. From the ACS, we observe the median household income, the share of adults holding bachelor degrees, and the share of residents with income below the poverty line at the county level. We also observe county-level Republican vote shares in the 2012 presidential elections. From the RCMS, also at the county level, we have the rates of Black Protestant, evangelical Protestant, mainline Protestant, Catholic, Muslim, and total adherents per 1,000 capita.<sup>53</sup>

The plots in Figures B.2–B.4 in the SI (supplementary information) compare our sample to the NCES population of 78,348 regular, non-charter public schools without missing NCES data in the 48 contiguous U.S. states. While our experimental sample is not truly a random sample from this NCES universe, it tracks the NCES population rather well in terms of observed covariates.

Our experiment contains four different treatments: (i) *Parent’s gender* (male/female), (ii) *child’s gender* (male/female), and (iii) *religious affiliation/non-affiliation* (no information given, Protestant, Catholic, Muslim, atheist). To unpack treatment effects and explore the mechanisms behind religious biases, we (iv) also randomize the *intensity* with which religious/atheist beliefs are communicated (low [identification], medium [compatibility inquiry], high [accommodation request]). This allows us to explore whether discrimination is

---

<sup>53</sup>Using Amelia II (Honaker, King and Blackwell, 2012), missing data in RCMS variables have been multiply imputed using the NCES and ACS variables listed above, the outcome variable, and an additional set of 18 ACS variables plausibly prognostic of religious adherence or missingness. All standard errors and statistical tests have been adjusted to account for multiple imputation.

indeed driven by higher perceived costs associated with families' greater intensity of belief, as we have theorized. As a secondary part of our design, we also randomized the parent's and child's gender to rule out the possibility that our causal inferences about religious affiliation/non-affiliation or intensity of belief are driven by a particular gender or gender combination. We signal the parent's and child's gender by using different names: Isaac and Rebekah Adam for the parents and Jonah and Sarah for the children. We chose these first and last names because they frequently appear in the Old Testament, an important religious text for both Christians and Muslims.<sup>54</sup> As each of these names is relatively common in the United States, atheists with these names are also not unusual. Given the large number of emails we had to send out we used eight different email accounts to contact principals; email account names also signal parents' gender by including either Isaac or Rebekah. We sent emails throughout a one-week period in April 2016; the order in which principals were contacted was randomized. The text of our emails is shown in Figure 6.1.

Note: Emails revealing no information about the parent's religious affiliation/non-affiliation exclude text blocks A, B and C. Among emails that do reveal religious affiliation/non-affiliation, low intensity requests include C (but not A or B), medium intensity requests include A and C (but not B), and high intensity requests include B and C (but not A).

We include a Catholic treatment in our experiment to ensure that discrimination against Christians is not being driven by political hostility toward conservative Protestants. Catholicism is liturgically and theologically distinct from Protestantism and readily culturally identifiable. As a religious group, contemporary Catholics are a good benchmark because they are ethnically diverse and close to the U.S. mean on many demographic characteristics including education, income, and political preferences. In terms of the American religious

---

<sup>54</sup> Of course, Muslims might be more likely to have Arabic versions of these names; "Jonah", for example, might be rendered as "Yunus." However, if we had used different names for different religious affiliations we would have conflated signals of religious affiliation with signals of race/ethnicity.

spectrum, Catholics, on average, identify as religious moderates (Sherkat, 2014). They furnish a better reference category than “mainline” Protestants because liberal Protestants are, on average, similar to the unchurched in attitudes and values. Accordingly, a mainline Protestant identification as a religious signal would not be as resonant as Catholicism.

The literature on religious discrimination has often focused on the Jewish experience (Davidson and Pyle, 2011). When designing our study we decided to focus on Islam rather than Judaism as our non-Christian minority religion of interest. Obviously, we are not claiming that Jews do not experience religious discrimination. Rather, our reasoning was that (at the time of our study) survey evidence consistently showed that Americans were more favorable toward Jews than toward any other religious group, that discrimination was ebbing, anti-semitism was declining, and that stereotypes were fading (Rebhun, 2016).<sup>55</sup>

As an improvement over (most) audit studies, we also seek to experimentally test for mechanisms driving the effects we observe. To be abundantly clear, in doing so we are not claiming that we are trying to identify the *only* mechanism driving our effects. Doing so is inherently difficult — if not impossible — given the presence of unobservable mediators (Green, Ha and Bullock, 2010). However, given our large sample size we are able to build into our design experimentally assigned conditions that test a primary theoretical mechanism: that principals expect families with more intense beliefs to be more costly. To do so, we experimentally signal religious identity and the intensity with which beliefs are held in the following way. The low intensity condition signals religious affiliation/non-affiliation only through an email signature at the bottom of the email, in purple color. The email signature contains a modified version of a Richard Dawkins quote (“[...] teaches that

---

<sup>55</sup>We designed and conducted our study well in advance of the 2016 presidential election and failed to foresee the reappearance of anti-Semitic tropes during the election campaign and, after the election, among the so-called alt-right. Our experimental design was informed by Wright et al. (2013) and Wallace, Wright and Hyde (2014), who find minimal discrimination against Jews in the labor market. In hindsight, the inclusion of Jewish families would have been valuable.

life is precious and beautiful. We should live our lives to the fullest, to the end of our days.”). This quote is sufficiently bland (and obscure) that it could be reasonably attributed to virtually any source. We substitute “Christianity,” “Catholicism,” “Islam,” or “Atheism” into the quote, depending on the religious affiliation/non-affiliation treatment. We also change the purported author of the quote to Rev. Billy Graham, Pope Benedict, The Prophet Muhammad, or Richard Dawkins, again depending on the religious affiliation/non-affiliation treatment.

The medium intensity condition keeps the signature but adds the following sentence, which is designed to signal the desire for compatibility between the school and the beliefs of the family: “One of the reasons we would like to meet with you is that we are raising [Jonah/Sarah] to be a good [Christian/Catholic/Muslim/Atheist Humanist] and want to make sure that this would be possible at your school.”<sup>56</sup> The high intensity condition likewise keeps the signature but adds the following sentence, which is designed to signal a request for the accommodation of the family’s religious beliefs: “One of the reasons we would like to meet with you is that we are raising [Jonah/Sarah] to be a good [Christian/Catholic/Muslim/Atheist Humanist] and want to protect [him/her] from anything that runs counter to our beliefs. We want to make sure that this would be possible at your school.” The no information given condition only contains the gender treatments.

Treatments were randomly assigned within blocks defined by state, shares of Asian, Hispanic, Black, and White students, the percentage of students eligible for free or reduced price lunches, median household income, the share of adults holding bachelor degrees, the share of residents with income below the poverty line, and Republican vote share in the

---

<sup>56</sup>We used “Atheist Humanist” as opposed to merely “Atheist” in our emails since atheism as such does not have any ethical content. It would have sounded odd if parents had announced their intention to raise children to be “good Atheists.” Moreover, we used “Christian” as opposed to “Protestant” in our emails since American Protestants typically refer to themselves as “Christian” and not “Protestant.” In order to ensure that respondents would recognize the Protestant treatment in the signature line we attributed it to Billy Graham, one of the most famous Protestant clergymen of the late 20th Century.

2012 presidential elections. These design elements signal our *ex ante* interest in exploring treatment heterogeneities along these dimensions. Table B.1 in the SI shows that our sample is well-balanced.<sup>57</sup> We also compute an omnibus randomization inference *p*-value that tests for joint balance across all 18 covariates. This *p*-value equals 0.90, confirming that the blocked randomization procedure was successful in balancing observables.

We sent a single email to each principal, with no follow-up in case of non-response. We then observed whether principals replied to our email within a 14 day window from the time the email was sent.<sup>58</sup> Automatic replies such as out-of-office replies were discarded. We use receipt of a non-automated reply email as our binary outcome variable.<sup>59</sup> We recognize of course that non-response results from many sources besides bias. For example, each principal’s responsiveness is undoubtedly affected by factors such as his or her work load. That being said, we are interested in systematic differences across randomly assigned groups of principals exposed to different emails. While we cannot interpret non-reply by any individual principal as a sign of discrimination, the presence of systematic differences in responsiveness between randomly assigned treatment groups *is* evidence of discrimination

---

<sup>57</sup>We approximated exact randomization-based *p*-values using 10,000 randomly chosen blocked treatment assignments. The test statistic is the maximum Kolmogorov–Smirnov statistic across all two-way comparisons of treatment groups. The *p*-value is the fraction of test statistics at least as large as the test statistic in our sample.

<sup>58</sup>Most principals responded within three business days, so a 14 day window is conservative. Lengthening the window to four weeks does not change our results at all.

<sup>59</sup>We randomly sampled and read 500 reply emails. In almost all of them, principals either asked for times that we would be able to meet or proposed times for a meeting. In a few emails, principals asked us to provide additional information such as our moving date or our child’s grade level. In 8 emails, principals informed us that their schools did not offer school tours at the moment, typically because state testing was currently taking place. 17 reply emails were from former principals who suggested contacting the current principal, almost always either providing contact information for or *cc*’ing the current principal. After discarding automatic replies, we thus feel confident in treating the receipt of a reply email as indication of a principal’s willingness to meet with us. A typical reply was something like “Sure! When can you come in?” We should also point out that none of the replies suggested that principals found our emails suspicious in any way. We originally had planned to use text analysis tools to analyze and code the content and tone of reply emails, but the reply emails proved to be too uniform in content and too terse to make such an endeavor worthwhile. The major variation in responses is thus between getting a reply email and not getting a reply email.



(Bertrand and Duflo, 2017*b*; Butler and Broockman, 2011*a*). Of course, teasing apart the exact roots of discrimination, be they statistical or taste-based, is inherently difficult (Guryan and Charles, 2013) and in all likelihood requires a series of experiments and/or observational studies.

## 6.5. Empirical results

Among our 45,710 subjects, 19,696 sent at least one non-automated reply email within 14 days, for a response rate of 43.1%. Most replied within three business days; extending the reply window to four weeks does not change our results in any way. This response rate is in line with response rates from other internet audit experiments with elected and appointed public officials (Costa, 2017).<sup>60</sup>

Table B.2 in the SI shows results from a probit model. Because we are interested in the interaction between the religious affiliation/non-affiliation treatment and the intensity treatment, we include dummy variables representing all combinations of the religious affiliation/non-affiliation and intensity treatment levels in the model. The model also includes dummy variables for parent’s and child’s gender as well as fixed effects for the eight email accounts we used to send emails (coefficient estimates not shown). Robust standard errors are clustered at the school district level.

We use a plot to visualize the main empirical results of our experiment. Based on the probit estimates in Table B.2, Figure 6.2 plots treatment effects (i.e., differences in probabilities). Treatment effects of male names are in comparison to female names. Treatment effects for the twelve religious affiliation/non-affiliation and intensity combinations are in comparison to the baseline condition in which we do not provide any information about

---

<sup>60</sup>A small number of principals sent several emails, typically to update times they had mentioned in a previous email during which they would be available for a meeting.

the religious affiliation/non-affiliation of the family.<sup>61</sup>

While not of primary interest here, we find that compared to female names, male names for both the parent and the child reduce the probability of receiving a reply email. For parent’s gender, the estimated effect size is 1.6 percentage points; for child’s gender, it equals 0.8 percentage points. These effect estimates are statistically significant at the 0.05 and 0.10 levels, respectively.<sup>62</sup>

Turning to the religious affiliation/non-affiliation and intensity treatments and comparing to the baseline (no information) condition (while averaging over the gender factors), we find the following patterns. For the religious affiliation/non-affiliation treatment paired with the low intensity condition, the effects of Protestant and Catholic affiliation are slightly positive but not statistically significant at the 0.10 level. This suggests that simply signaling membership in a mainstream religious group has no effect on the probability of getting a reply — principals do not discriminate against families belonging to these mainstream religious groups when costs of enrollment are not signaled.

An affiliation with Islam, on the other hand, even if signaled solely through the email signature (and not the text of the email itself), reduces the probability of reply by 4.6 percentage points, an effect that is highly statistically significant ( $p < .001$ ). The effect size for atheist email signatures is very similar, reducing the probability of reply by 4.7 percentage points compared to the baseline condition ( $p < .001$ ). These effects are substantively meaningful. As a benchmark, they are three times as large as the gender effects we observe

---

<sup>61</sup>Probability estimates are simulated using the observed values approach. Estimated probabilities for a given factor or factor combination average over the remaining factors. Figure B.1 in the SI displays the underlying probability estimates.

<sup>62</sup>There is no evidence of an interaction between parent’s gender and child’s gender:  $\chi^2_1 = 0.19$ ,  $p$ -value = 0.67. There is also no evidence of interactions between the gender treatments and the religious affiliation/non-affiliation or intensity treatments:  $\chi^2_{24} = 26.73$ ,  $p$ -value = 0.32, suggesting that the discriminatory effects we observe are uniform across gender (a fact consistent with the other treatment heterogeneities we explore below). All Wald tests are two-sided.

in the same sample and are just over 10% of the baseline response rate. They are slightly smaller (but not statistically distinct from) the race/ethnicity effects reported in an audit study of state legislators (Butler and Broockman, 2011*a*). The results demonstrate that a clear bias exists against these minority groups and that (as best we can tell) this bias is of the order of the large race/ethnic biases found in studies of elected officials. This bias is present even when costliness is not explicitly signaled.

Our results are both substantively and statistically identical if we additionally control for block fixed effects (Table B.3). Using linear probability models instead of probit also does not affect our results (Table B.4). The same is true when we control for the covariates listed in Table B.1 (Table B.4). Finally, our findings are also completely unaffected by dropping Massachusetts from the sample (see discussion in footnote 52; Table B.5).

### **6.5.1. Potential mechanisms: Perceived costs of intense beliefs**

To unpack these effects, we next examine the intensity treatments that randomly assign principals to higher perceived costs either by inquiring about the compatibility of the school with the family’s beliefs or by requesting accommodation of the family’s beliefs. If the religious discrimination that we have observed were driven by perceived costs, we would expect to see larger effects among those who are in these treatment conditions.

In practice, this is exactly what we observe. In the medium intensity condition, in which parents inquire about the compatibility of the school with their beliefs, the extent of discrimination increases for all four religious affiliations/non-affiliations. Effect estimates increase to  $-8.7$  percentage points for Muslims and  $-13.8$  percentage points for atheists. This suggests that the public officials in our sample discriminate against Muslims and atheists (in part) because they perceive that serving such families will impose costs on them. Such costs could arise because these families are perceived to make illegitimate,

costly demands on schools or because other members of the school community might object to their presence, causing conflicts that principals would prefer to avoid completely.

Interestingly, we also observe a discriminatory effect for mainstream religious groups when they signal a greater intensity of belief. For Protestants, the estimated treatment effect is  $-5.4$  percentage points; for Catholics, it equals  $-6.6$  percentage points. These estimates (and those for Muslims/atheists) are statistically significant; all four are also more negative than the corresponding effect estimates in the low intensity condition ( $\chi_4^2 = 129.97, p < .001$ ). Moreover, the effect for Muslims is substantially larger (in absolute value) than the effects for Protestants and Catholics ( $\chi_2^2 = 6.80, p = 0.03$ ) and the effect for atheists is substantially larger than the effect for Muslims ( $\chi_1^2 = 17.72, p < .001$ ). This suggests that while mainstream religious groups are penalized for intense beliefs and the accompanying perception that they are difficult to deal with, Muslims and especially atheists are punished *even more so*. These results suggest that what we find for Muslims and atheists is religious bias magnified by perceived costs.<sup>63</sup>

Effect estimates for the high intensity condition are very similar to estimates for the medium intensity condition. We cannot reject the null hypothesis that effects in the medium and high intensity conditions are the same ( $\chi_4^2 = 0.95, p = 0.92$ ).

### 6.5.2. Benchmarking to theoretical expectations

Our results provide qualified support for all three hypotheses. Consistent with **H3**, discrimination against atheists is greater than discrimination against any religious group including Muslims, at least in the medium and high intensity conditions. When religious identity is merely signaled through email signatures, we still find sizable discrimination against athe-

---

<sup>63</sup>Like most audit studies, our design does not throw light on whether this perception of increased costs is itself rooted in anti-Muslim or anti-atheist sentiment on the part of the principals. This distinction does not matter for our finding of religious discrimination but would be important for policy interventions designed to reduce such discrimination.

ists, but it is not significantly larger than discrimination against Muslims (it is significantly larger than discrimination against Protestants and Catholics). As predicted by **H3**, we also find that discrimination against atheists increases as we move from the low intensity condition to the medium intensity condition. Unexpectedly, moving from medium intensity to high intensity did not further increase the extent of discrimination against atheists (or any other group). It is possible that differences in language between the medium and high intensity conditions were not sufficiently large to induce additional discrimination.<sup>64</sup> Alternatively, it is conceivable that the language we used for the medium intensity condition already implied a possible request for religious accommodation, so that principals refrained from responding to these emails at the same rate as in the high intensity condition.

In line with **H2**, we find that principals discriminate against Muslims even in the low intensity condition, where the only difference between Muslim parents' emails and Protestant and Catholic parents' emails is the email signature. Also consistent with **H2**, discrimination increases in the medium intensity condition. In all three intensity conditions, Muslim parents are less likely to receive a reply than Protestant or Catholic parents.

Finally, our results are also partially consistent with **H1**. While we fail to observe any discrimination against Protestant or Catholic parents in the low intensity condition, we do find that principals are significantly less likely to reply if parents inquire about the compatibility of the school with their beliefs or request accommodation of the family's beliefs. Overall, our results demonstrate more severe discrimination against Muslims and especially atheists than mainstream religious groups.

In the SI, we investigate treatment effect heterogeneity but find no evidence that treatment effects vary with the social context in which principals are embedded. We also

---

<sup>64</sup>When designing the experiment we considered even more strident language for the high intensity condition but decided against it in order to safeguard the realism of our emails and the internal and external validity of our experiment.

formally generalize our results to the NCES population of 78,348 regular, non-charter public schools in the 48 contiguous U.S. states without missing data. Results are virtually identical.

### 6.5.3. Limitations

Audit studies, and randomized trials more generally, are designed to identify the average causal effects of specific treatments — in our study, signals of religious affiliation/non-affiliation and intensity of belief embedded in parents’ emails. Our study is motivated by three theoretical frameworks drawn from the sociology literature (i.e., secularism, Judeo-Christian nationalism, civil religion). We have also tested one of the most theoretically compelling mechanisms (i.e., perceived costs) that could explain our results, but our experiment cannot conclusively demonstrate that this is the only mechanism in operation. Nor does our audit study speak directly to the many other situations in which principals (or other public officials) might engage in religious discrimination; indeed, our study does not speak directly to the question of whether students of different religious backgrounds are treated differently once they are enrolled in public school. That is a different question over which our research design does not afford us any leverage. Our study shares these limitations with other audit studies and randomized trials more generally (Green, Ha and Bullock, 2010; Guryan and Charles, 2013).<sup>65</sup> While it is important to note these limitations, our paper makes an important contribution by documenting for the first time that significant religious discrimination takes place when American parents interact with street-level bureaucrats.

---

<sup>65</sup>Take audit studies of wage discrimination for example. Bertrand and Mullainathan (2004*b*) show that black and white job applicants with otherwise identical resumes are treated differently. It is possible of course that employers discriminate against black workers when they review resumes but that black workers are treated the same as white workers once they are hired. This possibility does not negate the fact that employers engage in illegal discrimination in at least one domain.

## 6.6. Conclusion and implications

In this paper, we have demonstrated that in spite of legal protections, systemic bias against Muslims and atheists exists in the U.S. public domain. Our large-scale audit study provides us with clear evidence that religious discrimination is large and widespread. Such discrimination appears to be driven, in part, by a perception that these groups make illegitimate demands that impose costs on public officials. Our results are fundamentally important. They demonstrate that qualitative/survey-based evidence of religious discrimination is, indeed, evidence of a broader pattern of discrimination towards religious out-groups. Our results find express meaning in a context of growing hostility towards groups like Muslims and atheists, while also suggesting that mainstream religious adherents are penalized as well when they expressly mention their faith in interactions with public bureaucrats.

Our work also speaks to public policy in important ways. PK-12 principals occupy a challenging role as mediators between teachers and parents and schools and the public. Their task is complicated by societal changes that unsettle the established moral consensus and strain the capacity of public schools to ensure equal and fair treatment. Principals must simultaneously respect the separation of church and state and safeguard the protection of individuals' religious liberties. Discrimination on the basis of religious affiliation/non-affiliation could damage educational performance and attainment by undercutting school attachment and academic engagement (Johnson, Crosnoe and Elder Jr, 2001) and by posing a substantial barrier to parental involvement in schooling (Turney and Kao, 2009).

Because education is a primary factor in occupational and income mobility as well as formal socialization, the potential consequences of discrimination in this domain are far-reaching. As leading experts in education law observe, the U.S. is “a country that has developed an extraordinary reliance on public schools as a mechanism for social and economic justice and improvement” (Alexander and Alexander, 2011, xxxvii). A prominent

historian of education observes that “[i]n the early twenty-first century, America’s schools remain central to most public debates over how to define and secure the good life for the nation’s children” (Reese, 2011, 8). Religious discrimination by public school principals raises serious concerns about one of the fundamental institutions of modern democratic government and its capacity for integration and equity.

Although previous research has shown a prevalence of hostile and exclusive attitudes toward religious minorities and atheists and demonstrated religious biases in hiring, to the best of our knowledge our study reports the first experimental research examining religious discrimination in the American public school system. Now that we have established that such discrimination is common when parents interact with school principals in the process of school enrollment, future research should make it a priority to investigate further *when* and *why* principals engage in such discrimination, *whether* other public officials — both elected and appointed — also engage in religious discrimination, and *how*, if at all, such discriminatory behavior may be best remedied.



**Figure 6.1: Email to principals**

Subject: School visit?

Dear principal,

Hello. My family and I will be moving into the area sometime this summer. Right now, we are deciding where exactly to move and are looking at schools for our [son/daughter], [Jonah/Sarah]. Before we pick a place to live, we would like to meet with you or a member of your staff and chat a bit about your school. Would that be possible?

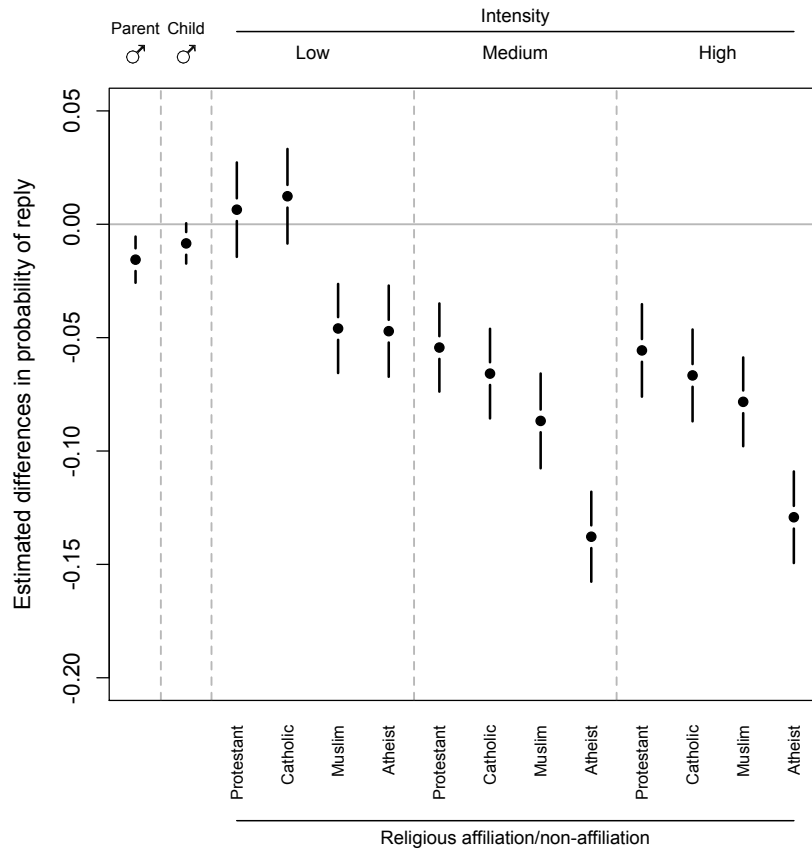
[A] [One of the reasons we would like to meet with you is that we are raising [Jonah/Sarah] to be a good [Christian/Catholic/Muslim/Atheist Humanist] and want to make sure that this would be possible at your school.]

[B] [One of the reasons we would like to meet with you is that we are raising [Jonah/Sarah] to be a good [Christian/Catholic/Muslim/Atheist Humanist] and want to protect [him/her] from anything that runs counter to our beliefs. We want to make sure that this would be possible at your school.]

Sincerely,  
[Isaac Adam/Rebekah Adam]

[C] [Catholicism/Christianity/Islam/Atheism teaches that life is precious and beautiful. We should live our lives to the fullest, to the end of our days. - Pope Benedict/Rev. Billy Graham/The Prophet Muhammad/Richard Dawkins]

Figure 6.2: Estimated treatment effects based on model in Table B.2



Note: The plot shows estimated differences in probabilities of receiving a reply (i.e., treatment effects) and 95% confidence intervals based on the probit model in Table B.2. Robust standard errors are clustered at the school district level.

# **Chapter 7. Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions**

This paper discusses how audit studies can be adapted to test the effectiveness of interventions aimed at reducing discrimination. While the approach we describe can be applied to studies of discrimination in many areas, we focus on racial discrimination because it affects numerous facets of life including — but not limited to — housing Turner et al. (2002), employment (Bertrand and Mullainathan, 2004*a*; Gaddis, 2014; Pager, 2007*b*), health (Burgess et al., 2007), and civic life (Griffin and Newman, 2008). Audit studies have become an increasingly popular tool for measuring discrimination across the social sciences (Vuolo, Uggen and Lageson, 2016)). Most audit studies, however, simply focus on measuring the extent to which bias exists. We build on those works by showing how audit studies can be combined with randomly-assigned interventions as a way of testing

---

This chapter is adapted from Butler and Crabtree (2017*a*).

the effectiveness of those interventions.

We implemented an adapted audit study to test whether information can be used to reduce levels of racial discrimination that local officials exhibit towards constituents seeking answers to basic questions. We focus on an informational treatment because evidence suggests that many people do not realize that they are acting in a biased manner (Dovidio and Gaertner, 2004). Consequently, some have advocated using information to make people aware of their biases (Burgess et al., 2007; Devine and Monteith, 1999; Rudman, Ashmore and Gary, 2001; Pope, Price and Wolfers, 2013). For our study, we provided a random subset of the local officials in our sample with information about previous evidence of racial bias in how elected officials respond to citizens' requests. A few weeks later we then conducted an audit study to measure levels of racial bias among all the officials in our sample (Adida, Laitin and Valfort, 2010; Broockman, 2013; McClendon, 2012), both those who were sent the information and those who were not sent the information. We did this by sending an email to each public official in our sample, randomizing whether the request for help came from a putatively black or putatively white constituent. We then measured the level of racial bias by comparing the response rates to the white and black constituents.

Our audit study successfully replicates prior work in this area by showing that officials exhibit bias against black constituents. Indeed, the extent of bias is comparable to levels found by previous studies of officials at other levels of government. Because local officials often help constituents with the services that they use most regularly, discrimination at this level of government raises concerns about fairness in representation.

Our results regarding the effect of information on reducing bias are less conclusive. In our study, the information treatment group and the control group exhibited the same level of racial discrimination. While this is consistent with the possibility that information alone is insufficient to reduce bias, there are limitations with the design that may also be

responsible for the null result. After presenting the details and results of our adapted audit study, we discuss those limitations and what may have been done differently.

We end our paper with a list of concrete recommendations for how to improve future audit studies. Our points draw on the lessons learned from both the limitations and strengths of our particular study. We provide these suggestions as a way to improve future work on identifying and reducing discrimination.

## **7.1. Testing the Effect of Information**

We use an audit study to test whether information can reduce racial bias. In our audit study, we sent elected, municipal officials (i.e., mayors and city councilors) in the United States requests for assistance from putative constituents, randomizing whether the request came from someone with a distinctively black name or a distinctively white name.

We chose our sample to learn about how white, elected officials at the municipal level treat white and black constituents. We focus on local officials because they are responsible for the provision of many important public services and oversee the government employees with whom citizens most regularly interact. Because we could not identify the race of all city officials, we restricted our sample to the types of cities where, based on the racial make up of the city, the vast majority of elected officials are likely to be non-Latino, whites. This is important from a methodological perspective because if the sample represented a more racially diverse population, the in-group bias exhibited by officials from different racial and ethnic groups may have cancelled each other out (Broockman, 2013).

We used the 2011 International City/County Management Association (ICMA) city survey, which includes information on the racial demographics of the elected officials, and data from the U.S. census to determine which cities to include in our sample. Based on these data, we restricted our sample to cities where 75 percent or more of the population is

white and less than 15 percent of the population is Latino. On average, 96 percent of the officials serving in cities with these racial demographics are non-Latino, whites (see Figure C.1 in the SI). We also restricted our sample to officials from cities with at least 50 black individuals and a total population of 3,000 or more.<sup>66</sup> We use these cut points to increase the plausibility that municipal officials would receive an email from a black constituent that they did not know. Our final sample included 11,801 city officials<sup>67</sup> from 2,160 cities from across the United States.

The email addresses for the elected city officials were collected by research assistants through web searches. Research assistants first searched for the website of each town or city taken from the census. If the research assistants were able to identify the city's website, they then collected the name and email address of the city's mayor and council members (or the equivalent).

To measure levels of differential treatment by race, we emailed all of the 11,801 officials in the sample, randomizing whether the email was sent from either a putatively black or a putatively white individual. We used information from previous studies and the U.S. Census to identify common first and last names for black and white individuals (Butler 2014). In total we used 76 different aliases: 35 black aliases and 41 white aliases (see Appendix E in the SI for list of aliases). For each alias we created a separate Gmail account. When the emails were sent, the alias, which signaled the individual's race, was included both in the field identifying the sender and in the salutation in the body of the message. We carried out the study during the Fall of 2014.

We carried out the study so that no two officials from the same city received a request from the same putative constituent. In particular, we first assigned officials to receive either

---

<sup>66</sup>The mean city in our sample has 23,949 residents, of whom 4.23 percent are black. Cities range in size from 3,011 to 583,776 people, and the black population within cities ranges from 50 to 42,188.

<sup>67</sup>This excludes the 338 officials with bad email addresses.

the black or white constituent email and then block randomized, by city, aliases from the assigned racial treatment. Similarly, the request found in each email was randomly drawn from a list of simple requests adapted from the ‘frequently asked questions’ sections of various city websites. We used 27 different requests, because all of the cities in the sample had 27 or fewer elected officials, and again block-randomized so that no two officials from the same city received the same request (see Appendix F in the SI for the full list of requests used). We also randomized when we sent the emails so that they went out over a five-day period. We spread the emails out to help ensure that no city in the study received all of the emails on the same day.

About two weeks earlier, we sent the elected officials assigned to the treatment group ( $n = 4,004$ ) a pair of emails with information about prior research on racial bias exhibited by public officials.<sup>68</sup> The information email we sent to the treatment group was ostensibly an invitation to take a survey about recent research on racial bias in how officials deal with constituent communications. We used this approach because it provided a rationale for sharing information about previous research, while minimizing the possibility that officials would assume that we were monitoring their behavior. The email states that research has shown “that office holders respond more often and provide better advice to individuals like them. White legislators, for instance, respond at higher rates to inquiries from white constituents than from black constituents.” We underlined this sentence and placed it in bold to draw attention to it. Within the main body of the email, we also provided links to (a) an NPR report on research showing racial bias among public officials and (b) a website we created to further highlight the findings from the racial bias literature (see <http://n.pr/1SRr6VA> for the NPR report and Appendix C in the SI for screenshots of the website). The sites that we linked to focused on the empirical finding that white legislators

---

<sup>68</sup>The only difference between the waves was that the second email began with “Thanks to those who have already read this email and taken the survey. If you have not, please continue reading”.

respond at higher rates to inquiries from white constituents than from black constituents, reinforcing the information provided in the email. Below the valediction of the email, we also included a one-paragraph summary of findings from the racial bias literature (see Appendix B in the SI for full email text).

Finally, we designed the study so that we could test both the direct effect of receiving the information on the recipient's behavior, and also assess whether there was any spillover effect that changed the behavior of the other officials in the same city. We are able to test for this type of contagion effect because we used a multilevel design (Sinclair, McConnell and Green, 2012) where we first randomly assigned the cities into two groups: in one-third of the cities no officials received information; half of the officials in the remaining cities were randomly chosen to receive the information. After excluding the bad email addresses, we had three randomly chosen groups of officials: (a) 4,004 officials who received the information, (b) 3,963 officials who did not receive the information, but who serve with officials that did, and (c) 3,834 officials who did not receive the information and whose colleagues also did not receive the information. If the information had a direct effect on reducing bias, but no spillover effect, then the officials in group (a) should exhibit less bias than those in (b) and (c) and there should be no difference between (b) and (c). If there is a spillover effect then the people in (a) and (b) should both exhibit less bias than officials in group (c). And if there is no effect at all, then all three groups should exhibit similar levels of bias.

## 7.2. Results

Of the 11,801 contacted elected officials, 7,135 sent at least one email within two weeks of receiving our requests, for a response rate of approximately 60 percent.<sup>69</sup> However,

---

<sup>69</sup>We do not count autoresponses as replies.



in line with previous research, not all putative constituents were treated equally. Elected municipal officials responded to emails from white constituents about 63 percent of the time (3,732/5,908) but only responded to emails from black constituents about 58 percent of the time (3,407/5,893). The five percentage point difference [3.6-7.1, 95 percent confident interval] between response rates is statistically significant<sup>70</sup> and in line with other findings on racial bias among elected officials (Butler, 2014; Costa, 2017). A treatment effect of this size suggests that blacks and whites receive quite different treatment by local elected officials. All discrimination by elected officials harms the individuals discriminated against as well as the broader society, but discrimination by *local* elected officials can be particularly harmful. This is because the actions of local officials influence the lives of their constituents in fundamental ways: they administer many of the public services that constituents use, they oversee many of the street-level bureaucrats that constituents interact with, and they are often the first to be contacted by constituents who need assistance. It should then cause considerable concern that local officials treat constituents differently on the basis of race.

Unfortunately, providing information about racial bias to public officials did not appear to change their behavior. Table 7.1 gives the mean response rates by treatment, as well as the differences between how they treated black and white aliases. All three groups in our sample were about five percentage points more likely to respond to an email from a white alias than they were to respond to an email from a black alias. Because there does not seem to be evidence of a contagion effect, Table 7.2 directly compares the level of bias among those who were assigned to be made aware of the research and those who were not sent any information. Individuals assigned to receive the information treatment exhibited racial bias that was about 0.7 percentage points greater than those exhibited by

---

<sup>70</sup> $t = 5.96, p \leq 0.00.$

the control group. Regardless of which way the control group is defined, the level of racial discrimination is comparable across the different groups. There are small differences in the level of bias (ranging from 4.7 to 5.8 percentage points), but the differences between these groups are statistically insignificant and point in the wrong direction: the largest difference was observed among those who were sent information. Being made aware of previous research on bias did not decrease the level of racial bias and so did not appear to have much of a spillover effect.

**Table 7.1: Email Replies from Elected Municipal Officials**

	Directly Informed	Colleagues Informed	No one in city Informed
Response Rate to Black Names	57.0% [N=2,017]	58.6% [N=1,949]	57.8% [N=1,927]
Response Rate to White Names	62.8% [N=1,987]	63.3% [N=2,014]	63.4% [N=1,907]
Racial Bias	-5.8* (1.5)	-4.7* (1.6)	-5.6* (1.6)
Bias Difference: Directly - No One (Std. Error)		-0.2 (2.2)	
Bias Difference: Colleagues - No One (Std. Error)		0.9 (2.2)	

*Note:* Standard errors are shown in parentheses. \*  $p < 0.05$ , two-tailed. The dependent variable, *Email Reply*, is coded as 1 if an elected municipal official sends a non-automated response to an information request within two weeks of receiving our email and 0 otherwise. While it appears that the directly treated group responded at a lower rate than the spillover or control groups, this difference is not statistically significant at any meaningful level.

**Table 7.2: Email Replies from Elected Municipal Officials**

	Directly Informed	Not Directly Informed
Response Rate to Black Names	57.0% [N=2,017]	58.2% [N=3,876]
Response Rate to White Names	62.8% [N=1,987]	63.4% [N=3,921]
Racial Bias	-5.8* (1.5)	-5.1* (1.1)
Bias Difference: Directly Informed - Not Informed (Std. Error)	-0.7 (1.9)	

*Note:* Standard errors are shown in parentheses. \*  $p < 0.05$ , two-tailed. The dependent variable, *Email Reply*, is coded as 1 if an elected municipal official sends a non-automated response to an information request within two weeks of receiving our email and 0 otherwise.

### 7.3. Limitations

Our treatment, however, might have been too weak to cause a noticeable effect. First, not every official assigned to receive the information about previous results was actually exposed to that information. Some individuals simply did not open either of the two emails we sent. We know this because we sent the emails with information on previous research through an email service that tracks whether recipients open their email. In our sample, 53 percent of the group assigned to receive the information treatment opened one or more of the emails with this treatment.

Second, even those who opened our treatment emails might not have read far enough to see the treatment. The key piece of our informational treatment appears in the third sentence of the email. Although we bolded and underlined the treatment text, we cannot

know if people read past the first two sentences.

Third, even if officials did read our informational treatment, they might not have paid attention to it. One reason why is because the email referred to scholarly research and officials might generally be uninterested in research. Another reason is that officials might not believe that the findings discussed in the email apply to them personally. While we cannot know for certain whether individuals carefully read our treatment emails, or thought that the content applied to them, we have anecdotal evidence that this might not be the case. We know that only 18 percent of the treatment group visited the survey link that was in the email, that only 15 percent of the treatment group completed the survey, and that only 4 percent of the treatment group click on the link to the research summary. Taken together, these statistics suggest that officials were not particularly interested in the information that we provided them.

The timing between our treatment and audit emails is another potential limiting factor. We sent the informational email to the treatment group a couple of weeks before conducting the audit because we wanted to assess whether our informational treatment had a meaningful long-term effect. Because our treatment was weak, it might have been unreasonable to expect that it had such a lasting effect. Perhaps if we would have sent the audit emails shortly after our treatment emails, we would have found an effect.

Further, when conducting audit studies with elites' email addresses, researchers can rarely be certain that the individual they email is the one that reads and responds to their email. Elites often employ staffers to handle their correspondence. It is possible that one staffer read the informational, treatment email and that another dealt the audit email. This might occur if staffers rotate email correspondence duties, there is turnover in staff, or staffers go on vacation.

## 7.4. Suggestions for Future Audit Studies

Below is a list of suggestions for future audit studies. These suggestions are drawn from both what worked in our study and from what could have been improved. We discuss our suggestions in relation to our study to provide concrete examples of how these points could be implemented.

1. **Put the treatment at the forefront.** The information treatment may simply have been missed. If we were to do this again, we would place the treatment language as close to the top of the email as possible. If it could not be placed in the first couple of sentences, we would put the message in color or highlight it.
2. **Increase the relevance of the treatment.** Individuals might not have thought that the information in the treatment applied to them. One could address this issue by increasing the perceived importance or influence of the sender. Our treatment may have been more powerful if the sender had been an individual or representative of a group from the same general geographic area, or from the local officials own constituency. It also might have been more powerful if it was sent from someone who could credibly sanction the officials if their future behavior was deemed inappropriate.
3. **Carefully consider how long to wait between the treatment intervention and the audit study.** Researchers are more likely to find treatment effects if the interval between treatment and the outcome is short, but in many cases scholars might care more about the long-term effects of treatment. It is not always clear, unfortunately, how long treatment effects should last to be substantively meaningful. The proper length of time should be carefully considered at the design stage.
4. **Consider automatically coding the outcome measure.** We created our dependent variable by reading 8000 email responses, identifying the sender, and then

coding whether they had replied to our email.<sup>71</sup> As researchers who conduct email audit studies know, this can be a time-consuming procedure. If we were to do this again, we would use software to automatically process email responses and construct a response indicator. Our replication files contain an annotated R script and example data file that researchers can adapt for this purpose.

5. **Have a sufficiently large pool of email texts.** If officials think they are being studied, they are less likely to exhibit socially undesirable behavior (Findley, Nielson and Sharman, 2015). This can be a problem for audit studies that send the exact same text to all officials. This is because officials might share the communications they receive (or share staff who respond to emails). If this is the case, individuals might know that they are being studied and may simply respond differently to those communications. Having a sufficiently large number of number of different email texts helps minimize this threat to validity. Ideally one would have enough different texts so that the officials who talk with each other do not receive the exact same email message. In our case, we had enough different email texts so that no one serving in the same city received the same message. In developing these questions, we looked for city pages that included answers to “frequently asked questions”. Those questions then helped us to write questions that would be relevant to our target population, thereby increasing the believability of our deception.

6. **Have a sufficiently large number of aliases and email accounts.** This will also minimize the possibility that officials respond to the emails differently than they normally would.

7. **Consider spillover at the design stage.** One of the advantages of studying public

---

<sup>71</sup>Some officials sent more than one reply.

officials is that they have publicly observed social networks. Researchers can use the design proposed by (Sinclair, McConnell and Green, 2012) to test for spillover in those networks.

8. **If the treatment designed to lower bias is delivered via email, track who opens the email.** Because we tracked who opened the emails, we were in a position to discuss what the bounds of the effect might be. If we were to do it over again, we would have also sent a placebo email (which included an invitation to participate in a survey on a different topic) to the control group and tracked who opened those emails. This would have given us a placebo design that would have increased the power of our analysis (Nickerson, 2005; Gerber and Green, 2012).

## 7.5. Conclusion

We conducted an adapted audit experiment to test whether making officials aware of bias could reduce levels of racial bias. Although limitations in our design make it difficult to assess whether information alone can reduce bias, our study makes two important contributions. First, we find that white, local, elected officials in the United States are less responsive to black constituents. The extent of this bias is in line with prior studies (Butler, 2014; Broockman, 2013; Costa, 2017), suggesting that blacks face a similar degree of discrimination across governmental contexts. The fact that we find such bias among local government officials is worrying, however, as local government is often the level that most directly affects citizens' daily lives. Reducing this bias should be an important goal.

Second, we have described how audit studies can be adapted to help learn what measures help reduce bias. We draw these lessons from what worked in our experience and from what did not work. We offer these suggestions as a way to help improve future work on

identifying and reducing discrimination.



# Chapter 8. How Public Opinion Shapes Discriminatory Policing

## 8.1. Introduction

On July 17, 2014, a police officer put Eric Garner, a 33-year-old, unarmed black man who was selling untaxed cigarettes, into a deadly chokehold (Marcus, 2016). Less than a month later, a law enforcement officer in Ferguson, Missouri fatally shot Michael Brown, an 18-year-old, unarmed black man (Buchanan et al., 2014). Garner and Brown's deaths sparked a series of protests that provoked an ongoing national debate about racial biases in policing and propelled the Black Lives Matter movement to political prominence (Jee-Lyn García and Sharif, 2015). Despite the administrative and legislative reforms that followed from these incidents and the public reaction to them, police killings of racial minorities continue to make the news with alarming frequency (Bonilla-Silva, 2017, 40-43). While racial discrimination by the police is seen vividly in the United States, similar forms of discrimination are common within a whole host of democracies. For example, there is evidence that domestic security forces in countries as diverse as Canada (Maynard, 2017), India (Subramanian, 2007), Israel (Davis, 2016), and Sweden (Kauff, Wölfer and Hewstone, 2017) disproportionately punish members of minority groups. In some cases, this means

that individuals from these groups are shot or ‘tased’ more frequently, while in others it means that they are harassed, dispossessed of their property, or ticketed at higher rates.

Why do the police discriminate against racial minorities?<sup>72</sup> Over the last two decades, social scientists have used new data sources and an increasingly sophisticated set of methods to study racial discrimination across many contexts (Sen and Wasow, 2016).<sup>73</sup> Some of this empirical work investigates the extent of racial bias among non-state actors, such as employers and landlords (Gaddis, 2018*b*). Researchers, for example, have documented discrimination with respect to labor (Bertrand and Mullainathan, 2004*a*; Gaddis, 2014; Pager, 2007*a*; Pager and Quillian, 2005; Pager, Bonikowski and Western, 2009; Crabtree, Hou and Liu, 2018; Maurer-Fazio, 2012), goods (Michelitch, 2015), housing (Berry et al., 2011; Einstein and Glick, 2017*b*; Ghoshal and Gaddis, 2015), credit (Ross and Yinger, 2002), education (Milkman, Akinola and Chugh, 2012; Crabtree, Hou and Liu, 2018), and even ride-share markets (Ge et al., 2018). Other work examines discrimination by state actors, such as political elites. This line of research has identified racial biases among elected (Butler, 2014; Butler and Broockman, 2011*a*; Butler and Crabtree, 2017*b*; Alizade, Dancygier and Dittmann, 2018; Costa, 2017; Gell-Redman et al., 2018*b*; Grose, 2014) and unelected officials (Einstein and Glick, 2017*b*; White, Nathan and Faller, 2015*b*; Hughes et al., 2017; Distelhorst and Hou, 2017). Taken together, the results of these and many other studies show that racial discrimination on the part of state and non-state actors has a significant impact on the quality of citizens’ lives (Crabtree, 2018).

---

<sup>72</sup>By police, I refer here to domestic security forces broadly conceived. This definition encompasses police, militia, para-police, and other agents of the state authorized to use force in the pursuit of crime or social control. Unlike the military, which states typically use to exert coercion outside the state or protect the state’s boundaries, the state typically uses the police to exert coercion or threat within the state. In line with the literature (Butler, 2014), I refer to bias and discrimination interchangeably. Following Sen and Wasow (2016), I use race as a shorthand for both race and ethnicity.

<sup>73</sup>There has been a similar growth in research on discrimination against immigrants (Hainmueller and Hangartner, 2013; Hainmueller and Hopkins, 2015; Dancygier and Laitin, 2014; Baert and Vujić, 2016) and religious minorities (Crabtree et al., 2018; Pierné, 2013; Adida, Laitin and Valfort, 2010), reflecting an increasing concern with group-based inequalities across the social sciences.

Until recently, there has been relatively little research that focuses specifically on racial discrimination by the police.<sup>74</sup> This is surprising because the police are an important state actor. Almost all contemporary definitions of the modern state emphasize its reliance on coercion and its use of violence (Weber, 1965). State coercion and violence in the domestic arena is primarily enacted by the police. Since the police have the capacity to restrict and violate the physical integrity of others, we might be particularly concerned about whether law enforcement personnel exhibit discrimination. They often have the opportunity to do so. For example, about 20 percent of all American adults have some kind of encounter with the police in a given year (Eith and Durose, 2011). Public-police interactions occur about as frequently in other countries. It has been estimated, for instance, that one in two citizens in Belgium, Finland, and Sweden are approached, stopped, or contacted by domestic security forces in a two-year time period (Staubli, 2017). While similar statistics for other countries, particularly autocracies, are often unavailable, we can imagine that these types of interactions are just as common elsewhere. This would mean that every year millions of people, perhaps hundreds of millions, cross paths with the domestic security services of their states. Since the police wield considerable power and exercise it frequently, it matters if the police are treating different groups unequally.

What research there is on the police tends to focus on identifying whether racial discrimination exists (Edwards, Esposito and Lee, 2018; Ray, Ortiz and Nash, 2018; Ross, 2015; Gelman, Fagan and Kiss, 2007; Baumgartner, Christiani, Epp, Roach and Shoub, 2017; Epp, Maynard-Moody and Haider-Markel, 2014, 2017; Baumgartner, Epp, Shoub and Love, 2017; Bowling, 1990; Sun, Wu and Hu, 2013; Cashmore and McLaughlin, 2013; Bigo and Guild, 2005). A handful of studies have begun to go further and look at whether

---

<sup>74</sup>A quantitative analysis of the literature shows that while there was some early work on this subject (Lohman and Reitzes, 1952; Furstenberg and Wellford, 1973; Hadar and Snortum, 1975; Rafky, 1975; Hahn, 1971), there has been considerably more attention to it in recent recent years, though most of this been in sociology. For more details, see Appendix A.

certain behavioral interventions, such as implicit bias trainings (Smith, 2015*b*; Nix et al., 2017; Fridell, 2016; Spencer, Charbonneau and Glaser, 2016), or specific institutions, such as oversight boards and the courts (Kennedy et al., 2017*a*), can reduce racial discrimination in the police. While studies on racial discrimination in the police have occurred in many countries, such as Canada (Maynard, 2017) and Brazil (Mitchell and Wood, 1999), the majority have focused in some way on the American context.<sup>75</sup>

While these new studies have produced important insights, notably that racial discrimination in the police is common in many settings, this work has at least four limitations. First, many studies rely on administrative data provided by law enforcement agencies themselves. In most countries, though, we know that these self-reported data are often biased in important ways (Davenport, 2005). For example, they tend to exaggerate potential threats to police officers and downplay the violence of police actions (Sullivan and O’Keeffe, 2017; Davenport, 2005).<sup>76</sup> Second, most studies rely on observational data. This means that it is hard to know if the differences we see in how the police treat racial groups are actually caused by race as opposed to some other reason, such as socioeconomic status, class, or education. Third, research that does leverage creative designs to isolate the causal effect of group identity on policing practices has reached conflicting conclusions. In some cases, it appears that the police do discriminate (Zimring, 2017). In other cases, though, there is no evidence of police discrimination (Fryer, 2016). One explanation for these conflicting findings is that discrimination is context dependent. If this is true, then researchers need to use research designs that eschew standard additive models and experimental designs in favor of interactive models and factorial designs. The fourth and possibly largest limitation with

---

<sup>75</sup>Over half of all police related articles published in a political science journal since 1980 have examined the police in the American context (Crabtree, 2018, 5).

<sup>76</sup>As Eck and Crabtree (2018) note, Sweden represents an important exception to this general trend. The Swedish government collects comprehensive data on policing practices and provides nearly all of it to the public upon request.

existing research, though, is that it does not isolate the mechanism(s) driving any observed discrimination. In other words, the empirical literature has largely focused on identifying where and when discriminatory policing occurs, rather than on investigating *why* it occurs. Without a better understanding of the factors that drive police discrimination it is hard to determine the most appropriate policy interventions to reduce it.

According to the existing literature, which frequently employs a principal-agent perspective (Fagan, 2017; Balko, 2013),<sup>77</sup> the primary explanation for racial discrimination in policing is that officers are bad agents. This could be because individuals prone to implicit or even explicit biases against racial groups choose to enter into policing (Engel and Swartz, 2013; Smith, 2015*b*; James, Vila and Daratha, 2013) or because bad institutional environments and poor training might instill or reinforce racially biased views in law enforcement officers (Wilson, 1978; Shusta et al., 2002; Smith and Alpert, 2007). In either case, the result is agency loss.<sup>78</sup> According to this framework, police discrimination can be reduced by adopting more stringent tests at the hiring stage, increasing the diversity of police personnel, developing effective implicit bias training, and actively promoting a more racially tolerant organizational culture. In effect, we need to increase the supply of good agents — unbiased officers. Doing this is difficult, though, and the effectiveness of implicit bias training, cultural reforms, and increased diversity within the police force remain contested. Police discrimination appears to have been reduced in some contexts, but not in others (Vedantam, 2008; Peruche and Plant, 2006; Bregman, 2012; Butz et al., 2018).

An assumption, often left implicit, in these studies is that the principal — the public

---

<sup>77</sup>Contributions to this literature typically evaluate law enforcement officers as agents, even if they do not adopt principal-agent language.

<sup>78</sup>Agency loss is typically defined as the “difference between the actual consequence of delegation and what the consequence would have been had the agent been perfect” (Clark, Golder and Golder, 2017, 501). In the context of policing, agency loss is then the difference between what the police do and what the public would want them to do.

— wants racial equality with respect to policing outcomes. This is, after all, why police discrimination is considered a form of agency loss. But this premise is rarely justified theoretically or empirically. There are reasons to believe that much of the public may not always hold egalitarian preferences (Staats et al., 2015; Stephens-Davidowitz, 2014; Northup, 2010). In comparison to earlier work, I argue that discriminatory policing depends on the interaction between (1) unbiased police officers, (2) public egalitarianism, and (3) police accountability. Racial discrimination will be highest when both the supply of unbiased officers and the public demand for equality are low. It will be lowest when both the supply of unbiased officers and the public demand for equality are high. And it will be moderately high if either the supply of unbiased officers is low or the public demand for equality is low. The impact of public preferences on police discrimination is strongest when when the police can be held accountable for their actions through electoral means or when there are electoral incentives for intermediaries, such as political officials, to oversee police actions.

This theoretical framework highlights the role that public opinion plays in discriminatory policing. There is a large literature that examines the congruence and responsiveness of various policy and distributional outcomes to citizen preferences (Kelly and Enns, 2010; Burstein, 2003; Risse-Kappen, 1991; Lax and Phillips, 2009; Soroka and Wlezien, 2010; Erikson, Wright and McIver, 1989; Golder and Stramski, 2010; Golder and Ferland, 2017). This literature has not examined whether the level of police discrimination is responsive to citizen preferences. Similarly, the growing literature on racial discrimination in policing has not taken into account the possibility that public attitudes might in part fuel or limit the racial injustices in policing that we observe.

To test the implications of my theory, which states that the interaction between supply of unbiased police officers and public demand for racial inequality drive police discrimina-

tion, I conduct survey experiments with two samples of American political elites — law enforcement administrators and elected officials who oversee the police. To the best of my knowledge, this is the first experiment on racial discrimination conducted with law enforcement administrators, a particularly consequential class of public bureaucrats.<sup>79</sup> I focus on these two groups because they are important intermediaries between the public (i.e. the ultimate principal) and front-line police (i.e. the ultimate agents). I focus on the US because this is a particularly important case given the historical significance of race relations in general and because the US is the most examined country in the policing literature (Crabtree, 2018). In each experiment, I randomly provide information to respondents about public demand for racial equality in their jurisdiction. I then ask subjects to evaluate two citizen-police interactions, randomly varying the race of the citizen. As predicted, elected mayors and state legislators who oversee the police, exhibit racial discrimination with respect to Blacks and Hispanics, but less so when they are informed that voters in their districts support racial equality in policing. In partial contradiction to my theory, though, police chiefs and sheriffs, do not exhibit discrimination against minorities and their behavior is not influenced by public preferences for racial equality. Overall, my results suggest that public views about racial equality influence discrimination by the police only indirectly, through the elected institutions that monitor and check their power. This implies that if we desire substantial change in police discrimination, we need to change the racial biases held by both the police *and* the public. Without addressing the public’s attitudes, racial discrimination is likely to endure since the police and their elected supervisors respond to and are held accountable by public preferences. This has important implications for future work as it suggests that we need to better understand what the public wants the police to

---

<sup>79</sup>While police chiefs, sheriffs, and other local-level law enforcement administrators sometimes appear in general discussions of policing (Brehm and Gates, 1993; Wilson, 1978), the most research on policing in the United States and elsewhere focuses on front-line officers (Manning, 2005; Earle, 1988; Smith, 1940).

do before evaluating police behavior.

## 8.2. Theory

The extent of racial discrimination in policing outcomes depends not only on whether law enforcement personnel think in biased ways but also on whether the public wants groups to be treated equally. In other words, it depends on the interaction between the supply of unbiased officers and the demand for unbiased policing. It also depends on the extent to which the police and their supervisors are responsive to citizen preferences.

### 8.2.1. Supply of Unbiased Policing

The literature on policing often adopts a principal-agent framework (Waterman and Meier, 1998; Brehm and Gates, 1999; Dharmapala, Garoupa and McAdams, 2016; McAdams, Dharmapala and Garoupa, 2015; Conrad, 2018). The general view is that the ultimate principal, the public in a democratic context, delegates authority via intermediaries to domestic security forces so that they can accomplish some set of tasks (Brehm and Gates, 1999). These duties typically include maintaining order, enforcing the law, and providing community service (Wilson, 1989). The public, however, faces two primary problems when they cede authority to the police. The first, adverse selection, occurs because the public has incomplete information about possible law enforcement personnel and cannot screen out ‘bad’ types, who might not act in accordance with their preferences (Akerlof, 1978). The second, moral hazard, occurs when the public cannot fully observe, and therefore sanction, ‘bad’ police actions.<sup>80</sup> While both problems can occur when voters delegate to

---

<sup>80</sup>A second-order problem is that sanctioning the police for bad behavior is difficult. The public typically relies on the police to enforce the delegation contracts that they make with other agents, such as bureaucrats and elected officials. The issue in the law enforcement case, though, is that the police have little incentive to enforce something against themselves (Monkkonen, 1981).



police forces, the likelihood of moral hazard problems is particularly high for several reasons — the police need a great deal of discretion to perform their duties (Lipsky, 2010), they possess much better information about their actions than principals (Hölmstrom, 1979), and they have historically been able to avoid direct oversight since that would be either prohibitively costly or impractical (Wilson, 1978). Some combination of adverse selection (Brehm and Gates, 1999) and moral hazard (Mas, 2006) can lead to agency loss (McAdams, Dharmapala and Garoupa, 2015), defined in this context as the distance between what the public wants the police to do and what they actually do (Gailmard, 2012).

Arguably, agency loss in policing is a more serious problem than in most other delegation contexts, as the consequences of it are potentially very high. When the public view the police as acting independently (as ‘rogue’ rather than perfect agents), they are more likely to question their legitimacy and that of other state institutions (Davis, 2017). The downstream consequences of a disaffected citizenry are manifold, particularly in democracies where challenges to the state’s legitimacy can lead to support for undemocratic alternatives (Foa and Mounk, 2016). In recent years, researchers across political science, economics, criminology, public administration, public policy, and sociology have begun to pay more attention to adverse selection and moral hazard problems in the context of policing and their possible remedies (Butler, Gluch and Mitchell, 2007; Soss and Weaver, 2017; Reiner, 2010).

In the context of racial discrimination among the police, a growing literature centers on identifying the extent to which domestic security officers engage in what is often termed ‘taste-based discrimination’ (Becker, 2010) or ‘biased mental processes’ (Correll et al., 2007, 2014; Warren et al., 2006; Tomaskovic-Devey, Mason and Zingraff, 2004; Knowles, Persico and Todd, 2001).<sup>81</sup> According to this research, discrimination might occur because

---

<sup>81</sup>Taste-based, or preference-based, discrimination occurs when individuals treat the members of some groups worse than others because they are unwilling to pay the psychic costs of treating everyone the

individuals who possess implicit or even explicit biases against racial groups enter into policing (Engel and Swartz, 2013; Smith, 2015*b*; James, Vila and Daratha, 2013), or because bad institutional environments and poor training might instill or reinforce racially biased views (Wilson, 1978; Shusta et al., 2002; Sim, Correll and Sadler, 2013; Smith and Alpert, 2007). To remedy these biases, we might develop effective implicit bias training or more racially tolerant organizational cultures. Many policy reports, academic publications, and police training modules argue that doing this can ‘fix’ police personnel (Smith, 2015*b*; Spencer, Charbonneau and Glaser, 2016; Da Silva, 2018; James, 2017). In sum, this view holds that to reduce racial discrimination in policing, we need to increase the supply of good agents — unbiased officers. This logic is encapsulated in the *Officer Supply Hypothesis*.

**Officer Supply Hypothesis:** The greater the supply of unbiased law enforcement officers, the less policing practices will be discriminatory.

Practically speaking, increasing the supply of unbiased police officers is difficult, though, as the effectiveness of implicit bias training and cultural reforms remains contested. In a wide range of areas, such as in workplaces and in policing, there is little consistent evidence that we can reduce implicit bias (Vedantam, 2008; Mak, 2018; Peruche and Plant, 2006; Bregman, 2012; Lebrecht et al., 2009). It remains an open question, though, why this sort of training works only some of the time — it might be because the interventions are not well-designed, because biases cannot be reduced, or because they can only be reduced in some contexts.

---

same. Statistical discrimination, on the other hand, occurs when individuals use information about the average member of a group to make decisions about individual members (Arrow, 1972). Statistical discrimination does not necessarily reflect biased mental processes, as individuals might have reasons to use group-level attributes to infer individual-level qualities.

### 8.2.2. Demand for Unbiased Policing

One reason why the effectiveness of implicit bias training programs for police might change across contexts is because public preferences for police discrimination do so as well. Most studies state or assume that racial discrimination is evidence of agency loss. However, observing racial discrimination on the part of the police does not on its own indicate the existence or extent of any agency loss. To evaluate the extent of agency loss, must know something about the agent's actions *relative to the preferences of the principal*.<sup>82</sup> It may be the case that there is no agency loss and that the police engage in racial discrimination in part as an attempt to be responsive to the public's preferences, potentially in anticipation of future sanctions (Mansbridge, 2003). In most circumstances, we think that congruence and responsiveness on the part of state agents with respect to public preferences is a good thing (Pitkin, 1967; Dahl, 1989), even if they are unelected (Rehfeld, 2006; Näsström, 2015; Kuyper, 2016). It is important to remember, though, that congruence and responsiveness might not always lead to socially or normatively beneficial outcomes. Many would argue that it is bad if agents are responsive to public preferences for racial discrimination. (Dryzek, 1996; Saward, 2008; Pateman, 1970). The point here is not to engage in a debate about whether responsiveness in this context is good or bad, but rather to highlight that police responsiveness to public preferences may explain variation in racial discrimination across contexts. This suggests that we need to consider how the preferences of the police and the public interact.

The public have heterogeneous preferences over the punitiveness of policing and we know that these preferences have racial overtones (Enns, 2014, 2016; Dovidio and Gaertner, 2004). It seems reasonable to think that the public also has specific preferences for the racial fairness of policing. However this set of preferences is considered, I refer to them

---

<sup>82</sup>Appendix B uses a simple principal agent model to provide additional intuition for this point.

as *public demand for racial equality*. Since we know that punitiveness (Enns, 2016) and racial animus (Stephens-Davidowitz, 2014; Hadden, 2001) vary across contexts, we might think that demands for racial equality do as well. For example, Stephens-Davidowitz (2014) shows that racist views appear to be fairly common in places like West Virginia, rural Illinois, and southern Oklahoma, but less common in areas like Hawaii, sections of California, and Colorado's Front Range. It follows that the people in these places likely have different ideas about how the police should interact with and punish members of minority groups.

We can imagine that several factors feed into public demand for racial equality. Some individuals, for instance, might believe that certain racial groups deserve better treatment than others (Becker, 2010). Others might come to associate racial groups with criminal activity and thus endorse or accept (as necessary) higher rates of police violence (Hjorth, 2017). While racial biases might explain quite a bit of the variation in public demand, other concerns might matter as well (Hehman, Flake and Calanchini, 2017). For example, even when individuals are not biased against racial groups, we can imagine that they might not care much about stopping discriminatory policing. This could be, for example, because they might believe in a tradeoff between security and racial equality (Dietrich and Crabtree, 2018). Indeed, in high-crime areas, citizens are likely to be less concerned about abstract notions like 'fairness' and 'equity' than about the police using whatever means necessary to ensure safety (Chevigny and Chevigny, 1995).

We know that public preferences influence the behavior of state actors such as judges (Canes-Wrone, Clark and Kelly, 2014; Bright and Keenan, 1995; Hall, 1995). It is reasonable to think that public opinion might also influence the police. Like other street-level bureaucrats, law enforcement officers labor under heavy workloads (Jauregui, 2016; Knight, 1990). They are asked to maintain order, enforce the law, and serve the community (Wil-

son, 1989). In an average day, an officer might need to break up a fight, apprehend a robber, and provide first-aid or directions to local citizens (Monkkonen, 1981). To manage their many diverse responsibilities, officers use their broad discretionary powers to decide where and how they should work (Muir, 1977). They often focus on tasks and otherwise conduct their duties in a way that is most likely to maximize the satisfaction of their supervisors (Lipsky, 2010), who have preferences over policing practices, punitiveness, and other outcomes (Brehm and Gates, 1999; Monkkonen, 1981; Enns, 2016). Since these supervisors are typically directly or indirectly accountable to the public, the police can be expected to act in line with public opinion. In a similar vein, the police might also want to satisfy the elected officials who oversee them. In addition to these reasons, the police, like other street-level bureaucrats, also try to minimize their effort and any public dissatisfaction that could make their life at work less costly (Lipsky, 2010; Wilson, 1978).

Law enforcement personnel have many opportunities to assess public satisfaction and to update their actions in line with community preferences. One way they can do this is through patrolling communities and talking with citizens, a core component of law enforcement duties in many localities (Reiss, 1973).<sup>83</sup> Another way that they can monitor public opinion is through Community-Police Advisory Boards, which typically exist to increase communication between law enforcement and the areas that they serve (Chevigny and Chevigny, 1995). Other avenues through which police can gauge public attitudes include local news coverage, social media interactions, and official complaints.

One potential issue, however, is that police are unlikely to understand or perhaps care about the preferences of the median citizen in their jurisdiction. While police come into contact with many individuals, the non-criminal sample that they interact with is skewed

---

<sup>83</sup>One law enforcement administrator who completed my survey experiment offered a jaundiced view of that process: “The older citizens tend to remind you that they “know” the law, they “know” your boss, and they tend to tell you what your job is.”

— more given to needing help, desiring punishment, and prioritizing safety over other goals (Muir, 1977). As a result, the police might have biased information over public preferences regarding racial equality in policing practices. In addition, police administrators may be responsive to those who have political clout. Individuals from these groups might value protection over racial equality in policing practices. This means that the police might not tailor their work to satisfy the preferences of the median individual in their jurisdiction but rather the median individual with whom they interact or the median individual in their preferred group of citizens.

The importance of public demand on racial equality in policing outcomes is captured in my *Public Demand Hypothesis*.

**Public Demand Hypothesis:** The greater the public demand for racial equality, the less discriminatory policing practices will be.

Theoretically, the supply of unbiased officers and public demand for racial equality should interact to determine the level of police discrimination. When the supply and demand are low, we should observe high levels of discrimination. This is because ‘bad’ police officers are inclined to discriminate and they do not face public pressure to apply the rule of law equally across races. Similarly, when supply and demand are high, we should observe low levels of discrimination. This is because ‘good’ police prioritize racial equality in their citizen interactions and the public’s egalitarian views reinforce their views. In all other cases, we should observe moderate levels of discrimination. When supply is high but demand is low, unbiased police officers face public demand for racial inequality and may compromise their views to minimize dissatisfaction. And when supply is low but demand is high, an egalitarian public pressures biased officers to engage in less discrimination.

### 8.2.3. Police Responsiveness

As discussed above, the police should care about the public's preferences. The extent to which public demand influences police officers should depend on the degree to which they can be held accountable to the public. The extent to which the public can hold the police accountable depends on two things. One, the public needs to know if the police are responsible for the good or bad things that they observe. This can be difficult since there is often little independent evidence about policing actions to identify who is responsible.<sup>84</sup> When the police kill someone, they can simply say that the victim carried a gun or threatened them, and the 'blue wall of silence' dissuades other officers from exposing false accounts of citizen interactions. Two, the public needs to be able to influence the police. This is difficult, however, as front-line police and administrators are typically appointed to office. In many cases then, the public cannot use electoral institutions to directly reward or punish bad police.

In some cases, there is limited accountability in practice. When this is the case, the police might ignore public preferences entirely if they can get away with it, acting instead based on their own preferences (Skolnick and Fyfe, 1993). While little systematic data exists on the preferences among police for racial equality, we can imagine that the police would provide preferential treatment to members of their own racial group. This would be in line with the large literature on implicit biases (Dovidio, Kawakami and Gaertner, 2002; Green et al., 2007), and is also the basis for many calls to diversify police personnel (Kennedy et al., 2017*b*).

In other cases, though, the police can be held directly or indirectly accountable by the public (Chevigny and Chevigny, 1995). The public, for example, can influence police behavior through local oversight bodies, such as Los Angeles' Civilian Oversight Commission

---

<sup>84</sup>This issue is one of the reasons why some advocate for the police to wear body-worn cameras.

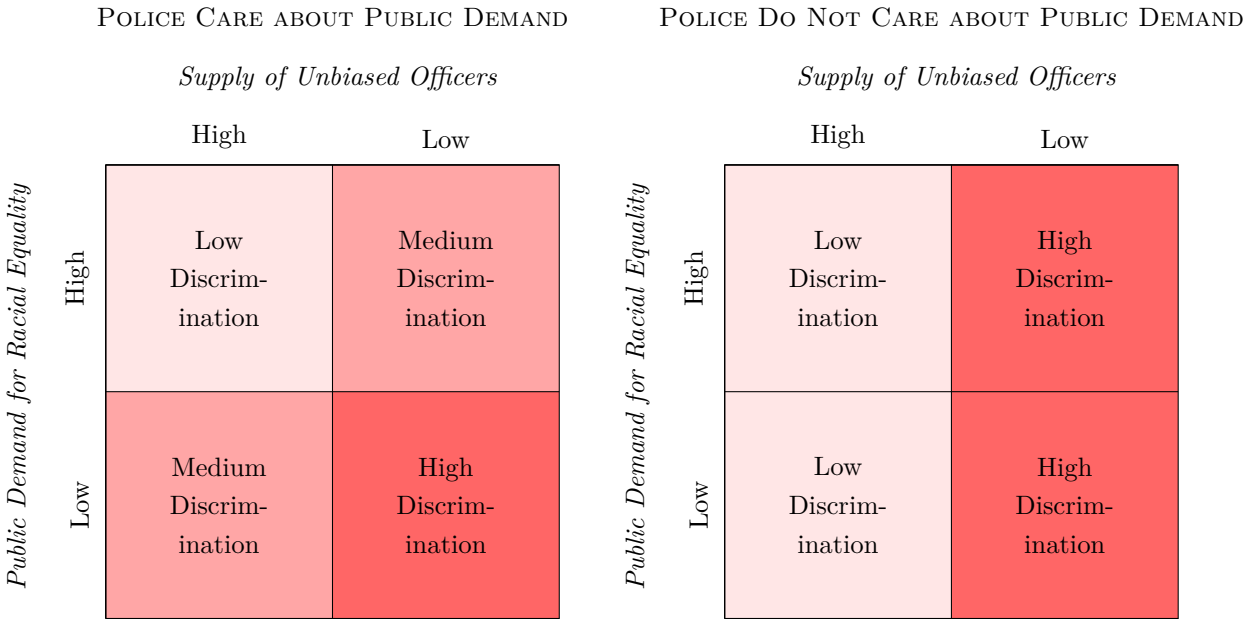
(Chan, 1999). The public can also hold law enforcement accountable through elections. While the public typically do not vote for front-line police, they often elect law enforcement administrators who can indirectly exert pressure on street-level officers to act in accordance with public wishes. They also elect officials such as mayors and state legislators whose job it is to oversee the police. We know that citizens call or write to mayors and city managers about their concerns with policing practices and that they sometimes elect representatives because of their criminal justice policies (Eckhouse, 2016; Kennedy et al., 2017*a*; Surette, 1985). It is reasonable that they do the same with elected law enforcement administrators, and that those important intermediaries between the public and police care about citizen preferences regarding racial equality. This would be in line with the judicial politics literature, which finds that public opinion influences how elected judges decide cases. For example, elected judges are more likely than unelected judges to act in line with public preferences about punitiveness, particularly as elections approach (Streb, 2007; Brace and Boyea, 2008; Berry, 2015). In other words, the effect of public preferences only matter to the extent that police have an incentive to be responsive.

In comparison to earlier work, my framework suggests that discriminatory policing depends on the interaction between (1) unbiased police officers, (2) public egalitarianism, and (3) police accountability. Figure 8.1 graphically represents the full theory. The table on the left captures a world in which the police and the people who oversee them care about public preferences. The table on the right captures a world in which the police and people who oversee them do not care about public opinion. In each of these worlds, the level of police discrimination depends on the supply of unbiased police officers and the demand for racial equality in policing. The cells denote the predicted level of discriminatory policing with darker colors indicating more discrimination.

As Figure 8.1 illustrates, the effect of public demand on discriminatory policing varies



**Figure 8.1: Theoretical Expectations about Discriminatory Policing - Concern about Public Demand**



**Note:** Cell entries denote the predicted level of discriminatory policing.

depending on whether the police care about it. I have already described the predictions for the world in which police care about public preferences. In the world where police officers do not care about public preferences (the right table), the level of police discrimination is driven entirely by the supply of unbiased police officers. When the supply of unbiased police officers is high, discrimination is low, and when the supply of unbiased police officers is low, discrimination is high. This is the world implicitly assumed in existing studies of policing. This framework leads to two conditional hypotheses (Brambor, Clark and Golder, 2006a; Berry, Golder and Milton, 2012), the *Conditional Officer Supply Hypothesis* and the *Conditional Public Demand Hypothesis*.

**Conditional Officer Supply Hypothesis:** The greater the supply of unbiased law enforcement officers, the less policing practices will be discriminatory. This negative effect increases with public demand for racial equality, particularly when the police care about citizen preferences.

**Conditional Public Demand Hypothesis:** The greater the public demand for racial equality, the less policing practices will be discriminatory. This negative effect increases with the supply of unbiased officers, particularly when the police care about citizen preferences.

### 8.3. Empirics

Most analyses of racial discrimination ignore the demand for racial discrimination. That is, existing research focuses on the supply of unbiased police officers and fails to recognize the interaction between supply and demand side factors that influence racial discrimination by the police. My theory suggests that this focus on the additive effect of unbiased policing might explain some of the conflicting results about police discrimination in the literature. My theory also suggests that researchers should examine how the supply of unbiased police officers interacts with public preferences for racial egalitarianism across different accountability structures to affect police discrimination.

It is difficult to collect data on the public's preferences for racial discrimination by the police and police incentives to care about these public preferences. In what follows, I focus on evaluating how public opinion affects the level of racial discrimination exhibited by agents involved in policing, and the extent to which this effect is stronger when those agents have greater incentives to care about the preferences of the public. To test my hypotheses, I need variation in (1) racial discrimination by agents involved in policing, (2) public demand for racial egalitarianism on the part of the police, and (3) incentive structures that encourage agents to be congruent with public opinion. To address these data needs, I design and conduct two novel survey experiments that examine potential discrimination in the use of police power in the American context. These experiments help provide variation in (1) racial discrimination by agents involved in policing and (2) public demand for racial egalitarianism from the police. They focus on the use of excessive force and traffic stops by the police. I conducted the survey experiments in June and July 2018.

While my theory is general and should apply across many contexts, I test it in the United States for two reasons. First, America provides a useful laboratory to think about the principal-agent relationship in policing. Unlike the United States, many countries have a unified policing system or a unitary system of government. This means that the policing units share the same principal — the national public and often the same intermediary principals, national legislators. In America, though, each local policing unit is responsible primarily to the citizens in their own jurisdiction and are overseen by a heterogeneous set of local- and state-level officials.<sup>85</sup> This means that there are thousands of different ultimate principals across the United States and hundreds of different intermediaries, all facing differential demands for public equality in policing. In other words, the United

---

<sup>85</sup>There is some overlap in these communities, as the jurisdictions served by municipal law enforcement agencies are often one part of the jurisdictions served by county law enforcement agencies. No two local law enforcement units, though, share completely overlapping jurisdictions.

States provides the much needed variation in public preferences for racial egalitarianism needed to test my hypotheses. Second, the American case rests at the heart of the growing comparative policing literature. Roughly half of all the articles published on policing in the last 40 years have focused in some way on the United States.<sup>86</sup> As a result, my empirical analysis speaks directly to a large portion of the literature.

### 8.3.1. Sample

While the traditional focus of the policing literature is on the attitudes and behaviors of front-line personnel, the ultimate agents who act on behalf of the public, I focus instead on two alternative sets of actors that operate in the delegation chain linking police officers to the public.<sup>87</sup> The first set of actors are law enforcement administrators. Law enforcement administrators are important for two reasons. One is that they influence front-line police through both the orders they give and the organizational culture they help create. This is particularly true in federal systems, such as the U.S., where local-level administrators possess tremendous discretion within their jurisdictions (Wilson, 1989). Another reason is that they — not rank-and-file officers — are the ones who might be held politically accountable if local police services fail to satisfy members of the community. Importantly for my theory, the administrators in my sample differ in the extent to which they should care about public demand for racial equality. Some have been appointed (i.e. municipal administrators and some sheriffs), while others have been elected (i.e. most sheriffs). This is important for testing my *Conditional Officer Supply Hypothesis* and my *Conditional Public Demand Hypothesis*. The second set of actors are elected officials. Elected officials are

---

<sup>86</sup>A quantitative review of the political science literature shows that the United States appears in 787 of the 1,492 articles (52.75 percent) published on policing from 1980-2018. The next two countries that appear the most frequently in this literature are Germany (572) and China (480). Appendix A contains more details about this review.

<sup>87</sup>Appendix C presents a stylized version of the delegation chain that links the public to front-line officers.

important for two reasons. One is that while there is some variation in the elected nature of law enforcement administrators, this is in practice quite minimal; the vast majority of law enforcement administrators are appointed. Including elected officials, thus, provides me important variation in the extent to which the actors in my sample care about public preferences with respect to policing outcomes. Given their elected status, we should see that elected officials are much more responsive to public preferences than law enforcement administrators. The second reason is that there is an increasing demand for police reform and a growing recognition that this might only come through legislation. This means that studying how elected officials respond to the public's preferences for racial egalitarianism is potentially important for understanding the policymaking process that leads to police reforms.

My law enforcement sample comprises the 11,251 administrators at the municipal and county levels who possess valid personal email addresses listed in a popular directory service for public safety professionals, Safety Source's *National Directory of Law Enforcement Officers*. This sample comprises approximately 73 percent of American law enforcement administrators. My elected officials sample comprises the 5,852 state legislators and 1,281 mayors who have valid personal email addresses listed on openstates.org and provided by the United States Conference of Mayors. This sample comprises about 79 and 91 percent of American state legislators and mayors (of cities larger than 40,000 people). I invited law enforcement administrators and elected officials to participate in my research by sending them an email.<sup>88</sup>

The fact that law enforcement issues in the United States have become increasingly contentious likely means that law enforcement administrators and elected officials are reluctant to participate in and complete surveys about these issues and that response rates

---

<sup>88</sup>Appendix D contains the email I sent to each sample.

are likely to be low. To ameliorate this situation, I did several things to maximize my response rate. First, I pre-tested the language I used in my email invitation and found that it was consistently viewed as respectful, polite, and warm in tone.<sup>89</sup> Second, I offered a small financial incentive for recipients to complete the survey. Ethical and legal issues prevent me from directly paying respondents for their participation, so I instead pledged to donate \$1 for every completed survey response to the National Law Enforcement Officers Memorial Fund (NLEOMF). Butler and Pereira (2018) find that the promise of charitable donations can encourage political elites to perform better on surveys. Since performance is related to work effort, I expect that charitable donations will encourage survey completion. I selected this police-related charity because, unlike others, it engages in only limited political advocacy.<sup>90</sup> This is important because I do not want administrators and officials to decide whether to participate in the survey based on whether they support the political goals of the charity. Third, I re-invited individuals in my samples who did not open the survey invitation at three different points. I used Mailchimp to deliver the emails and to track who opens my emails and the survey links that they contain (Butler and Crabtree, 2017b).<sup>91</sup> Fourth, I purposefully constrained the length of my survey to about 10 minutes.<sup>92</sup>

Despite my best efforts, I received only 349 completed responses to my survey from law enforcement administrators and 237 responses from elected officials, for a response rate of approximately 3 percent per sample. According to MailChimp, only about 33 percent

---

<sup>89</sup>Appendix E contains additional details about this pre-test.

<sup>90</sup>One indication of this is that the organization has not taken at least 5 clear for or against positions on bills that themselves have experienced interest-group position-taking five times (Crosson and Lorenz, 2018).

<sup>91</sup>Appendix F presents the implementation timeline for each survey.

<sup>92</sup>Both survey experiments were embedded in longer surveys conducted by the Annual Law Enforcement Survey. The ALES conducts annual surveys to collect data on the sociodemographic characteristics of law enforcement administrators and elected officials, as well as information about their views on police policies, problems, and future priorities.

of each sample opened at least one of my emails. This means that the response rate conditional on an email being opened is approximately 9 percent. At first glance, this reply rate might seem low. There are several reasons why we might expect a low response rate. The most important has to do with the particular focus of my survey. As I suggested earlier, there is some evidence to support the view that law enforcement personnel and elected officials are hesitant to share their views on policing practices. For instance, I received many emails from law enforcement administrators, state legislators, and mayors indicating that they would prefer not to complete my survey because policing is a hot-button topic. Some email senders went so far as to indicate that they worried their replies could somehow be tracked and later used against them, even though the survey notes in several places that all answers would be confidential. This is likely a reaction to the public outcry over police violence that followed in the wake of the 2014 death of Michael Brown and the national emergence of the Black Lives Matter movement in 2015. It might also be a reaction to the public's increasing use of social media as a means of 'naming and shaming' law enforcement officers (Hafner-Burton, 2008; Salter, 2016). Administrators might worry that their responses could be used against them by being posted online without permission.

Even though relatively few administrators and officials completed my survey, we should care about the attitudes and actions of those who did. The 292 appointed municipal law enforcement administrators and 57 elected county sheriffs who completed my survey represent an important subset of all criminal justice personnel. Collectively, they serve 6,208,017 Americans across 46 different states and command 9,549 front-line, sworn officers.<sup>93</sup> National statistics suggest that every year one in five Americans interacts with a law enforcement officer (Eith and Durose, 2011). If this statistic travels to the contexts policed by the administrators in my sample, then these individuals make decisions that

---

<sup>93</sup>Appendix G shows a map of the law enforcement administrator respondents.

inform the lives of over 1,200,000 residents annually. While Institutional Review Board restrictions prevented me from collecting detailed data on the elected official respondents, there is reason to believe that they also represent a politically consequential sample. For example, the 57 mayors who completed the survey together oversee cities that contain, at a minimum, more than 2.28 million individuals across 27 states. The 176 state legislators in my sample serve populations across 47 states.<sup>94</sup>

### 8.3.2. Experimental Research Design

My samples of law enforcement officers and elected officials completed the same basic survey experiment, which has a fully crossed 2 by 3 factorial design that mirrors the conditional theoretical framework shown in Figure 8.1. The experiment is comprised of three parts. First, respondents are asked a series of basic demographic questions. The answers to these questions provide me with important pre-treatment covariates and, for law enforcement administrators, allow me to identify whether administrators have been elected or appointed. This data is necessary to test my *Conditional Officer Supply Hypothesis* and my *Conditional Public Demand Hypothesis*.

Second, I randomly provide respondents with polling information about public attitudes in their jurisdiction towards policing outcomes. All respondents are told that citizens in their jurisdiction are concerned about lowering crime and making police work less dangerous. Pew surveys indicate that these are the public's primary concerns about policing (Ekins, 2016). Some respondents are provided no additional polling information and some are told that their citizens also care about "increasing racial equality in policing."<sup>95</sup> I mix information about public attitudes toward racial equality in policing in with polling

---

<sup>94</sup>Appendices H and I contain additional details about the respondents.

<sup>95</sup>Some law enforcement administrators were also told "that citizens in their jurisdiction would like to increase harsher prison sentences." I examine the effect of the punitive language in a different paper and average over this treatment and the no information treatment in my empirical analyses.



data about other police outcomes with the goal of minimizing the probability that respondents would be primed to think about race or that they would infer the purpose of my study. While there are good reasons for mixing the racial equality information in with other polling data, this does increase the possibility that the respondents will miss this treatment.<sup>96</sup> Here is the exact wording for the different EQUALITY treatments.<sup>97</sup> Items (a) and (b) are randomly assigned.

*...polling data indicate that citizens in your jurisdiction are concerned about:*

- *Lowering the crime rate*
- *(a) Increasing racial equality in policing, (b) No text*
- *Making police work less dangerous*

Third, after providing information about public attitudes towards policing in their jurisdiction, I ask respondents to evaluate two vignettes that describe an interaction between an individual who has violated (or been thought to violate) a law and police officers. These vignettes can be thought of as two separate experiments. To maximize the ecological validity of the experiment, I modeled each vignette on ones used in publicly available police training materials (Shadish, Cook and Campbell, 2001). One vignette focuses on police use of force and the other involves a traffic stop. The excessive use of force by police has received increased scrutiny in American media after a series of police killings and beatings that occurred in 2013 spurred nationwide protests. Racial discrimination with respect to

---

<sup>96</sup>For my informational treatment to work, respondents need to notice and believe it. The racial equality treatment is only several words long and respondents might easily overlook it or quickly forget it while reading the rest of the paragraph (160 words). Prior to implementation, I pre-tested this treatment with a sample of 400 Amazon Mechanical Turkers (MTurkers). The results of this analysis, presented in Appendix J, suggest that the treatment was noticed and remembered. It is important to consider that if respondents do not believe that the public in their jurisdiction possesses egalitarian attitudes, or if they have strong prior beliefs about these attitudes, I will not find the effects predicted by my theory.

<sup>97</sup>Appendix K contains the full description seen by respondents.

traffic stops is beginning to attract more public attention as a growing number of individuals record their traffic encounters with police and post these online. Anecdotal and empirical evidence suggest that racial discrimination often occurs in both of these policing contexts (Zimring, 2017; Ross, 2015; Epp, Maynard-Moody and Haider-Markel, 2014; Baumgartner, Christiani, Epp, Roach and Shoub, 2017).

To measure discrimination among the respondents, I randomize the race of the individual with whom the police officers interact in each vignette. This treatment factor has three levels: (1) BLACK, (2) LATINO, and (3) WHITE. In the use of force vignette, I vary race by stating that the individual is Black, Latino, or White. In the traffic stop vignette, I indicate race by varying whether the individual has a Black-, Latino-, or White-sounding name.<sup>98</sup> I use different strategies for signaling the race of the individual in the vignettes in order to minimize the possibility that respondents figure out the real purpose of each vignette. The three race treatments in each vignette combine with the two EQUALITY treatments to create six treatment combinations per vignette. All treatment conditions are assigned with equal probability in each vignette.

The vignettes and response questions are presented on the same page. This is to help ensure that respondents can refer back to the vignette while considering their response. The vignettes and questions seen by law enforcement administrators are displayed below. Similar vignettes and questions were presented to elected officials.<sup>99</sup> The questions for elected officials differ only in that they ask about the extent to which they would investigate,

---

<sup>98</sup>Using names to indicate racial identities is common practice in the larger literature on discrimination (Heckman, 1998*b*; Crabtree, 2018; Crabtree and Chykina, 2018; Crabtree, Hou and Liu, 2018), and there has been considerable methodological work on the appropriate use of names for signaling racial identity (Butler and Crabtree, 2017*b*; Gaddis, 2018*b*; Crabtree, 2018; Crabtree and Chykina, 2018). To mitigate possible name effects, each group identity condition is signaled by 10 unique names (Butler and Crabtree, 2017*b*). I pre-tested the extent to which these names cue the intended racial identity and found wide agreement among my human coders. The full results of this construct validity test are available in Appendix L.

<sup>99</sup>Appendix M presents these vignettes.

in the use of force case, or accept, in the traffic stop case, policing actions. Words in brackets were randomly assigned, as was the order in which the vignettes appeared.<sup>100</sup>

**[Use of Force vignette]** *Two police officers on foot patrol surprise a 6'2" <Black / Latino / White> man with a large build who appears to be breaking into an apartment. The man flees, shouting at the officers. The officers chase after him for a while, eventually succeeding in tackling him to the ground. After he is under control, both officers punch him a couple of times as punishment for fleeing and resisting.*

**How likely would you be on a scale of 1 to 10 (where 10 is most likely) to reprimand these officers?**

**[Traffic stop vignette]** *You pull over a silver 2010 Ford Taurus for driving 40mph in a 30mph zone. When you approach the vehicle, you notice that the driver seems very concerned. He identifies himself as <Black / Latino / White name> and explains that he is rushing to pick up his child from school. It's mid-afternoon, the school he mentions is nearby, and the driver is very apologetic.*

**How likely would you be on a scale of 1 to 10 (where 10 is most likely) to issue him a traffic ticket?**

### **8.3.3. Empirical Analysis**

My theory addresses how public egalitarian attitudes change the level of racial discrimination in policing outcomes. Do law enforcement administrators and elected officials treat minorities better if they know that the public wants greater racial equality in policing?

---

<sup>100</sup>Appendix N presents the full survey instruments used for each sample.

To answer this question, I vary public demand for egalitarianism in the second part of my survey, and I identify whether there is racial discrimination among respondents in the third part of my survey. By combining these two steps, we can see if public demand for egalitarianism *changes* the level of racial discrimination. Factorial experiments are designed to test conditional theoretical claims such as these. This means that the results from these experiments can usefully be evaluated with an interaction model (Brambor, Clark and Golder, 2006*b*; Berry, Golder and Milton, 2012).

I use the following specification to test the *Conditional Public Demand Hypothesis* for my law enforcement administrator and elected official samples.

$$\begin{aligned} \text{Beating/Ticketing} = & \beta_0 + \beta_1 \text{Black} + \beta_2 \text{Hispanic} + \beta_3 \text{Equality} + \beta_4 \text{Black} \times \text{Equality} \\ & + \beta_5 \text{Hispanic} \times \text{Equality} + \mathbf{X} + \epsilon \end{aligned} \quad (8.1)$$

My dependent variables in each specification, BEATING or TICKETING, capture responses to the two survey vignettes. Each is measured on a 1 – 10 scale. For law enforcement administrators, greater values originally indicated that the respondent was less likely to reprimand police officers for using excessive force and more likely to issue a traffic ticket. For elected officials, high values originally indicated that the respondent was less likely to investigate police officers for using excessive force and more likely to approve a police officer issuing a ticket.<sup>101</sup> Given the continuous nature of my dependent variables, and in line with my pre-registration plan, I estimate Model (1) using ordinary least squares models. Also in line with my pre-registration plan, I report robust standard errors.<sup>102</sup>

Turning to my independent variables, BLACK and HISPANIC are binary variables that

---

<sup>101</sup>To ease model comparison later, I reverse code the use of force measure so that higher values signify greater acceptance of excessive force. Appendix O presents descriptive statistics for these measures for each sample.

<sup>102</sup>My results do not substantively change here or in later models if I use classic standard errors instead.

equal 1 if the respondent receives the Black or Hispanic citizen treatment in the policing vignette. EQUALITY is a dichotomous indicator that equals 1 if the respondent receives information that the public cares about racial equality in policing outcomes. The interaction terms are included to test the conditionality of my hypotheses.  $\mathbf{X}$  denotes a vector of mean-centered pre-treatment covariates plausibly related to the outcome measures (Lin and Green, 2015). These include administrator or elected official age, education, gender, party identification, and race.<sup>103</sup> Since I randomize my treatments, I know the estimated treatment effects are unbiased. I include these respondent-level covariates, though, to increase the precision of my estimates (Gerber and Green, 2012).<sup>104</sup>

The literature indicates that the police and other actors in the criminal justice system discriminate. If this is true,  $\beta_1$  and  $\beta_2$ ,  $\beta_1 + \beta_4 Equality$  and  $\beta_2 + \beta_5 Equality$  should always be positive. In fact, since the existing literature assumes that police discrimination is not affected by public preferences  $\beta_4$  and  $\beta_5$  should be 0.

In contrast, my theory suggests that the police are less likely to discriminate if they receive the EQUALITY treatment. Thus, I predict that  $\beta_4$  and  $\beta_5$  are negative.<sup>105</sup> One problem for testing my theory, though, is that police and elected officials will have some prior knowledge about public preferences for racial discrimination. As indicated above, this could be because they interact with the public on a daily basis, read the local news, or interact with the public in formal institutional settings, such as citizen oversight boards. If police and elected officials believe that citizens value racial egalitarianism, my theory suggests that we might not observe racial discrimination. On the other hand, if police and elected officials believe that the public do not value racial equality in policing, then we should see racial discrimination. In either of these cases, we would expect that my

---

<sup>103</sup> Appendices H and I contain descriptive statistics for these measures.

<sup>104</sup> I obtain similar results if I omit these covariates from my model.

<sup>105</sup> Whether the negative effect of  $\beta_4$  or  $\beta_5$  is enough to eliminate the positive effect of  $\beta_1$  or  $\beta_2$  is not specified by my theory.

EQUALITY treatment would not influence how law enforcement administrators and elected officials respond to my vignettes.

In addition to these predictions, my theory suggests that the negative effects of  $\beta_4$  and  $\beta_5$  should be larger when the public can hold the police accountable. The model specification shown in Eq. (1) does not take into account the incentives that police face to take into account the public's preferences. I expect the negative modifying effect of public preferences to be stronger, and perhaps only matter, when the police, or other intermediate agents, have an incentive to be congruent with citizen preferences. This means that the effects of  $\beta_4$  and  $\beta_5$  should be higher for elected officials and elected law enforcement administrators than appointed law enforcement administrators. To test this, I essentially estimate Eq. (1) on two different samples — elected agents and unelected agents.<sup>106</sup>

Figure 8.2 presents the results from an OLS model that uses BEATING as the dependent variable (top plot) and another that uses TICKETING as the dependent variable (bottom plot).<sup>107</sup> The vertical axis of each figure denotes the different treatment indicators and combinations and the horizontal axis indicates the effect of these treatments, with higher values indicating greater acceptance of beating or ticketing by the police. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals. Since my hypotheses provide predictions about coefficient signs, I conduct one-tailed tests of significance and declare statistical significance when  $p < 0.05$ .<sup>108</sup>

The results from Figure 8.2 provide partial support for the *Conditional Public Demand Hypothesis*. My theory predicts that the coefficients for HISPANIC and BLACK should be positive or 0. This is because the extent to which law enforcement administrators discriminate against racial minorities is conditional on their prior beliefs about the extent

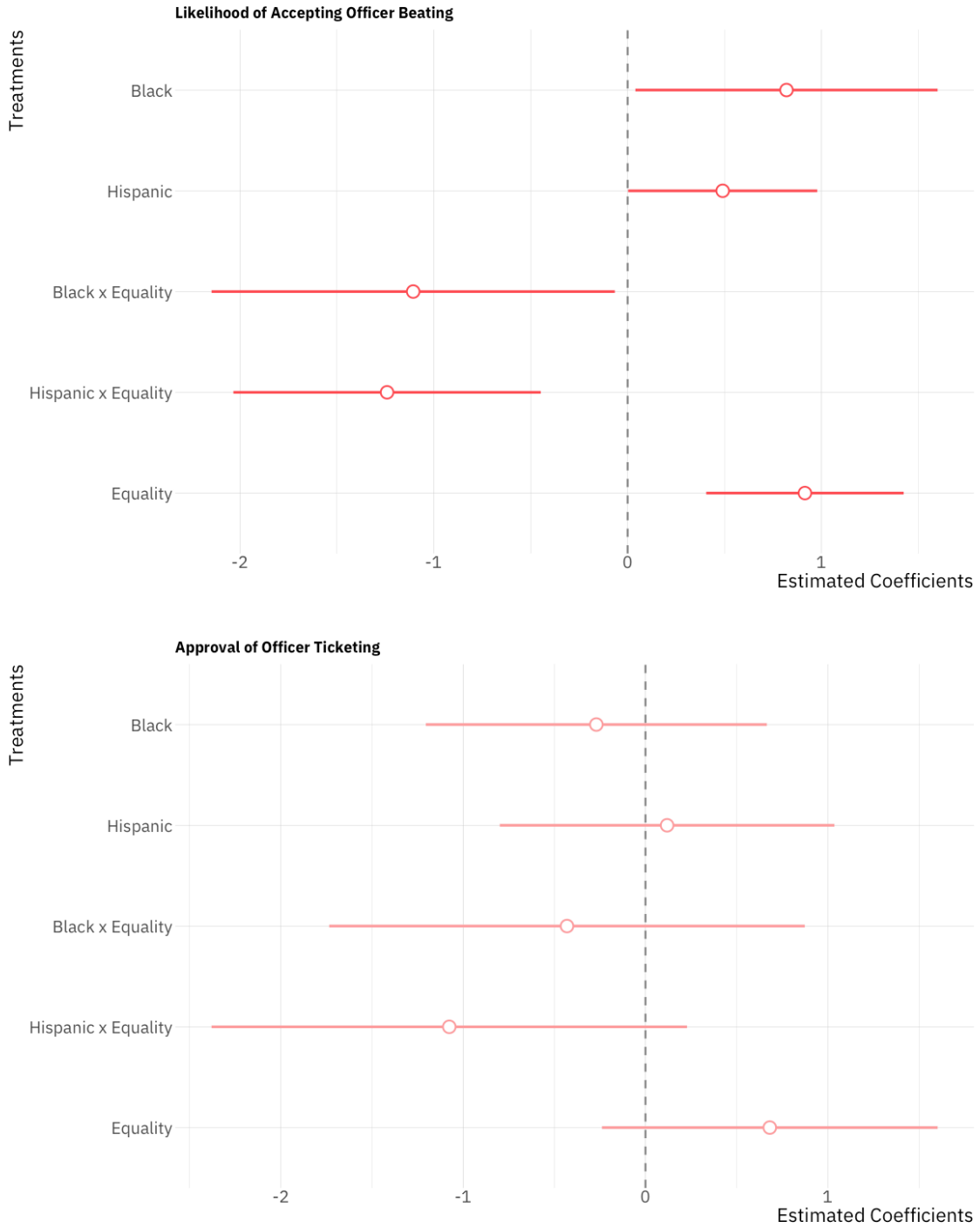
---

<sup>106</sup>In the case of law enforcement administrators, I use a triple interaction model to evaluate whether the results of my models are different for elected and appointment respondents.

<sup>107</sup>Appendix P presents these results in table format.

<sup>108</sup>I pre-registered this decision.

**Figure 8.2: OLS Models with Data from Elected Officials**



*Note:* Figure 8.2 presents the results from an OLS model that uses BEATING as the dependent variable (top plot) and another that uses TICKETING as the dependent variable (bottom plot). The vertical axis of each figure denotes the different treatment indicators and combinations and the horizontal axis indicates the effect of these treatments, with higher values indicating greater acceptance for beating or ticketing by the police. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals. Data come from 57 elected mayors and 176 elected state legislators.

to which the public supports racial equality. The slopes for HISPANIC and BLACK are positive in the model with BEATING as the dependent variable. The coefficient for BLACK is both statistically and substantively significant. It indicates that elected officials are half a standard deviation more likely to tolerate the police using force against Blacks than they are the police using force against Whites. Taken together, these results provide some evidence that elected officials discriminate against Blacks and Hispanics in their capacity of overseeing the police. Importantly, we observe the strongest evidence of discrimination here when police use excessive force, which is problematic normatively as it suggests that the officials charged with overseeing and potentially punishing front-line officers and policing units for using violence against civilians are likely to fulfill these functions in a biased way. Also as predicted, the coefficients on the interaction terms, BLACK  $\times$  EQUALITY and HISPANIC  $\times$  EQUALITY, are negative in both models. They are always negative and statistically significant in the beating case. This indicates that the positive effect of the BLACK and HISPANIC treatments on the police punishment is always smaller and sometimes disappears if the respondent believes that their constituency holds egalitarian preferences. In other words, perceptions about public attitudes toward racial equality in policing outcomes influence the extent of discrimination exhibited by elected officials.

The results from the model with TICKETING as the outcome measure are less supportive of my theory. Importantly, I cannot reject the null hypothesis that the coefficients for the BLACK and HISPANIC terms are different from 0. This means that elected officials do not appear to discriminate against racial minorities in this policing context. In line with my theory, though, the coefficients on the interaction terms, BLACK  $\times$  EQUALITY and HISPANIC  $\times$  EQUALITY, are negative. They are not statistically significant at conventional levels, though.

To help ease the interpretation of these results for elected officials, I plot the marginal



effects of the different racial minority treatments in Figure 8.3. The top row plots the marginal effects of BLACK and HISPANIC on BEATING for respondents who did or did not receive the racial equality treatment. The bottom row plots these marginal effects in relation to TICKETING for the same groups. The vertical axis denotes the marginal effect of the different racial minority treatments, with higher values indicating greater punishment by the police, while the horizontal axis denotes whether respondents were treated with the racial equality message. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals.

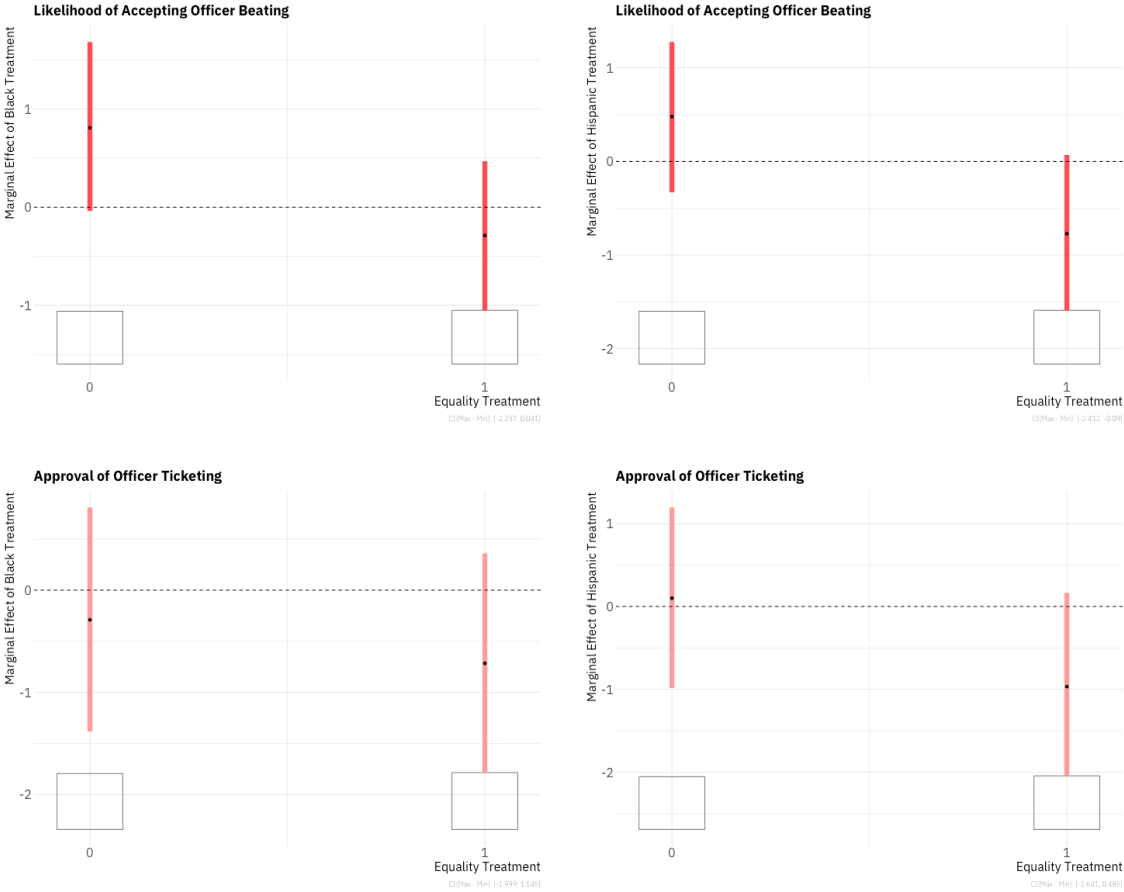
Viewed together, we see that the results for elected officials are largely in line with my theory. In 3 of the 4 plots, the estimated effect of the racial minority cues are positive when the respondent does not receive information about whether the public holds egalitarian views. In line with my theory, we see that effect is negative when elected officials do receive information that the public values racial equality in policing outcomes. These plots also show that the substantive effects of the treatments is large in some cases. This can be seen in the right column of Figure 8.3. Both the top and bottom plot indicate that respondents who received the EQUALITY treatment were approximately half a standard deviation less punitive of Hispanic drivers and suspects than White drivers and suspects. Another way of thinking about this is that elected officials treated with my racial equality message punished Whites at a higher rate compared to Hispanics.

Next I test the predictions from my hypotheses with data from law enforcement administrators. Figure 8.4 presents the results from my analysis of the law enforcement administrator sample in a series of four plots. The plots in the top row present results from OLS models that use BEATING as the dependent variable, and the plots in the bottom row present results from OLS models that use TICKETING as the dependent variable.<sup>109</sup> The

---

<sup>109</sup>Appendix Q presents these results in table format.

**Figure 8.3: Perceptions of Public Demand for Racial Equality and Discrimination**



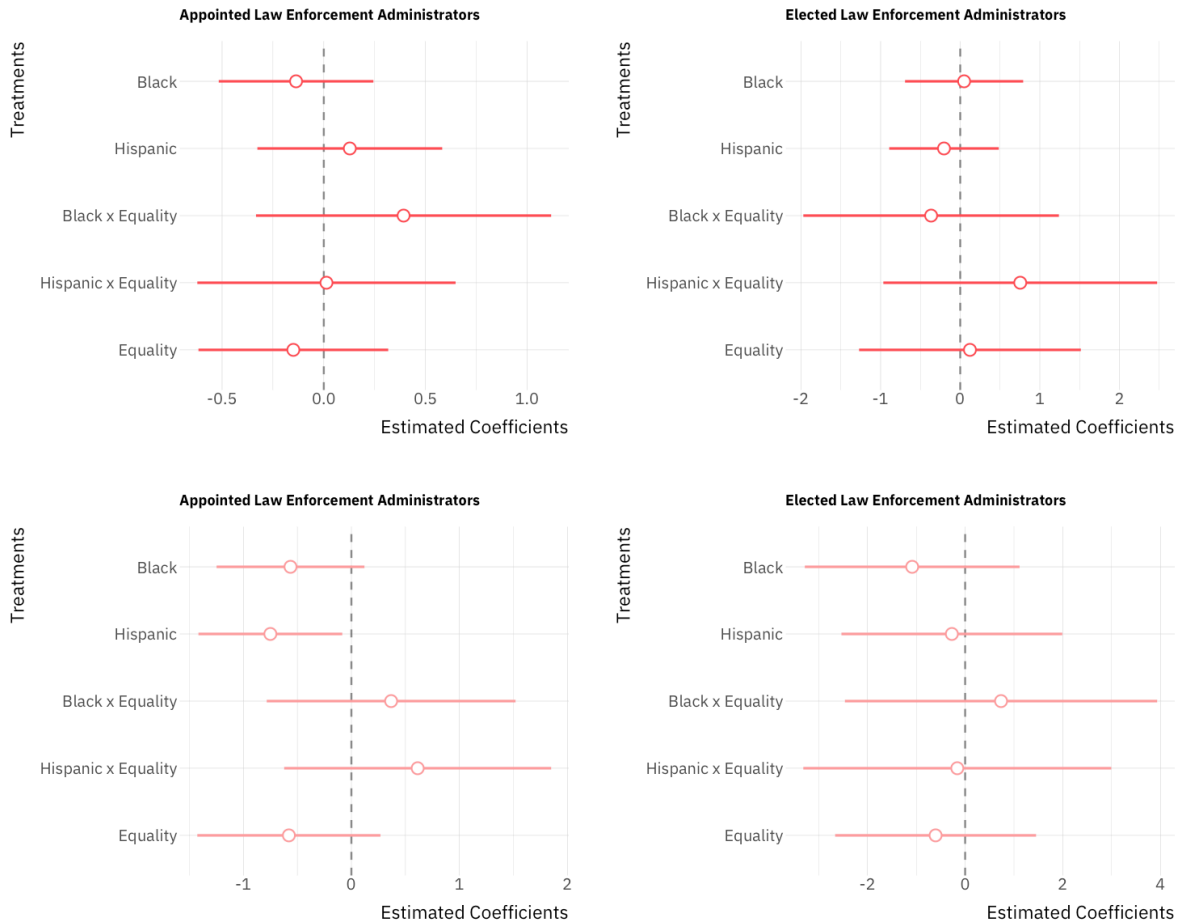
*Note:* Figure 8.3 presents several marginal effects plots created using the models displayed in Figure 8.2. The top row plots the marginal effects of the BLACK and HISPANIC treatments on BEATING, while the bottom row plots the effects of these treatments on TICKETING. The vertical axis denotes the marginal effect of the different racial minority treatments, with higher values indicating greater punishment by the police, while the horizontal axis denotes whether respondents received the racial equality treatment. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals. The bars on the horizontal axis indicate how many respondents were assigned to each level of the EQUALITY factor. Data come from 57 elected mayors and 176 elected state legislators.

plots in the left column present results for models estimated with data from 292 appointed law enforcement administrators, while the plots in the right column present results for models estimated with data from 57 elected law enforcement administrators. The vertical axis in each plot denotes the different treatment combinations and the horizontal axis indicates the effect of these treatments, with higher values indicating greater punishment by the police. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals.

The results in Figure 8.4 do not provide support for the *Conditional Public Demand Hypothesis*. Recall that this hypothesis predicts that law enforcement administrators will discriminate against racial minorities in the absence of information about public demand for racial equality. Contrary to this prediction, the coefficients on BLACK and HISPANIC are negative in 3 of the 4 models. However, these coefficients are not statistically significant, with one exception. This means that I cannot reject the null hypotheses that the coefficients for BLACK and HISPANIC are 0. In other words, there is no evidence that the police ever discriminate in their replies to these vignettes. This is a surprising finding given prior empirical studies on racial discrimination in policing. The coefficients on the interaction terms in these models also indicate that the police do not seem to respond to public preferences, irrespective of whether they are elected or appointed.

There are at least three possible explanations for why there is no evidence that law enforcement administrators discriminate against racial groups in their evaluation of the police behaviors presented in the two vignettes. One explanation is that law enforcement administrators do not have the same biases as front-line personnel. Most of the work on racial discrimination among the police has focused on street-level officers (Persico, 2009; Rafky, 1975; Baumgartner, Christiani, Epp, Roach and Shoub, 2017; Smith, 2015b). In contrast, we know little about the extent of bias among law enforcement administrators.

**Figure 8.4: OLS Models with Data from Law Enforcement Administrators**



*Note:* The plots in the top row present results from OLS models that use BEATING as the dependent variable, and the plots in the bottom row present results from OLS models that use TICKETING as the dependent variable. The plots in the left column present results for models estimated with data from 292 appointed law enforcement administrators, while the plots in the right column present results for models estimated with data from 57 elected law enforcement administrators. The vertical axis in each plot denotes the different treatment combinations and the horizontal axis indicates the effect of these treatments, with higher values indicating greater punishment by the police. Plotted points represent estimated coefficients and solid bars represent two-tailed 95 percent confidence intervals.

Another explanation is that the group of law enforcement administrators who completed the survey are less biased than other groups of law enforcement administrators. A third explanation is that the law enforcement administrators in my study perceived *a priori* that their publics preferred racial equality in policing outcomes. This would explain both why they do not discriminate and why they do not update when they are told that the public wants racial equality. Unfortunately, my experimental design does not allow me to adjudicate among these explanations.

On balance then, I find partial support for my theoretical framework in the data from elected officials but little support for it in the data from law enforcement administrators. A productive area of future work would be to re-examine my theory with a different set of law enforcement administrators. This would allow a determination of whether the null results I find here for that sample travel to other groups of this politically important class of bureaucrats.

## 8.4. Conclusion

In this article, I have situated the police in a political context. Previous studies often examine the police in isolation and ignore the fact that they take actions in a situation surrounded by citizens who have preferences about their attitudes behaviors. In effect, previous studies ignore the political nature of police actions. I have tried to bring politics into this equation by looking at public preferences for racial egalitarianism and the incentives that the police have to respond to these preferences. Departing from prior work, I theorize that racial discrimination depends not only on the extent of egalitarian views among the police but also in a conditional way on the degree of racial egalitarianism among the public and the degree to which police must account for those preferences.

While my theoretical framework is hard to test, I have found partial support for my the-

ory. Specifically, I found that elected politicians exhibit less racial discrimination in law enforcement oversight when informed that the public supports racial equality in policing. Contrary to my theory, though, police do not react to perceived public demand for egalitarianism. Overall, my results suggest that public views about racial equality influence discrimination by the police only indirectly, through the elected institutions that monitor and check their power.

These results have several policy implications. One is that training aimed at reducing implicit and explicit biases among the police is not enough on its own to reduce racial discrimination. If we desire substantial change in police discrimination, we need to change the racial biases held by both the police *and* the public. Without addressing the public's attitudes, racial discrimination is likely to endure since the police and their elected supervisors respond to and are held accountable by public preferences.

This has important implications for future work as it suggests that we need to better understand what the public wants the police to do before evaluating police behavior. We know relatively little, though, about under what circumstances the public condones or supports racial discrimination in policing. One productive area for future research then is to identify the determinants of public attitudes toward racial egalitarianism in policing.

## **Part IV.**

# **Supplementary Materials**

# **Appendix A. Appendix for ‘Persistent Bias Among Local Election Officials’**

## **A.1. Email Scraping**

We collected email and personal contact information from local election officials by programmatically visiting state-maintained sites of local election official contact information. We do not include the following states’ local election officials in our assignment to treatment: Alaska, Hawaii, Maine, Maryland, Missouri, and New Jersey. We exclude Alaska because local election official jurisdictions were not mappable onto census area delineations for covariate data. We exclude Hawaii because a single board member represented each island, and the state did not provide individual email addresses for each island; rather, there was a single catch-all address. We do not include Maine, Missouri, or New Jersey because these states do not make email addresses of local election officials available. We do not include Maryland due to a clerical oversight.

We report other individual officials that were excluded from randomization, as well as reasons for these exclusions in Table A.1. Local election officials were excluded from the



study for concerns related to spillover, or multiple local election officials overseeing a single jurisdiction. All determinations were made prior to randomization. Figure A.1 reports the Consort enrollment and randomization chart for this project.

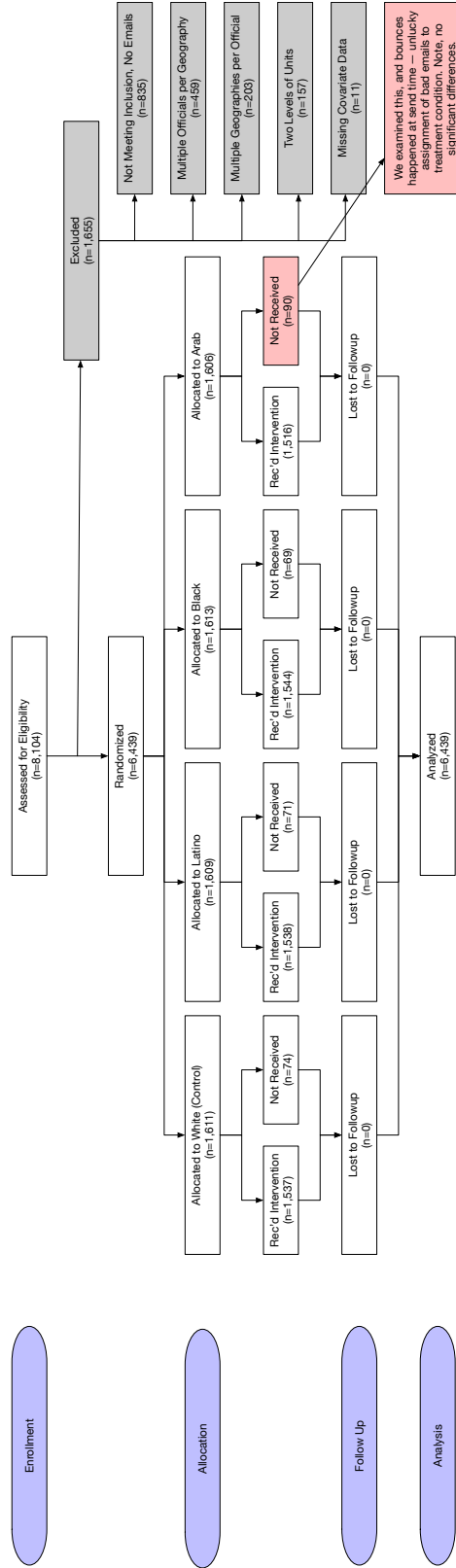
Table A.1: Local Election Officials excluded prior to randomization

Attrition by Study Exclusion Criteria

Exclusion Criteria Category	Exclusion Criteria Details	Number of deleted registrars or units of treatment (n)	Number of subjects remaining in cohort after exclusion (N)
<b>Initial Count</b>	Registrars from whom we collected public information		8104
<b>Two levels of units per state</b>	<b>County and municipality</b>		
	Delete registrars at county level - Wisconsin	(72)	8032
	Delete registrars at county level - Michigan	(83)	7949
	<b>State and county</b>		
	Delete registrars at state level - Delaware	(2)	7947
<b>Missing emails</b>	Delete registrars at county level with no email address - California, Idaho, Indiana, Maine, Missouri, Mississippi, New York, Pennsylvania	(652)	7295
	Delete registrars at municipality level with no email address - Connecticut, Michigan, New Hampshire, Rhode Island, Wisconsin	(183)	7112
<b>Multiple registrars per unit of treatment</b>	Randomly select one registrar per county and delete remaining duplicates:		
	Alabama	(3)	7109
	Arkansas	(19)	7090
	Connecticut	(79)	7011
	Louisiana	(15)	6996
	New Hampshire	(4)	6992
	Keep registrar with name and delete registrar with no name - Nevada	(2)	6990
	Keep registrar with job title "County Director" and delete registrar with job title "Deputy County" - Delaware	(6)	6984
	Keep registrar with job title "City Clerks" and delete registrars with job title "Town Clerks" - Michigan	(68)	6916
	For registrars with no job title, randomly select one and delete remaining duplicates - Michigan	(33)	6883
Randomly select registrar based on ranking of job title (1- "city clerk", 2- "town clerk", 3- "village clerk"), delete remaining duplicates - Wisconsin	(230)	6653	
<b>Spillover - Registrars responsible of multiple units of treatment or registrars sharing email address</b>	Randomly select one county, delete remaining counties for each registrar:		
	Georgia	(155)	6498
	Hawaii	(3)	6495
	Michigan	(31)	6464
	New York	(4)	6460
	South Dakota	(2)	6458
	West Virginia	(1)	6457
Wisconsin	(7)	6450	
<b>Missing data</b>	Unable to assign to treatment due to missing covariate data	(11)	6439
<b>Total</b>		<b>(1665)</b>	<b>6,439</b>

**Figure A.1: CONSORT Document**

CONSORT Flow Document for Main Analysis



## A.2. Email Server Construction

At the design phase of the experiment, informed by the experience of White, Nathan and Faller (2015a) we were concerned about the possibility that local elections officials might become aware of the conduct of our experiment.

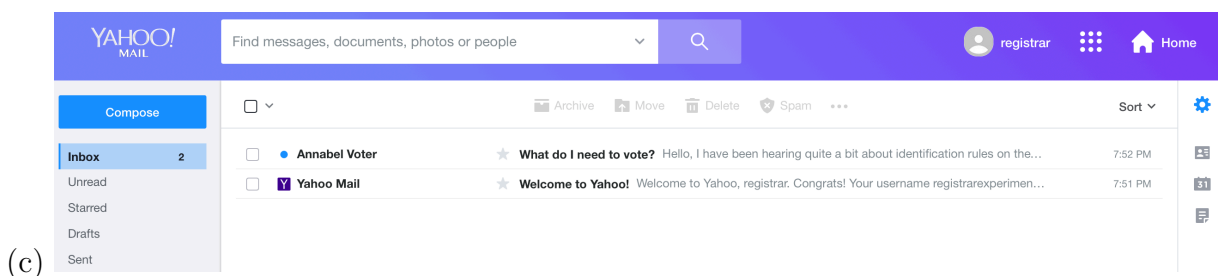
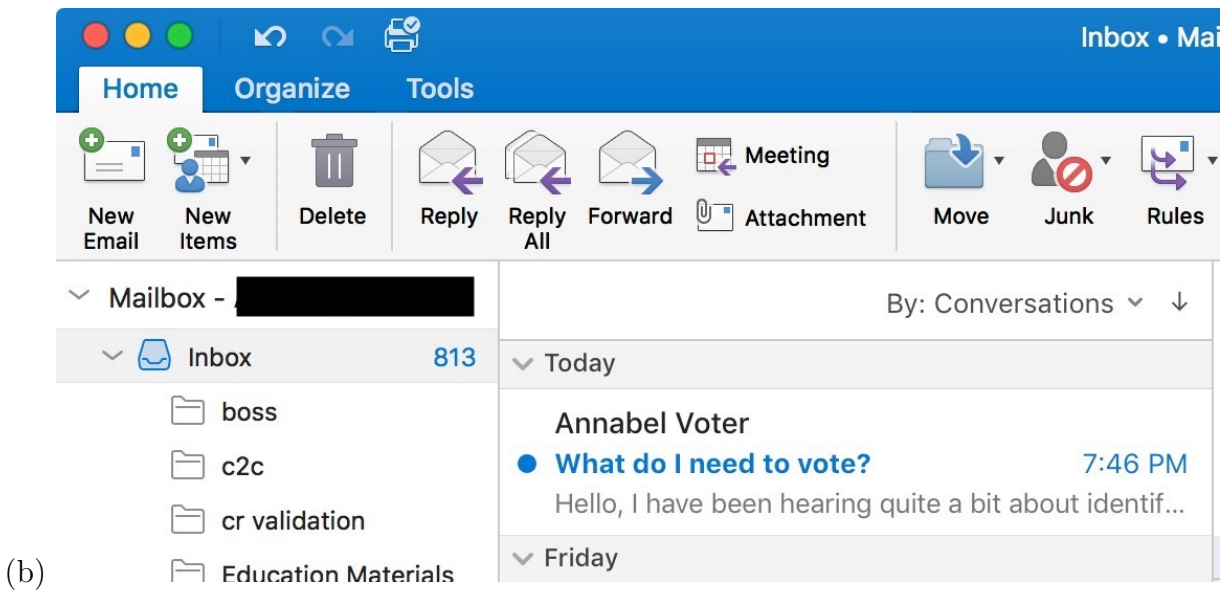
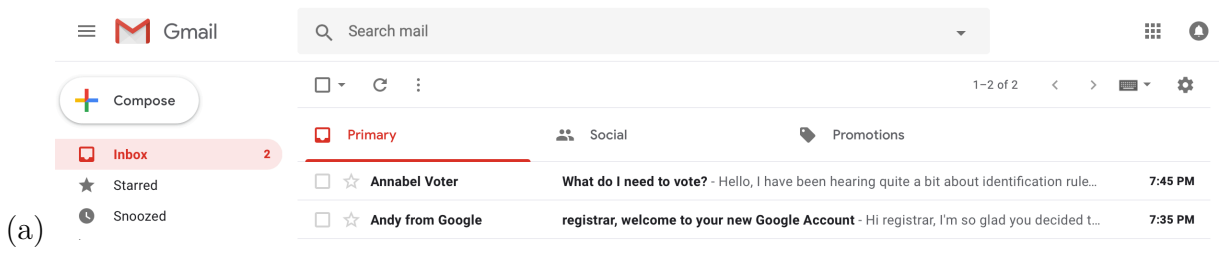
A leading concern was that the domain name `ez-webmail` might structure election officials' responses. However, during the design phase of this experiment, we were surprised to make the observation that most client-side email services do not make the sender domain visible to the user. As we present in Figure A.2, because we engineered our email server to match the `From:` name to the experimental stimulus, local election officials saw the sender name, not the email address in their Inbox. As a result, local election officials using most email programs would most likely not have seen the domain name of our sending server. We note, however that upon opening, all client-side programs make domain information visible to the election official (Figure A.3).

Through the design of this experiment, our research into the front-end and back-end structure of how these emails are processed assuaged many of our concerns about the imperfect delivery of treatment. The following section describes this process.

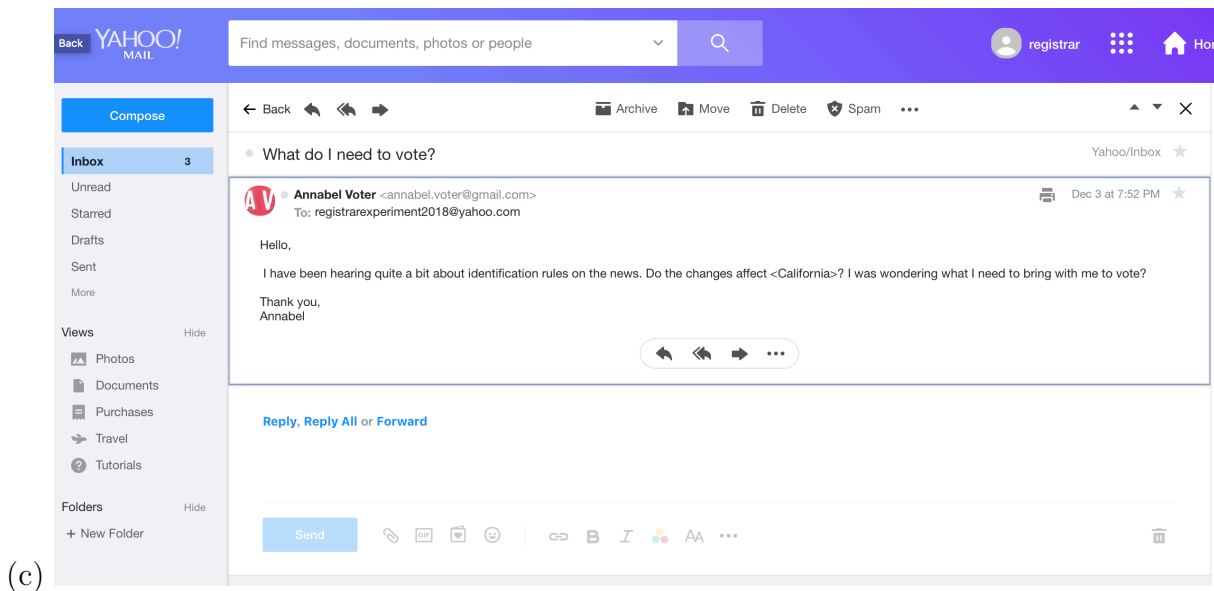
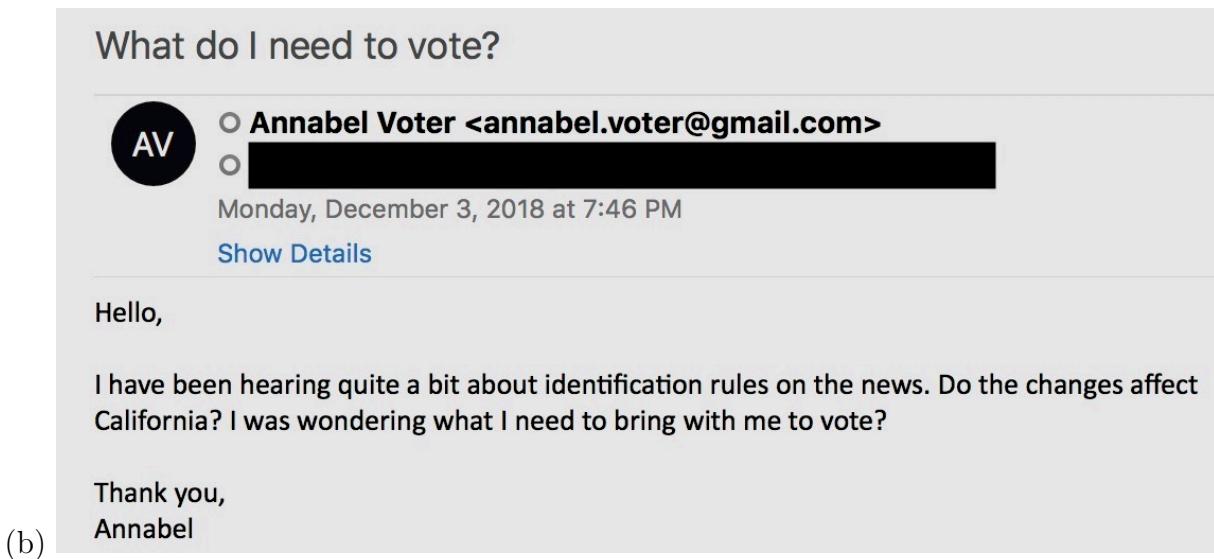
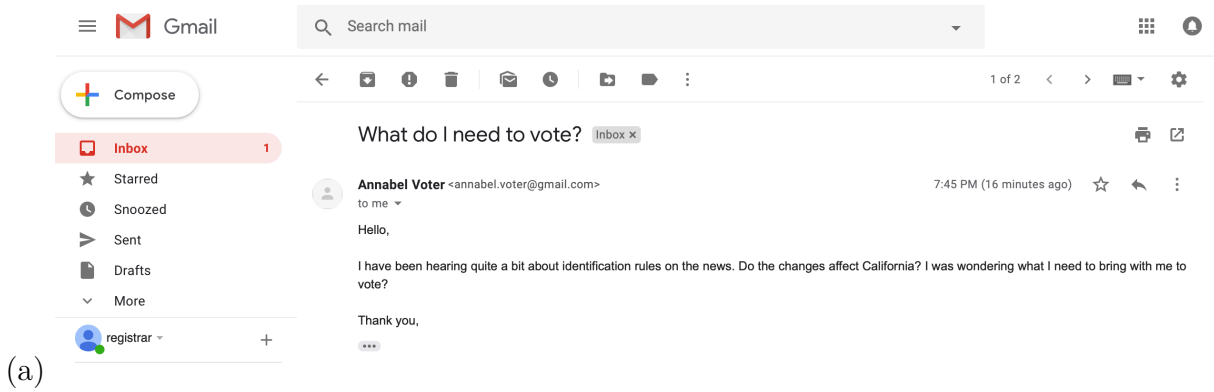
## A.3. Email Back-End Considerations

Our leading concern was that these forms of contact would not reach election officials' inboxes. The primary cause of this failure is being captured by spam filters. To mitigate this concern, we expended significant IT effort to construct an email serving system that would be "well-respected" by client-side (i.e. election official side) email servers.

While the full technical specifications are beyond the scope of this article, we built the email sending server such that it was whitelisted for use on client-side email servers that



**Figure A.2: A likely view of our stimulus in local election officials' email inboxes. Subfigure (a) presents the view in gmail, (b) outlook, and (c) yahoo inboxes.**



**Figure A.3:** A likely view of our stimulus, once opened, in local election officials' email inboxes. Subfigure (a) presents the view in gmail, (b) outlook, and (c) yahoo inboxes.

included Gmail, Outlook, and Yahoo. We confirmed this, before sending, in two ways.

First, we use the *Return Path* sender score to evaluate that we had built sufficient sender score to have a high likelihood of reaching inboxes. Presently the leading indicator for delivering to client inboxes, the Return Path *senderscore* characterizes the reputation, and therefore probability of successful delivery of an email server. The identification of this product, as well as a considerable part of the sending architecture was influenced by interviews we conducted with leadership at a major direct-to-consumer (i.e. email) marketing firm.

Second, we tested that emails were actually arriving at inboxes. Specifically, we sent stimulus emails from our servers to a convenience sample of individuals associated with the research team, in an effort to cover a large part of the client-side landscape. We contacted colleagues, friends, and family using Microsoft Outlook at several different companies, and also contacted several people on each of Gmail and Yahoo email providers. These trials were instructive and serve as a cautionary tale for future researchers: in first rounds of pilot sending – trials where we had relatively low senderscore for our email server – we were not able to deliver any mail to any inbox.

Upon this realization, we took additional steps to improve the reputation of our email server. This involved server certificate signing, as well as ensuring that we had met specific (i.e. DKIM) authorization protocols. Although this is relatively routine for IT professionals, we would like to point out to experimentalists and those considering future audit studies that the process involved considerable work even for individuals with a background in this form of Information Technology.

Despite the cost and challenges of setting up a unique server, we would like to include this piece of dictum: It is our opinion that researchers who are going to engage in future audit studies should undertake the cost. On the one hand, the ability to flexibly define sender

identity, include custom email headers and tracking infrastructure, and design custom data fields permits the estimation of theoretically interesting causal quantities (e.g. open rates). On the other hand, the increased cost of setting up this sending infrastructure serves to rebalance the costs bourn by experimenters and their audit/correspondence study subjects.

## A.4. Mailer Content

Unlike White, Nathan and Faller (2015a), we did not vary whether the local election official receives a request directly related to voter identification. Because previous results establish that prejudicial behavior occurred almost exclusively in response to emails related to voter identification, we focus only on requests of that type.

To minimize the chance that local elections officials would become aware of the study, we took care to develop many versions of email language. In particular, all content that we mailed was a variant of a simple, three sentence paragraph that took the form: (1) Preamble; (2) Question One; (3) Question two.

By asking the same question in multiple ways, we achieve greater certainty that the resulting behavior is a response to the main causal variable of interest, the race of the putative voter, rather than any idiosyncratic feature of our request. Table A.2 presents the different values for the preamble and the two questions. These elements were combined at random, to produce 27 variants of the message text delivered to local officials. These variants were scored by 171 humans for “*clarity*”, “*warmth*” and “*appropriateness*”. Data resulting from these evaluations suggest that the language variants would not be evaluated differently by readers.

As an example, one particular realization of our stimulus might draw the first cue each section, forming the email:



Dear <John Adams>,

I have been hearing quite a bit about identification rules on the news. Do the changes affect <California>? I was wondering what I need to bring with me to vote?

Thank you,

<Daniel Nash>

Cue Type	Cue Text
Preamble	I have been hearing quite a bit about identification rules on the news.
Preamble	I have heard a lot on the news about identification.
Preamble	The news has talked a lot about identification rules.
Question 1	Do the changes affect <b>state</b> ?
Question 1	Are these changes happening in <b>state</b> ?
Question 1	Do these affect <b>state</b> ?
Question 2	I was wondering what I need to bring with me to vote?
Question 2	I was wondering if I need to bring anything specific with me to vote?
Question 2	Is there anything specific I need to bring to vote?

**Table A.2: Features manipulated for random assignment of messages to registrars of voters.**

## A.5. Pilot

We conducted three pilots prior to deploying the experiment. The first pilot was conducted in Minnesota, chosen because it was the locale utilized as a pilot in previous studies White, Nathan and Faller (2015a). Infrastructure problems meant that no emails were received by elections officials in the first pilot. We made changes, and conducted a second pilot in MN that successfully delivered emails. Finally, we conducted a third pilot in the western states of Washington, Oregon, California, and Nevada. These states were chosen due to their physical distance from other states, relatively small number of election officials, and peculiarities in election administration (e.g. Oregon does not conduct in-person elections).

## A.6. No Question Effects

In the following models, we report that the causal effects are invariant to including fixed effects for the specific questions asked.

**Table A.3: Question FE Model**

	<i>Dependent variable:</i>	
	GotResponse	
	(1)	(2)
Minority	-0.047*** (0.014)	
Latino		-0.030* (0.017)
Black		-0.0001 (0.017)
Arab		-0.111*** (0.017)
Question Fixed Effect	Yes	Yes
Observations	6,439	6,439
R <sup>2</sup>	0.006	0.013
Adjusted R <sup>2</sup>	0.002	0.009
Residual Std. Error	0.493 (df = 6411)	0.492 (df = 6409)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01	

## A.7. Name Selection

In this appendix, we describe our approach to selecting the names of constituents. Our primary intent in choosing names from population lists was to eliminate the possibility of any name-based confounds to be responsible for differences in the behavior of local elections officials. Injecting variation in this facet of the treatment also lowers the likelihood that officials would become aware of the intervention by observing messages across offices sent from the same alias. By varying the names used to signal identity, we break from the general practice in political science, which has been to select a small number of names – frequently one or two for each racial/ethnic group (e.g., Butler and Broockman, 2011*a*; White, Nathan and Faller, 2015*a*). Nevertheless our approach is in line with practices in the audit literature more broadly (see especially, Bertrand and Mullainathan, 2004*a*)

In line with previous work on election official responsiveness, we exclusively use male names (White, Nathan and Faller, 2015*a*). Using names from a single gender reduces the variance in outcomes that is not associated with race or ethnicity signals, increasing the efficiency of the experimental design.

We draw white first names from the social security administration’s records of births in Oregon in 1990. We utilize a list of distinctly African American names to produce our Black first names (Fryer and Levitt, 2004). Latino names are sourced from New York City baby names for children born between 2011 and 2014. Finally Arab/Muslim names were sourced from a list of common names (<http://www.behindthename.com/names/usage/arabic/>). Our intent in using this varied set of name sources was twofold. First and foremost, we wanted to generate plausible first names as an experimental stimulus. Second, we took care to ensure that the list of names we utilized was unlikely to match other name lists used in name-based audit studies.

We generate non-Hispanic White, Black, and Latino surnames from a US Census list of

the 1000 most commonly occurring surnames (Word et al., 2008). This dataset provides information about the distribution of racial and ethnic groups by each surname. For example, among individuals with the most commonly occurring surname, Smith, the census data identifies that 73% identify as non-Hispanic White, 22% identify as Black, and 1.5% identify as Hispanic. To select names, we set minimum levels within each category. For a surname to be chosen as a white surname, more than 70% with that name needed claim a non-Hispanic White identity. For a surname to be chosen as a black surname, 30% or more of people with that surname needed claim a black identity; for Latino surnames we set this threshold at 60%. We note that this choice was made to produce what were, in our estimation, names that strongly signaled racial/ethnic group, without utilizing the *most* common surnames associated with these groups.

Arab/Muslim names, and indeed demographic and health statistics are difficult to identify (Al-Sayed, Lauderdale and Galea, 2010). Consequently, we sourced surnames from <http://surnames.behindthename.com/names/usage/arabic>. This site does not provide frequency counts for names, so we assigned a uniform probability to each name being assigned.

With the set of first and last names created, we join the names together to produce a *given name* and *surname* pair that signals senders' racial/ethnic identities.

After constructing and curating a list of names to be sent as racial and ethnic primes, we recruited a set of workers through Amazon's *Mechanical Turk* (mTurk) worker platform. We paid mTurk workers a small amount to guess the probability that a particular name was of one or another ethnic group. Specifically, for each of 25 randomly selected names (from the set of  $\approx 400$ ) we asked workers to estimate their confidence (ranging from 0 percent to 100 percent) that an individual with a given name belonged to a particular racial or ethnic group.

As an example – the example we used in training workers for the mTurk task – we provided the name Yao Ming, a famous Chinese basketball player who played in the American NBA for 8 seasons. If a subject were certain that the name Yao Ming was a member of the Asian racial or ethnic group, the worker would place a certainty of 100 with this group. If the worker were mostly certain – for example 90 percent certain – that the name Yao Ming belonged to the Asian racial or ethnic group, she would place a 90 with that group and the remaining 10 percent certainty with other group(s) she thought the name may belong.

The results of this task are reported in section A.17, Table A.16.

## A.8. Blocking

We block on measures that are likely to predict whether a voting official will respond to (a) any form of contact and (b) forms of contact from minority voters. Specifically, we block on population density, proportion below 150 percent of the federal poverty line, proportion Black, proportion Latino, President Obama’s margin of victory in the 2012 Presidential Election, and previous coverage by §5 of the VRA.

Our blocking data was most commonly measured at the county level – e.g. county electoral returns. However, the relevant electoral area addressed by a local election official may, or may not also be a county. In some states local election officials execute elections across multiple counties; in other states local elections officials represent a single county; while in still others officials might work at the municipal level. When our blocking features were more geographically broad than the area covered by a local election official, we apply the county level values to the municipal level. When our blocking features were more narrowly measured than the political geography covered by an official, we simply average the county-level measurements. Details of implementation can be found in the notebooks

that accompany this work.

Blocking was implemented via the `blockTools` package written by Ryan Moore (Moore 2012.) Blocks of size four were created using an ‘optimalGreedy’ blocking algorithm. The algorithm begins by identifying the best pair of individual units to place in a single block, then identifies the best additional unit to include in that block, until the specified magnitude of the block is reached. It repeats the process until all units are blocked. We did not permit blocks from being formed between units in different states. In Table A.4 we report the results of our blocking strategy. In brief, blocking and subsequent randomization succeeded.

**Table A.4: Blocking**

	ethnic_cue	Mean Density	Mean Income	Mean Black	Mean Latino	Mean Obama	Mean VRA
1	White	1.860	0.044	0.043	0.055	-0.063	0.120
		0.019	0.001	0.003	0.003	0.007	0.008
2	Latino	1.850	0.045	0.043	0.055	-0.061	0.117
		0.020	0.001	0.003	0.003	0.007	0.008
3	Black	1.850	0.045	0.044	0.056	-0.060	0.120
		0.020	0.001	0.003	0.003	0.007	0.008
4	Arab	1.840	0.045	0.043	0.054	-0.065	0.118
		0.020	0.001	0.003	0.003	0.007	0.008

*Notes.* Standard errors are reported beneath variable means



**Table A.5: Response Rates by Experimental Condition**

<b>Ethnic Cue</b>	<b>White</b>	<b>Minority</b>	<b>Latino</b>	<b>Black</b>	<b>Arab</b>
Response Rate (%)	61.3	56.6	58.4	61.4	50.1
Standard Error	1.21	0.71	1.23	1.21	1.25
N	1,611	4,828	1,609	1,613	1,606

*Notes:* The *Minority* column includes all data from the *Latino*, *Black*, and *Arab* columns. Response rates and standard errors are reported in percentage terms.

## **A.9. Nonparametric Results**

The table reproduced in this section produces the non-parametric, difference in means between the white, minority, latino, black and Arab name-cues. As we report in Figure 5.1, minority, latino and Arab names receive responses at rates lower than white names. There is no detectable difference between the response rates of black and white names.

## A.10. Fixed Effects Models

Table A.6 presents linear probability models estimating the same causality quantities reported in Figure 5.1 in the main body of the paper, though we provide more information in this Appendix. Models 1 and 2 estimate the causal effect of voter contact sent by non-white voters (model 1) and specific racial and ethnic classes of voters (model 2), but without including block-specific fixed effects. Models 3 and 4 estimate these same relationships, but include block fixed effects. Models 1 and 2 estimate robust (HC3) standard errors; models 3 and 4 estimate robust standard errors as constructed in the `lfe`, version `lfe_2.5-1998`.

We note that, while all models reported herein use *HC3* standard errors, we obtain substantively similar results when using Bell-McCaffery small-sample standard errors recommended by Lin and Green (2015).

In Model 1, we estimate that the local election officials respond to 61.3 percent of the emails they received from white voters. Emails received from racial and ethnic minority voters received a response at a rate 4.7 percent lower than this baseline: 56.6 percent of emails sent by minority names received a local election official response. Model 3 estimates the same relationship, but de-means the estimates within each block. The estimate of the causal relationship between sending an email as a minority voter rather than a white voter does not change substantively, although the blocking does improve the efficiency of the estimator.

In Models 2 and 4 we examine whether different racial and ethnic minority groups are treated differently by the local election officials. We find evidence to support this hypothesis. Models that do (Model 4) and do not (Model 2) include block fixed effects both find that emails from a Latino voter are 3.0 percent less likely to receive a response than emails sent from a white voter. In contrast, emails sent from Black voters are treated very similarly as emails sent from white voters. The estimate of the causal relationship is

very nearly zero ( $\beta = 0.1$  percent), and is roughly 1/30 the magnitude of the latino effect. In both Models 2 and 4 we estimate Arab/Muslim aliases receive a response from elections officials at a rate 11.3 percentage points lower than the baseline response rate.

**Table A.6: Causal Estimates**

	GotResponse			
	(1)	(2)	(3)	(4)
Minority	-4.700*** (1.410)		-4.710*** (1.330)	
Latino		-2.970* (1.730)		-2.990* (1.630)
Black		0.110 (1.720)		0.167 (1.650)
Arab		-11.300*** (1.740)		-11.300*** (1.630)
Constant	61.300*** (1.210)	61.300*** (1.210)		
Block FE	No	No	Yes	Yes
Observations	6,439	6,439	6,439	6,439
R <sup>2</sup>	0.002	0.009	0.330	0.337

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## A.11. Robust to Link Function

While OLS estimators are unbiased estimates of the causal effect under this research design, we demonstrate that the choice of link function in a general linear model does not meaningfully alter estimates. In Table A.7 and Table A.8, we use a maximum likelihood approach to estimating these models, first with a gaussian link function, but also with logit and probit functions.

**Table A.7: Robust to Logit and Probit Specification**

	<i>Dependent variable:</i>		
	GotResponse		
	<i>normal</i>	<i>logistic</i>	<i>probit</i>
	(1)	(2)	(3)
Minority	−0.047*** (0.014)	−0.194*** (0.059)	−0.121*** (0.037)
Intercept	0.613*** (0.012)	0.461*** (0.051)	0.288*** (0.032)
Observations	6,439	6,439	6,439
Log Likelihood	−4,589.000	−4,379.000	−4,379.000
Akaike Inf. Crit.	9,183.000	8,762.000	8,762.000
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01		

**Table A.8: Robust to Logit and Probit Specification**

	<i>Dependent variable:</i>		
	GotResponse		
	<i>normal</i>	<i>logistic</i>	<i>probit</i>
	(1)	(2)	(3)
Latino	-0.030* (0.017)	-0.124* (0.072)	-0.077* (0.045)
Black	0.001 (0.017)	0.005 (0.072)	0.003 (0.045)
Arab	-0.113*** (0.017)	-0.459*** (0.072)	-0.286*** (0.045)
Intercept	0.613*** (0.012)	0.461*** (0.051)	0.288*** (0.032)
Observations	6,439	6,439	6,439
Log Likelihood	-4,567.000	-4,356.000	-4,356.000
Akaike Inf. Crit.	9,141.000	8,721.000	8,721.000

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## A.12. Pilot Inclusion

We piloted our delivery and intake engineering in two separate pilots. The first, executed in Minnesota, was initially met with technical implementation issues – we received server information that no emails from our system were being delivered to local election official addresses. We addressed this issue, and, because our forensics determined that it would not be possible for officials to be aware of our first pilot, we re-ran this pilot and were successful on this follow-up attempt. To ensure that our engineering was not only a Minnesota-specific success, we ran a second pilot in the Western states of Washington, Oregon, California, and Nevada. We chose these states because of their relatively small local election official population (233 total local election officials), and their distance from locales with many local election officials.

As we report in Table A.9 and Table A.10, neither including nor excluding these pilot states from the analysis changes the substance or the interpretation of the core results. In addition, there is no evidence that the causal effect is different in pilot compared to non-pilot states.

**Table A.9: Robust to Pilot Exclusion**

	<i>Dependent variable:</i>		
	GotResponse		
	(1)	(2)	(3)
Minority Cue	-0.047*** (0.014)	-0.046*** (0.014)	-0.046*** (0.014)
Pilot			0.120* (0.065)
Minority Cue * Pilot			-0.034 (0.076)
Constant	0.613*** (0.012)	0.609*** (0.013)	0.609*** (0.013)
Include Pilot	Yes	No	Yes
Observations	6,439	6,206	6,439
R <sup>2</sup>	0.002	0.002	0.003
Adjusted R <sup>2</sup>	0.002	0.001	0.003

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Table A.10: Robust to Pilot Exclusion**

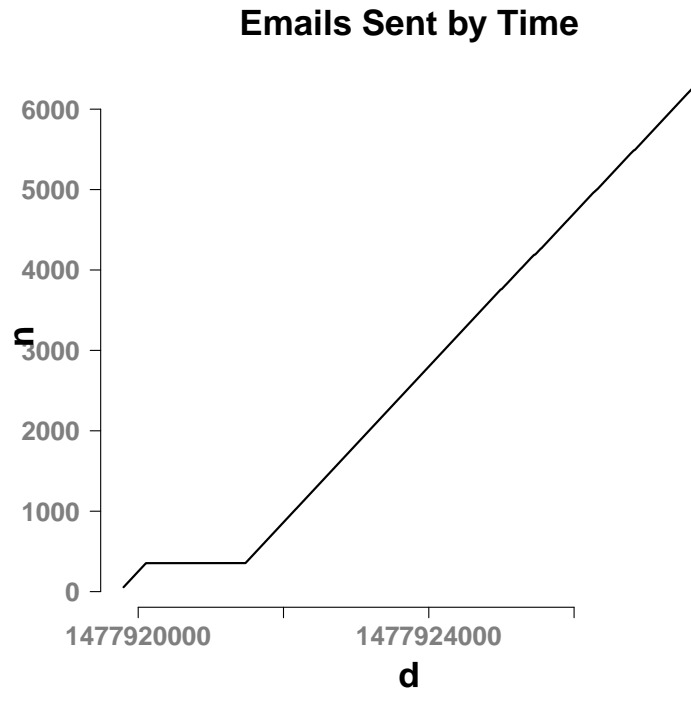
	<i>Dependent variable:</i>		
	GotResponse		
	(1)	(2)	(3)
Latino Cue	-0.030*	-0.030*	-0.030*
	(0.017)	(0.018)	(0.018)
Black Cue	0.001	0.005	0.005
	(0.017)	(0.018)	(0.018)
Arab Cue	-0.113***	-0.112***	-0.112***
	(0.017)	(0.018)	(0.018)
Pilot			0.120*
			(0.065)
Latino Cue * Pilot			0.021
			(0.093)
Black Cue * Pilot			-0.107
			(0.092)
Arab Cue * Pilot			-0.013
			(0.093)
Constant	0.613***	0.609***	0.609***
	(0.012)	(0.012)	(0.012)
Include Pilot	Yes	No	Yes
Observations	6,439	6,206	6,439
R <sup>2</sup>	0.009	0.009	0.010
Adjusted R <sup>2</sup>	0.008	0.009	0.009

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01



## **A.13. Email Send Timing**

In this appendix, we describe the timing of sending our emails. Emails were delivered in waves over a few hours to officials in the sample. We decided against emailing all local election officials at the same time to reduce the chance of unexpected results due to technical errors and to reduce possible spillover effects. We also considered emailing local election officials over a period of multiple days. Ultimately, we were concerned that the likelihood of differential response rates on different days outweighed the benefits to spreading email messages across several days. Note the 30 minute gap in sending. Here, we waited to ensure that emails were making it to officials' inboxes, before green-lighting the remainder of the production email run. We determined that our stimulus was making it to election officials inboxes when we received replies from officials in several states.



h

**Figure A.4:** The number of emails sent is marked on the y-axis, and the time (in UNIX seconds, in the UNIX epoch) are plotted on the x-axis. Note the 30 minute gap in sending. Here, we waited to ensure that emails were making it to officials' inboxes, before green-lighting the remainder of the production email run.

## A.14. Time to Response

In this appendix, we consider how much time was required for local election officials to respond to our email. To do so, we merge tracker hits from our server with the time that we received an email reply. The tracker hit records when a registrar opened the email, and the response effectively records when the task is complete.

We take some care in computing this, because election-official-side email clients handle our tracker hits differently. In particular, some email clients “cache” a version of our image on their own servers to speed up the loading of images in emails. When this occurs, we do not receive reliable information about when an email was opened.

We work around this problem by including only the *first* load that occurs on our sever. Not only does this preclude problems with individuals’ email clients, but at the same time we believe it also represents a conservative (long) estimate of the time to complete the task.

As we plot in Figure A.5, the task that we set before election officials did not require a substantial amount of time. Of those responses that we received, and have valid data for, the median time to respond was fewer than three minutes. It is, however, important to note that we neither have information about the time to respond for officials who do not respond to our stimulus, nor for officials whose email clients prohibit us from gathering reliable data.

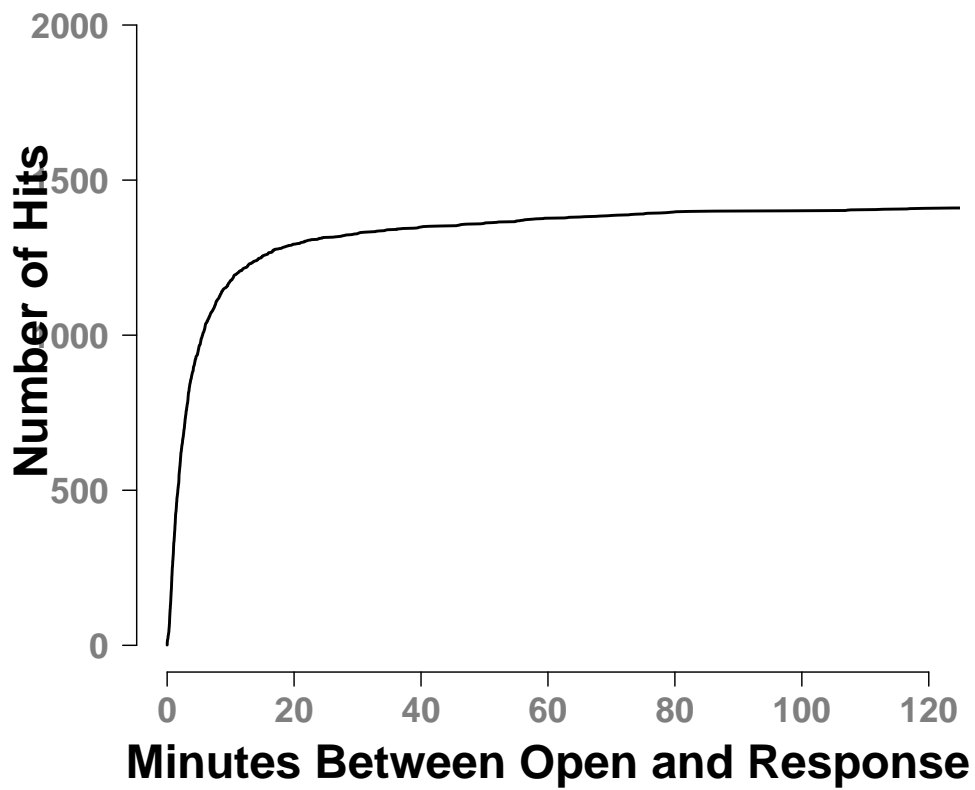


Figure A.5: On the x-axis are the minutes elapsed since the first time the local election officials opened our stimulus, until the time that we received a response from that election official. On the y-axis are the cumulative number of responses that have been received in that duration of time.

## A.15. No Damage from Spillover

After we collected outcome data, we learned that election officials in some states were suspicious about the emails, and contacted their state organization who, in turn, contacted the national organization. As well, we came to learn that at least one other research team was pursuing a substantively similar project, using the domain registered by White, Nathan and Faller (2015*a*).

To examine whether this notification seems to have affected the willingness of elections officials to respond we estimate two distinct robustness checks. First we estimate a number of Cox proportional hazard (duration) models. We choose this model class because they are unbiased in the presence of censored data. In particular, this model type permits us to estimate models that use the pre-registered end date of observation, as well as the timing of the NASS clerk email as the end date of observation. As we report in Table A.11, the coefficients estimated in all models are highly stable.

As a second robustness check, we estimate our core, pre-registered models again, but excluding states where the news reported early awareness: Michigan, New Hampshire, and Colorado. The results we report in Table A.12 retain their statistical significance and substantive interpretation. Although these are not dispositive tests, this set of results do not surface any evidence to suggest that the differences in response rates we observe are being caused by awareness.

Table A.11: Cox Proportional Hazards Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Minority Cue	-0.13*** (0.04)	-0.14*** (0.04)	-0.13*** (0.04)	-0.13*** (0.04)				
Latino Cue					-0.10* (0.05)	-0.10* (0.05)	-0.08* (0.05)	-0.07 (0.05)
Black Cue					-0.02 (0.05)	-0.03 (0.05)	-0.01 (0.04)	-0.02 (0.04)
Arab Cue					-0.29*** (0.05)	-0.29*** (0.05)	-0.31*** (0.05)	-0.30*** (0.05)
Data Subset	Clean	Clean	All	All	Clean	Clean	All	All
Censoring Date	Election	Clerk	Election	Clerk	Election	Clerk	Election	Clerk
Observations	4,548	4,548	6,435	6,435	4,548	4,548	6,435	6,435
R <sup>2</sup>	0.002	0.002	0.002	0.002	0.01	0.01	0.01	0.01

Notes. Cox proportional hazards models. Outcome is converting from no response to response. *Clean* data subset are states without known spillover, and exclude pilot data. *All* data subset includes all states' data. Two censoring points are estimated. *Election* is the pre-registered censoring date at election day; *Clerk* places the censoring date at the time of the NASS email notification. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Table A.12: No Difference in Estimates in Interference States**

	<i>Dependent variable:</i>							
	GotResponse				HitTracker			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Minority	-0.045*** (0.013)	-0.048*** (0.016)			-0.033** (0.014)	-0.036** (0.016)		
Latino			-0.028* (0.016)	-0.036* (0.019)			-0.005 (0.017)	-0.001 (0.020)
Black			0.004 (0.016)	-0.002 (0.019)			-0.006 (0.017)	-0.016 (0.020)
Arab			-0.110*** (0.016)	-0.107*** (0.019)			-0.088*** (0.017)	-0.092*** (0.020)
ln(pop dens)	0.152*** (0.029)	0.117*** (0.035)	0.149*** (0.029)	0.112*** (0.034)	0.063** (0.029)	0.059* (0.035)	0.060** (0.029)	0.054 (0.035)
Pct < 150	(0.693)	(0.766)	(0.689)	(0.763)	(0.700)	(0.785)	(0.698)	(0.783)
Pct Black	-0.498** (0.236)	-0.227 (0.301)	-0.483** (0.235)	-0.213 (0.300)	-0.033 (0.239)	0.508* (0.309)	-0.014 (0.238)	0.532* (0.308)
Pct Latino	-0.355* (0.202)	-0.161 (0.241)	-0.367* (0.201)	-0.179 (0.240)	-0.333 (0.204)	-0.177 (0.247)	-0.341* (0.203)	-0.192 (0.246)
Obama Margin	-0.006 (0.084)	0.058 (0.095)	-0.015 (0.083)	0.053 (0.095)	-0.031 (0.085)	0.006 (0.098)	-0.039 (0.084)	0.001 (0.097)
Observations	6,439	4,552	6,439	4,552	6,439	4,552	6,439	4,552
R <sup>2</sup>	0.334	0.327	0.341	0.332	0.282	0.284	0.287	0.289
Adjusted R <sup>2</sup>	0.109	0.097	0.118	0.104	0.039	0.040	0.046	0.046

*Note:*

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## A.16. Limited District Characteristic Heterogeneity

In the following models, reported in Table A.13, Table A.14, Table A.15, we examine whether officials' response to treatment is different conditional on characteristics of their district. In particular, one hypothesis is that officials who preside over jurisdictions that hold a relatively large share of minority voters may be more likely to respond to a question about voting from minority voters. Indeed, as we show in Table A.13 and Table A.14, while there is little change in the responsiveness of election officials as the proportion of voters in that jurisdiction becomes increasingly black (shown in *Model (2)* and *Model (3)* in both Table A.13 and Table A.14), as we report in *Model (1)* in Table A.13 and Table A.14, there is some evidence that officials' responsiveness changes as the proportion of Latinos in a jurisdiction increases.

Of particular interest is the possibility that the large treatment effects for the Arab/Muslim cue are driven by the implausibility of the treatment, due to the very small proportion of Arab Americans living in many jurisdictions. The results below are motivated by the following logic: if treatment effects for a given identity are driven by implausibility then they should be smaller in places where individuals who have been ascribed that identity are more numerous.

The distribution of Arab Americans is somewhat distinct from the distribution of blacks and Latinos. Indeed, data from the current CPS suggests that just 8 percent of U.S. counties have no Latino population, and 25 percent have no black population. In contrast, fully half of the counties in the U.S. have no residents who identify with an Arab heritage. Thus, it is possible that the lack of variation in the `pct_arab` population variable has made it mechanically impossible for a regression to detect a heterogeneous treatment effect.

To examine whether this is possible, we rescale the percent of Arab population into a three-level factor variable in the following way:



**Table A.13: TEH - Communities by Minority Share**

	<i>Dependent variable:</i>		
	GotResponse		
	(1)	(2)	(3)
Minority	-0.052*** (0.015)	-0.048*** (0.015)	-0.044*** (0.015)
Percent Latino	-0.241 (0.236)		
Percent Latino × Minority	0.093 (0.143)		
Percent Black		-0.163 (0.230)	
Percent Black × Minority		0.013 (0.133)	
Percent Arab			1.580 (2.440)
Percent Arab × Minority			-1.270 (2.530)
Observations	6,439	6,439	6,406
R <sup>2</sup>	0.330	0.330	0.329
Adjusted R <sup>2</sup>	0.104	0.103	0.101

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Table A.14: TEH - Communities by Ethnicity**

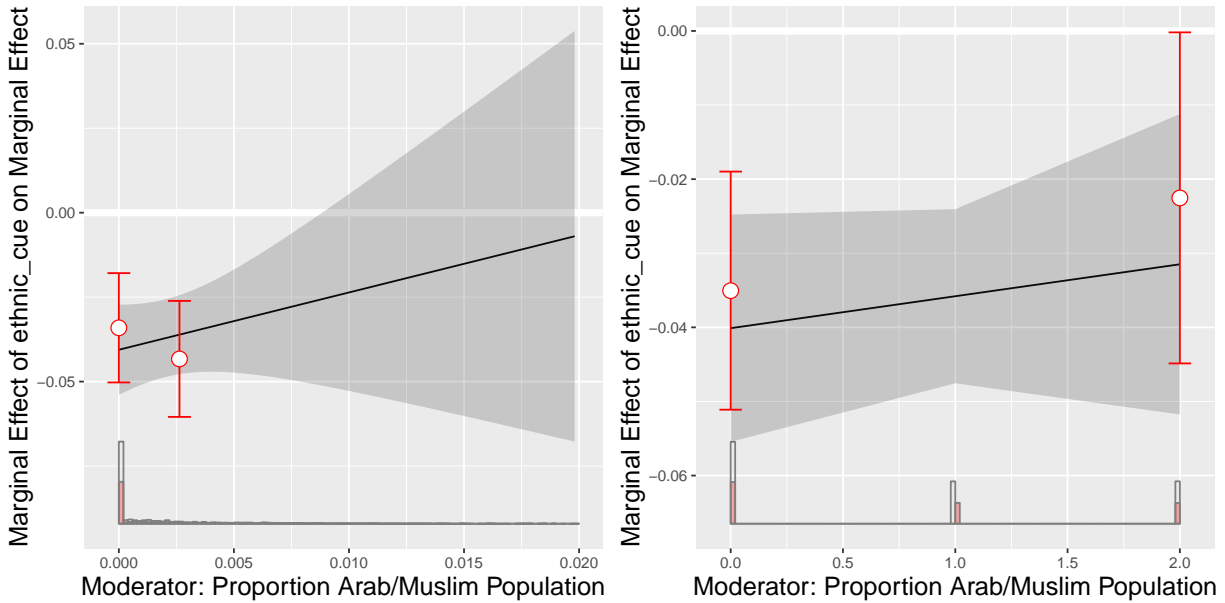
	<i>Dependent variable:</i>		
	GotResponse		
	(1)	(2)	(3)
Latino	-0.049*** (0.019)	-0.026 (0.018)	-0.028 (0.018)
Black	0.013 (0.019)	-0.003 (0.018)	0.003 (0.018)
Arab	-0.121*** (0.019)	-0.113*** (0.018)	-0.109*** (0.018)
Percent Latino	-0.227 (0.233)		
Percent Latino × Latino	0.345** (0.167)		
Percent Latino × Black	-0.199 (0.174)		
Percent Latino × Arab	0.138 (0.168)		
Percent Black		-0.173 (0.234)	
Percent Black × Latino		-0.098 (0.162)	
Percent Black × Black		0.119 (0.166)	
Percent Black × Arab		0.008 (0.156)	
Percent Arab			1.680 (2.460)
Percent Arab × Latino			-0.850 (2.780)
Percent Arab × Black			-0.657 (2.770)
Percent Arab × Arab			-1.740 (2.670)
Block FE	Yes	Yes	Yes
Observations	6,439	6,439	6,406
R <sup>2</sup>	0.339	0.337	0.337

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

- For geographies that have zero Arab population, we code the rescaled variable as 0. This represents the 0-50th percentile distribution of communities arranged by Arab-American population;
- Among geographies that have at least one person who identified an Arab heritage, we make a further split at the median.
  - The lower of the two groups, the set of communities that represent the 50-75th percentile distribution of communities; and,
  - The higher of the two groups, the set of communities that represent the 75-100th percentile distribution of communities.

As noted, this indicator splits the Arab population into three categories. The first category covers the 50 percent of U.S. counties with no Arab population. The second covers the 25 percent of U.S. counties whose Arab-American population is below the median value for those counties in which any Arabs live. In these counties, Arab Americans still represent a small part of the population: 0.12%. The third category covers the remaining 25 percent of counties whose Arab population is above this median. In these counties with the greatest presence of Arab-Americans, this group represents, on average, 1% of the county population. In the geographies corresponding to this quartile of the distribution, it would not be uncommon for a local election official to be in the presence of an Arab-American person or family at a community gathering of several hundred people – such as a parade or high school graduation.

As we report in Table A.15 after rescaling the data in this way, there is little evidence that our treatment effects were moderated in geographies in which census data records a greater number of Arab Americans. Neither of the terms interacting the treatment with the recoded covariate described above yield point estimates with p-values approaching



**Figure A.6: Flexible estimates of HTE across Arab/Muslim population (left plot) and the three-level factor that indicates (0.0) zero arab/muslim population; (1.0) Less than 1% Arab/Muslim population; and (2.0) 1% or more Arab/Muslim population.**

standard thresholds of statistical significance. (We recognize that this failure to reject could be driven by insufficient statistical power.)

To provide further evidence, Figure A.6 reports estimates of the treatment effect of receiving an email from an Arab/Muslim sender rather than a white sender, using the `iterflex` method that flexibly estimates and projects treatment effects across moderating variables (Hainmueller, Mummolo and Xu, 2018). We note in the left plot that the uncertainty estimates rapidly expand among counties with larger arab populations due to the sparse nature of the data: there are only 5 counties with an Arab/Muslim population larger than 10%. On balance, the evidence presented here conforms to the logic presented above in support of our argument: treatments were not more influential in those places where Arab Americans are less numerous.

**Table A.15: TEH - Arab Communities**

	<i>Dependent variable:</i>	
	GotResponse	
	(1)	(2)
Minority Cue	-0.040** (0.021)	
Latino Cue		-0.020 (0.025)
Black Cue		0.004 (0.026)
Arab Cue		-0.106*** (0.025)
1-50pct Arab	0.073** (0.032)	0.073** (0.032)
51-100pct Arab	0.082** (0.034)	0.081** (0.034)
Minority Cue * 1-50pct Arab	-0.026 (0.035)	
Minority Cue * 51-100pct Arab	0.005 (0.036)	
Latino Cue * 1-50pct Arab		-0.024 (0.043)
Black Cue * 1-50pct Arab		-0.004 (0.043)
Arab Cue * 1-50pct Arab		-0.050 (0.043)
Latino Cue * 51-100pct Arab		-0.008 (0.044)
Black Cue * 51-100pct Arab		-0.0004 (0.044)
Arab Cue * 51-100pct Arab		0.026 (0.044)
Block FE	Yes	Yes
Observations	6,439	6,439
R <sup>2</sup>	0.332	0.340
Adjusted R <sup>2</sup>	0.107	0.115

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## A.17. Names and Assessment of Racial and Ethnic Group

Table A.16: Name Score Table

Name	Ethnic Cue	Mean White	Mean Latino	Mean Black	Mean Arab
Daniel Nash	White	97.6	0.9	1	0
Mathew Roberts	White	95	0	3.7	0
Alex Steele	White	94.6	0.4	5	0
Nicholas Austin	White	94.6	0.4	4.6	0
Zachary Fitzpatrick	White	94.3	0.7	4.1	0
Christopher Schmidt	White	93.7	0.1	3.4	0.1
Ryan Thompson	White	93.1	0	6.2	0
Timothy Bartlett	White	93	0	6	0
Corey Kennedy	White	93	0	7	0
Garrett Riddle	White	92.9	0.4	6.6	0
Austin Walsh	White	92.4	0.3	5.8	0
Christopher Rogers	White	92.1	0	7.9	0
Jacob Gates	White	92	0	6.7	0
Kyle Caldwell	White	92	0	6	0
Matthew Pratt	White	91.4	0	8.6	0
Joseph Mayer	White	91.3	0	8.7	0
Ian Thornton	White	90.5	0	9.5	0
Scott Sherman	White	89.5	0.2	8.8	0
Daniel Horn	White	89.3	0	2.5	0
Zachary Proctor	White	89	0	7.5	0
Brandon Hart	White	88.8	0	11.2	0
Nathan Brewer	White	88.3	0	2.8	0
Garrett Allen	White	87.5	0.6	11.9	0
John Miller	White	87.3	0	10.9	0

Robert Peterson	White	87.2	0	11.7	0
Dylan Garrett	White	86.9	0	7.5	0
Michael Quinn	White	86.7	0	13.3	0
Justin Kramer	White	86.4	0	8.2	0
Robert Todd	White	86.1	0.4	12.1	0
Travis Roberts	White	85.7	0.7	10.7	0
Richard Bowers	White	85.7	1.3	6.7	0
Jason Gillespie	White	85.4	0.4	7.1	0
Garrett Miller	White	85.3	0	14.7	0
Kyle Thompson	White	84.4	0	15	0
Dustin Lawson	White	84.2	0	15.3	0
Sean Cooper	White	84.1	0	15.3	0
James McPherson	White	83.2	0	14.6	0
Brandon Pierce	White	83.2	0.5	14.7	0
John Gregory	White	83	2.9	10.2	0
David Cochran	White	82.9	0	17.1	0
Seth Rodgers	White	82.9	0.7	6.4	1.4
Christopher Anderson	White	82.9	0.2	16.8	0
Tyler Reeves	White	82.5	0.4	12.9	0
Justin McIntyre	White	82.5	5.6	6.4	0
Matthew Moore	White	82.4	0.7	16.6	0.1
Stephen Peterson	White	81.9	0	16.2	0
Kyle French	White	81.8	0.9	13.6	0
Timothy Middleton	White	81.4	0	17.7	0
Ian Smith	White	81.3	0	18.7	0
Tyler Larson	White	81.1	0	18.9	0
Gregory Leblanc	White	80.8	0.4	11.5	1.5
Ryan Chapman	White	80.7	0.2	16.8	0
William Humphrey	White	80.6	0	19.4	0
Justin Mullins	White	80.5	0	11.4	0

Joshua Burke	White	80.4	0	14.2	0
Jacob Haas	White	80	0	2.2	0
Levi Wolfe	White	80	0	0	0
Kevin Patterson	White	80	0	19.1	0
Jeremy Short	White	79.6	0	18.7	0
Cody Lang	White	79.4	0	3.1	0
Taylor Long	White	79	0	17.7	0
Zachary Bailey	White	78.8	0	12	0
Michael White	White	77.8	0	16.7	0
Jeffrey Phillips	White	77.1	0.4	21.7	0
Travis Miller	White	77.0	0	23.0	0
Brian Bennett	White	76.9	0	19.4	1.2
Robert Cochran	White	76.4	2.3	12.7	4.5
Michael Hendrix	White	76.2	0	17.9	0
Travis Osborn	White	75.4	0.8	7.1	0
Michael Boyer	White	75.3	0	15.3	1.3
Travis Collins	White	75	0	24.3	0
Christopher Hebert	White	74.7	0.7	22.7	0
Samuel Peters	White	74.5	0	18.2	0
Shane Page	White	74.4	1.2	24.4	0
Jeffrey Fox	White	74.4	0.8	8.1	0
Anthony Underwood	White	73.8	0	23.8	0
Justin Lyons	White	73.5	6.7	18.0	0
Michael Rose	White	71.9	3.8	23.1	0
Devin Foster	White	71	0	27	0
Joshua Clark	White	70	0	5	0
Jordan Rogers	White	69.7	0	21.6	0
Joseph Graves	White	68.8	0	17.8	6.2
Robert Reed	White	68.2	1.7	10.2	16.7
Tyler Murray	White	67.3	2	24	1.3



James Marsh	White	66.9	1.2	13.8	0
Travis Frye	White	66.8	0	24.1	0
Cameron Young	White	65.6	0	23.7	0
Stephen Sherman	White	64.6	0	26.9	0
Benjamin Wood	White	64	0	14.5	0
Eric Murray	White	61	0	29	0
Andrew Allen	White	60.9	0	28.4	0
Austin Hall	White	59.5	0	24.1	1.8
Samuel Wood	White	55.8	0	44.2	0
Marcus McFarland	White	55.5	0	44.5	0
Michael Lang	White	55.5	2.7	12.3	0
Samuel Hopkins	White	51.2	0	34.6	1.7
Brandon Estes	White	50.8	36.6	11.6	0
Sean Watts	White	40.4	1.8	50.7	1.4
Jordan Smith	White	39.6	0	50.4	0
Jose Hanson	White	9.5	77.5	12.5	0
Jose Cruz	Latino	0	100	0	0
Jorge Castro	Latino	0	100	0	0
Cesar Marquez	Latino	0	100	0	0
Jose Gutierrez	Latino	0	100	0	0
Juan Campos	Latino	0	100	0	0
Saul Gonzalez	Latino	0	100	0	0
Miguel Salazar	Latino	0	100	0	0
Jesus Perez	Latino	0	100	0	0
Diego Velazquez	Latino	0	100	0	0
Fernando Hernandez	Latino	0	100	0	0
Juan Ramos	Latino	0	99.6	0	0
Jose Valdez	Latino	0	99.6	0.4	0
Edwin Vasquez	Latino	0.6	99.4	0	0
Gerardo Escobar	Latino	0.8	99.2	0	0

Esteban Herrera	Latino	0	99.2	0	0
Jose Mendez	Latino	0	98.2	0.7	0
Luis Gomez	Latino	1.1	97.9	0.5	0
Fernando Acosta	Latino	1.1	97.8	0	0
Adriel Hernandez	Latino	0.8	97.3	1.2	0
Aldo Garcia	Latino	0	97.3	0	0
Jaime Gonzalez	Latino	1.4	97.1	1.4	0
Alejandro Rodriguez	Latino	0	96.9	3.1	0
Emilio Gonzalez	Latino	0.4	96.8	2.1	0
Esteban Contreras	Latino	2.3	96.6	0	0
Dariel Valdez	Latino	0	96.2	1.2	0
Enrique Lopez	Latino	3.8	96.2	0	0
Camilo Lopez	Latino	1.1	96.1	0	0
Miguel Barrera	Latino	0.7	95.7	1.8	0
Angel Ruiz	Latino	2	95.5	0.5	0
Roberto Reyes	Latino	0	95	5	0
Edwin Santiago	Latino	5.4	94.6	0	0
Angel Navarro	Latino	0	94.4	5.6	0
Ricardo Gomez	Latino	0.7	94.3	0.3	0
Marvin Lopez	Latino	3.6	92.7	2.7	0
Alejandro Ibarra	Latino	0.4	92.7	2.7	0
Jesus Hernandez	Latino	1.3	92.3	1.7	1.3
Emilio Cabrera	Latino	7.7	92.3	0	0
Cristian Ramirez	Latino	1.2	92.2	0	0
Jesus Martinez	Latino	2.1	92.1	1.4	1.4
Julio Morales	Latino	0.4	92.1	0	7.1
Adan Perez	Latino	2.5	91.5	0	0
Angel Maldonado	Latino	3.8	91.2	0	0
Darwin Gonzales	Latino	4.2	90.8	4.6	0
Dariel Garcia	Latino	2.1	90.7	6.4	0

Esteban Jimenez	Latino	0	90.4	1.9	0
Alberto Mendoza	Latino	0.7	90	1.4	0
Edgar Garcia	Latino	9	90	1	0
Miguel Rubio	Latino	0	89.1	9.1	0
Pablo Escobar	Latino	5.6	88.9	0	5.6
Luis Martinez	Latino	0	88.9	11.1	0
Carlos Villarreal	Latino	1.9	88.8	0.8	0
Luis Gonzalez	Latino	3.3	88.3	0	0
Jean Lopez	Latino	7.9	88.2	2.6	0
Carlos Ramos	Latino	1.4	88.2	0	0
Juan Perez	Latino	2.5	86.7	10.8	0
Ricardo Garza	Latino	5.8	86.7	1.7	1.7
Manuel Padilla	Latino	0	86.4	0	4.3
Miguel Rodriguez	Latino	1.8	86.4	0.9	0
Angel Pineda	Latino	5	85	1.2	1.2
Luis Moreno	Latino	2.5	84.6	0	0
Iker Martinez	Latino	3.2	83.9	1.1	0.7
Edgar Cardenas	Latino	8.7	83.7	1.7	0
Edwin Hernandez	Latino	11.1	83.5	3	0.5
Mario Chavez	Latino	3.6	82.1	1.4	1.4
Johan Estrada	Latino	8.3	80.7	0.9	0.7
Jefferson Sanchez	Latino	9.3	80.7	9.3	0
Johan Garcia	Latino	11.7	80.6	3.9	0
Emiliano Lopez	Latino	1.7	80	1.7	1.7
Erick Hernandez	Latino	13.8	79.4	5.3	0
Giovani Herrera	Latino	14.2	79.2	0	1.7
Luis Padilla	Latino	3.5	78.8	1.9	0
Randy Munoz	Latino	14.5	78.8	0	0
Jadiel Rodriguez	Latino	1.7	78.8	15.8	0.4
Brayan Estrada	Latino	2.8	78.2	9.5	1

Erik Rodriguez	Latino	7.7	78.2	0.5	0
Erick Suarez	Latino	13.5	76.9	2.7	1.5
Maximo Flores	Latino	9.7	76.1	3.2	0
Yaniel Campos	Latino	1.2	74.4	5.9	1.2
Miguel Trevino	Latino	0.9	72.6	5	0
Yair Fuentes	Latino	0	69.5	4.1	18.2
Matias Murillo	Latino	4.8	69	1	6
Anderson Guerrero	Latino	18.8	68.8	2.5	1.2
Edwin Castaneda	Latino	21.1	68.2	0	0
Kenny Rodriguez	Latino	27.1	67.4	0.9	1.2
Damian Martinez	Latino	13.7	66.8	18.2	0
Januel Aguilar	Latino	7.2	66.1	8.3	1.7
Noel Torres	Latino	22.3	65.9	11.8	0
Ismael Romero	Latino	5.8	60.4	4.2	24.6
Derick Torres	Latino	21.8	59.5	13.2	1.8
Julius Salazar	Latino	8.8	58.4	2.2	8.8
Angel Ponce	Latino	14.2	52.8	19.2	1.1
Thiago Zamora	Latino	2	52.5	6.5	6
Junior Delgado	Latino	15	50.4	30	0
Kenny Lozano	Latino	35.4	45.7	8.9	0
Jael Calderon	Latino	13.3	44	29.3	0
Darwin Guzman	Latino	26.0	42.4	17.4	0.7
Edwin Zuniga	Latino	12.7	38.7	22.7	3.3
Byron Salazar	Latino	34.2	31.5	24.6	6.9
Jean Barrera	Latino	45	23	5	2
Jefferson Ponce	Latino	55.9	0.5	28.2	0
DeShawn Jackson	Black	2.4	0	97.6	0
Tyrone Brown	Black	1.2	1.7	96.7	0
DeShawn Harris	Black	2.9	0.3	96.7	0
DeShawn Brown	Black	2.1	0	96.7	0

Darius Thomas	Black	2.5	0	96.2	1.2
DeAndre Jackson	Black	1.4	0.8	96.1	0
Jamal Jones	Black	1.8	0	95.4	0
DeShawn Glover	Black	4	1	95	0
Tyrone Thomas	Black	3.9	0.6	94.7	0
Terrell Turner	Black	4.4	0	94.4	0
Darnell Jackson	Black	5.7	0	94.3	0
Terrell Watkins	Black	5	0.8	93.1	0.4
Trevon Williams	Black	7.1	0	92.9	0
Darius Haynes	Black	6	0.7	92.7	0
DeAndre Wilkins	Black	5.3	0.3	92.3	0
Darnell Haynes	Black	7.5	1.1	91.4	0
DeShawn Ware	Black	5.4	0	91.2	0
DeAndre Scott	Black	5.8	0.4	91.2	0
Trevon Johnson	Black	0.9	0	90.9	0
Tyrone Jones	Black	9.2	0	90.8	0
Jalen Washington	Black	6.9	0	90.8	0
Darius Davis	Black	9.3	0	90.7	0
Darnell Alexander	Black	8.3	0.5	90.4	0
DeShawn Anthony	Black	3.5	0	90	0
Demetrius Jackson	Black	10	0	90	0
Darnell Davis	Black	11.8	0	88.2	0
Terrell Davis	Black	10.9	0	88.2	0.9
Jamal Coleman	Black	7.5	0.5	88	4
Tyrone Johnson	Black	8.5	0	87.7	0
Darius Washington	Black	11.8	0.6	87.6	0
Marquis Harris	Black	6.5	5	87	0
Malik Johnson	Black	5.5	0	86.4	6.4
Maurice Brown	Black	13.8	0	86.2	0
Tyrone Harris	Black	11.5	0.3	85.5	0

DeShawn Johnson	Black	13.6	0	85	0
DeAndre Davis	Black	12.7	1	85	0
Terrell Ware	Black	6	1.8	84.5	1.8
Andre Harris	Black	13.1	1.5	84.2	0
Jamal Williams	Black	10.5	1.1	84.2	1.1
Darnell Mitchell	Black	15.4	0	83.9	0
Darnell Carter	Black	10.3	0	83.8	0
Terrance Terrell	Black	13.5	1.2	83.5	0
Terrell Scott	Black	12.5	0.2	83	0
Terrance Johnson	Black	17.5	0	80.8	0
Andre Johnson	Black	19.3	0.2	80.4	0
Terrell Washington	Black	12.3	0	80.3	0
Demetrius Johnson	Black	14.5	0.5	79.1	0
Darryl Willis	Black	20	0	79	0
Dominique Richardson	Black	18.4	2.7	78.9	0
Darius Miles	Black	20.5	0.5	78.6	0
Darius Willis	Black	13	0	78.3	0
Dominique Brown	Black	16.2	0	77.2	0
Darius Bryant	Black	20	1.1	77.2	0
Trevon Grant	Black	20	1.7	77.1	0
Trevon Henry	Black	20.6	2.1	76.8	0
Reginald Brown	Black	13	8.5	76.5	0
Marquis Williams	Black	15	0.8	75.7	0
Dominique Walker	Black	21.8	1.6	75.5	0
Malik Hawkins	Black	15.9	0.3	75.3	8.3
Tyrone Dorsey	Black	25	0	75	0
Terrance Robinson	Black	16	0.2	73.8	0
Darius Byrd	Black	20.4	0	73.5	0
Malik Williams	Black	0.3	0.8	73.3	19.7
Jalen Walker	Black	27.1	0	72.3	0

Trevon Scott	Black	25.8	0	71.7	0
Maurice Miles	Black	25.2	0.5	71.5	0
Malik Mitchell	Black	6.7	0	71	14
Jamal Johnson	Black	6	0	71	3
Xavier Brown	Black	16.2	6.9	70.3	0
Dominique Jones	Black	22.7	4.5	70	0
DeAndre Mathis	Black	16.3	3.7	69.7	0
Maurice Davis	Black	29	0.6	69.4	0
Terrell Thomas	Black	8.3	8.3	69.2	8.3
Reginald Coleman	Black	33.3	0	66.7	0
Jalen Neal	Black	20	0	65.8	0
Jalen Harris	Black	17.8	2.8	65	0
Maurice Thomas	Black	27	1.3	64.3	0
Darryl Brooks	Black	28.9	7.1	62.1	0
Reginald Davis	Black	39.2	0	60.8	0
Malik Robinson	Black	14.4	0	60.6	18.9
Marquis Mitchell	Black	17.7	3.1	60.4	0
Terrance Woods	Black	39.3	0	60.4	0
Jalen Johnson	Black	10	0	60	3.3
Demetrius Fields	Black	23.5	2.4	60	0
Dominique Simmons	Black	27.7	11.2	59.6	0
Jalen Thomas	Black	26.8	4.5	59.5	0
Darryl Watkins	Black	39.1	0	57.7	0
Jalen Carter	Black	36	0	57.5	0
Xavier Scott	Black	37.8	0.6	56.7	3.3
Xavier Willis	Black	20.7	20	56.4	0
Willie Davis	Black	40	1	56	0
Malik Neal	Black	16.3	0	55.8	14.2
Xavier Brooks	Black	28.1	0.8	55	0
Dominique Alexander	Black	30.6	12.1	55	0

Willie Brown	Black	37.8	0.4	54.8	0.9
Darryl Williams	Black	28	0	54.5	0
Willie Jones	Black	39	2.5	54.5	0
Willie Williams	Black	43.3	0	54.3	0
Dominique Matthews	Black	34.7	8.8	53.5	0
Andre Miles	Black	35.8	9.2	52.3	0
Xavier Davis	Black	44	0.3	49	0
Darryl Brown	Black	44.4	0.6	47.8	0
Darryl Davis	Black	53.2	0	45	0
Willie Singleton	Black	46.2	0	43.8	0
Reginald Turner	Black	45	5.6	40.8	0
Jalen Holmes	Black	33.6	0	40.5	0
Darryl Walker	Black	57.3	0.7	40	0
Willie Nixon	Black	71.4	0	13.6	0
Basir Albaf	Arab	0	0	0	99.2
Botros Ahmed	Arab	0	0	0	98.4
Sami El-Amin	Arab	0	0	1.7	97.8
Salah Darzi	Arab	0	0	2.2	97.8
Abd El-Mofty	Arab	0	0.5	0.9	97.7
Sharif Abdullah	Arab	0	0	2.9	97.1
Shahnaz Hussain	Arab	0	0	0	96.8
Duha El-Amin	Arab	0	0	1.5	95.8
Shams El-Amin	Arab	0.1	0.1	3.3	95.6
Ibrahim El-Hashem	Arab	0	0	1.8	95.5
Mahdi Albaf	Arab	0	0	1.8	94.7
Bakr Abdullah	Arab	0	0	0	94.5
Husain Sultan	Arab	0	0	0	94.4
Sajjad Ahmed	Arab	0.6	0	1.2	94.1
Fayiz Muhammad	Arab	0	0	1	94
Ghassan Ahmed	Arab	6.2	0	0	93.8



Ghayth Abdullah	Arab	0	0	4.7	93.6
Ramadan Muhammad	Arab	0	0	4.4	93.3
Maalik El-Ghazzawy	Arab	0	0	1.9	93.1
Hafeez Saab	Arab	0	0	3	93
Tarik El-Amin	Arab	0	0	5	93
Abbas Abdullah	Arab	0	0	4.2	92.9
Imad Zaman	Arab	0	0	1.4	92.9
Mohammed Ahmed	Arab	0	0	3.8	92.5
Jabr Hussain	Arab	5.9	0	1.8	92.4
Hikmat Ahmad	Arab	1.2	0	0	92.2
Bahadur Abdullah	Arab	0.7	0	0	92.1
Al-Amir Bousaid	Arab	0	0	0.3	92.1
Shadi Bousaid	Arab	0	0	0	91.7
Jalal El-Amin	Arab	0	0	1.9	91.5
Nasim Abdullah	Arab	0	0	2.6	90.9
Salil Albaf	Arab	2.1	0	0.7	90.7
Hakim Ajam	Arab	0	0	8.7	90.7
Boulos Amjad	Arab	1.2	3.8	1.9	90.6
Baqir Ali	Arab	3.3	0	0.8	89.2
Mohammed Boulos	Arab	0	0	11.2	88.8
Bahij Nejem	Arab	0	0	0.9	88.6
Zahi El-Mofty	Arab	0	0	0.7	88.6
Gafar Hakim	Arab	0	0	2.9	88.6
Hussein Darzi	Arab	0.6	1.8	3.2	88.2
Basir Muhammad	Arab	0	2.1	8.6	88.2
Sa'Di Albaf	Arab	0	6.7	3.7	88
Mukhtar Amjad	Arab	0.5	0	6.5	87.8
Tahir El-Amin	Arab	0	4.6	2.4	87.6
Yuhanna El-Amin	Arab	0	0	6.2	86.9
Aamir Abujamal	Arab	0	0	0.8	86.7

Husain El-Mofty	Arab	10.9	0	0.9	86.4
Fadl Nejem	Arab	0	0	0	85.7
Halim Zaman	Arab	0	0	2	85.5
Imran Hakim	Arab	7.7	1.5	1.5	85.4
Samir Abdulrashid	Arab	0	0	1.1	84.6
Ihsan El-Mofty	Arab	0	0	0	84.5
Tarek Saqqaf	Arab	0.7	0	6	84
Abdul-Aziz El-Mofty	Arab	0	0	1.6	83.2
Wadud Hakim	Arab	1.2	0	13.8	82.5
Shukri Saqqaf	Arab	0	0	3.8	82.3
Yaser Karimi	Arab	0	0	3.2	81.6
Fakhri Ali	Arab	0.1	0	5.3	80.8
Nabil Saab	Arab	0.6	0	7.8	80.6
Ziauddin Muhammad	Arab	0	0	1.2	80
Rayyan Albaf	Arab	0	0	5	79.3
Rasul Ajam	Arab	0	0.3	1.5	78.8
Nour El-Ghazzawy	Arab	1.5	0	3.1	78.5
Rifat Alfarsi	Arab	0	0	6.7	78.3
Sajjad El-Amin	Arab	0	0	5	78.3
Sa'Di El-Ghazzawy	Arab	0.7	0	8	77.3
Fayiz Samara	Arab	1.5	0	2.3	76.2
Aali Hussain	Arab	0	11.1	1.1	75
Imran Mohammed	Arab	1.1	0	6.7	74.4
Nizar Kader	Arab	0	0	2.8	73.9
Jaffer Bousaid	Arab	6.9	0	1.2	73.8
Jafar Sultan	Arab	0.3	0	17.6	73.2
Shafiq Samara	Arab	0.9	0	16.8	73.2
Fayiz Nejem	Arab	0	0.3	2.6	72.4
Salim Kader	Arab	0	0	10.4	72.1
Wafi Sultan	Arab	0	0	3.7	71.6

Husni Zaman	Arab	0	0	18	71.3
Adam Ahmad	Arab	7.4	5.2	7.4	71.0
Khaled Samara	Arab	0	3.3	14.7	70
Rasheed Zaman	Arab	2.7	0.7	22.7	70
Fakhri El-Mofty	Arab	1.8	0.3	12.9	68.8
Sameer Sultan	Arab	6.2	0	9.6	68.5
Guda El-Mofty	Arab	0	11	7.5	66.5
'Abbas Nagi	Arab	0	0	15.5	65
Adnan El-Mofty	Arab	0	0	8.3	64.2
Zaki Karim	Arab	1.1	0	20.3	63.9
Mis'Id El-Ghazzawy	Arab	0	0	0	63.3
Nurullah Nejem	Arab	0	1.1	10.8	61.9
Latif El-Mofty	Arab	0.5	3.2	29.2	61.6
Safi Boulos	Arab	0.4	7.7	0.4	61.5
Tayeb Kader	Arab	3.8	0	21.8	59.8
Waheed Bousaid	Arab	1.5	0	14.4	58.5
Mansoor Amirmoez	Arab	0	21.2	5.6	58.1
Dawud Karim	Arab	0	1.2	35.6	52.9
Tal'At Tawfeek	Arab	7.1	0	20	46.4
Murtaza Nagi	Arab	0.4	0.7	4.6	42.5
Ayman Amirmoez	Arab	0	28.1	0	41.9
Rusul Samara	Arab	1.8	5.9	14.5	41.4
Rais Nagi	Arab	0	0.1	1.9	40
Wafi Kader	Arab	2.5	0	23.8	33.8

---

# **Appendix B. Appendix for ‘Does Religious Bias Shape Access to Public Services? A Large-Scale Audit Experiment Among Street-Level Bureaucrats’**

## **B.1. Treatment effect heterogeneity**

We might expect treatment effects to vary with the social context in which principals are embedded. Examining treatment effect heterogeneity is complicated by the fact that we have a number of different treatments and many covariates. We simplify our analysis by collapsing the medium and high intensity levels, which, as shown in Figure 6.2, lead to very similar effect estimates. We do the same with Protestant and Catholic religious affiliation,

again for the same reason. Moreover, parametric models such as probit are poorly suited for modeling a large number of interactions of unknown functional form (Berry, DeMeritt and Esarey, 2010; Hainmueller, Mummolo and Xu, 2016). We therefore use Bayesian Additive Regression Trees (BART), one of the best off-the-shelf statistical learning estimators, to nonparametrically model treatment effect heterogeneity (Chipman et al., 2010; Hill, 2011). We investigate how the treatment effects vary as a function of school-level characteristics (the share of Asian, Hispanic, Black, and White students as well as school type (primary, middle, high) and county-level characteristics (median household income, the share of adults holding bachelor degrees, the share of residents with income below the poverty line, Republican vote shares in the 2012 presidential elections, and religious adherence rates from the RCMS). We also include a dummy variable for the South. Finally, since it is possible that principals in schools/communities with greater diversity discriminate less, we also compute Herfindahl indices for racial and religious diversity and investigate whether treatment effects vary systematically with these indices. While it would be advantageous to also incorporate covariates related to the characteristics of the principals themselves, especially their religious identification (or lack thereof), such data are not available. We do have principals' names and so could potentially estimate their ethnicity using the method proposed by Imai and Khanna (2016), but this approach does not work with religious identification.

Figures B.5 and B.6 show treatment effect estimates from a BART fit with 95% credible intervals. Despite our large sample size we find very little treatment effect heterogeneity. This is perhaps not too surprising given that covariates are measured at the school- or county-level and do not directly measure attitudes toward minority religions, non-believers, or the separation of church and state. (Fitting a linear probability model regressing the experimental outcome on all covariates produces an adjusted  $R^2$  of merely 0.029, indicating

that covariates not only fail to capture treatment effect heterogeneity but also variation in principals' responsiveness more generally.) We thus conclude that we observe substantively large levels of discrimination on average and that discrimination does not appear to be unique to any single social context. The patterns of discrimination we observe appear to be similar across schools.

## **B.2. Generalizing impact estimates to NCES universe**

We formally generalize our results to the NCES population of 78,348 regular, non-charter public schools in the 48 contiguous U.S. states without missing data. Following Kern et al. (2016), we generalize effect estimates by reweighting our experimental sample so that it matches the NCES target population in terms of covariate means for the covariates listed in Table B.1. Weights are generated using maximum entropy weighting (Hainmueller, 2012), which guarantees that reweighted sample covariate means equal the covariate means in the NCES population. Estimates from Weighted Least Squares regressions are reported in Table B.6. Reweighting the sample so that it matches the NCES population has a negligible effect on our treatment effect estimates, in line with the earlier findings that our sample and the NCES population match closely (Figures B.2–B.4) and that treatment effect heterogeneity is limited (Figures B.5 and B.6). Based on these results, it seems plausible that the patterns of discrimination we observe in our experimental sample are not unique to the schools in our experiment. It is likely that similar discrimination occurs across public schools throughout the U.S.

## B.3. Ethics

Butler and Broockman (2011*a*) and McClendon (2012) provide a detailed discussion of the ethical concerns involved in audit experiments like ours. We believe that our study has minimized any potential risk to principals and the communities that they serve and that this minimal risk was justified given the important contribution of our article. In what follows, we touch on several potential ethical concerns with our audit study—and audit studies in general—and summarize some responses to these issues. While our goal is not to address all potential issues with audit studies—a worthy goal for trained political theorists/ethicists—we do argue that our study was justified and well within bounds ethically.

Before we begin, however, we briefly note that our sample of study (public school principals) may be exempt from IRB review. Ethical principles outlined in federal guidelines often label “elected or *appointed* public officials or candidates for public office” as exempt from review requirements.<sup>110</sup> While we think it reasonable to argue that public school principals qualify under this category, the boundaries of what exactly constitutes an appointed public figure are somewhat fuzzy. Out of an abundance of caution, we still applied for and were given approval of two IRBs before conducting our study. And we were very careful in our study to try and minimize ethical concerns (as we discuss below). That said, we think this important to briefly note.

### Deception and Informed Consent

Perhaps the most common ethical concern with audit studies is that they involve deception of experimental subjects. Another related ethical concern is that they are conducted

---

<sup>110</sup>See, for example, “Exempt Review”, Northwestern University IRB.



without the informed consent of participants (Desposato, 2018).<sup>111</sup> In reply to both objections, researchers point out that they cannot answer their research question(s) if subjects are aware of the goals of the research since such awareness would change subjects' behavior. Deception and informed consent are thus necessary to answer some theoretically and normatively important questions. Federal ethics guidelines explicitly allow research to be conducted without informed consent and using deception under such circumstances. (Especially for public officials who are often labeled as exempt all together.) We believe that our study falls into this category of studies (and both of the IRBs that approved our project agree). Informing participants about the goals of our research would make it impossible to study illegal behavior such as religious discrimination.

At the core of the defense of studies such as ours is the argument that the benefit of knowing something new about the world outweighs the possible harms caused by deception or enrolling participants without consent. This might not always be true, however. We can imagine cases in which conducting research in this way might create considerable psychological or physical harm for participants. There is little evidence that this is typically the case with an audit study, however, since all that is usually required of respondents is that they reply to basic questions via email or mail. The same is true for our study, which simply contacted principals to ask for a face-to-face meeting with a prospective parent concerning the enrollment of their child. We can also imagine the case in which what is learned through deception and without informed consent does not outweigh even the slightest ethical harm. We do not think that this is the case for our experimental intervention, as it is the first to offer robust evidence of religious discrimination in the American public school system.<sup>112</sup> We believe that the results we obtain in our study are

---

<sup>111</sup>Using surveys of academics and potential study participants, Desposato (2018) shows that academics place a much higher ethical premium on studies with informed consent and no deception than study participants themselves.

<sup>112</sup>Even if the study were not novel, though, we might still think that conducting it would be worthwhile. A

substantively and normatively significant given the constitutional and legal prohibitions against religious discrimination by state institutions such as public schools and the lack of rigorous experimental research documenting the extent and nature of this discrimination. Consequently, we consider the minimal harms that our intervention could have caused to be reasonable in relation to the benefits of our work to society. Both of our IRBs agreed with this assessment.

## **Participant Duration**

One additional ethical concern with audit studies is that they incur a time cost on participants without their consent. The usual reply to this objection is that the time required of participants in audit studies is minimal—participants are only asked to read a short email and reply to it—and that deception is warranted (for the reasons outlined above). The same applies to our study. In order to minimize the time principals spent in our study, we only sent a single email to principals, with no follow-up in case of non-response. Moreover, replies were entirely voluntary and failure to reply did not impose any costs on subjects. We estimate that the average participant in our study spent much less than a minute reading our email inquiry. Based on the length of email replies we received, we estimate that those who responded to our emails probably took less than a minute doing so. In total, the amount of time that any individual spent participating in our study was very small and likely inconsequential to participants. As a result, we do not think that the amount of time respondents spent in our study raises any ethical concerns.

A separate, but related, issue is the total time that all respondents combined spend

---

growing literature suggests that the results from many studies (experimental or observational) cannot be reproduced. One reason why this might be the case is that study results are unique to specific samples or geographic or temporal contexts. To help ensure that the results from any one audit study are reproducible, we should want researchers to conduct similar studies with different participants and in different places and times.

in an audit study. The concern here is that large-scale audit studies of public officials, such as ours, might impose a burden on the public by impeding officials from working on more important tasks. This concern highlights a trade-off in audit study research. On the one hand, researchers want to maximize the number of subjects enrolled in their studies. Doing so helps ensure that researchers have sufficient power not only to estimate treatment effects but also to investigate how treatment effects vary across subgroups or how well they generalize to a given target population. On the other hand, the larger the subject pool enrolled in an audit study, the more of a collective time burden the study imposes on the public. Given that the time that any principal spent in our study was minimal, we do not think that the total time burden was unreasonable either—especially given the potentially large combined benefits. Indeed, we designed our experiment to take as little time of principals as was possible. One should also keep in mind that the costs to society of discriminatory behavior likewise scale with the size of the target population. The larger the target population, the larger, *ceteris paribus*, the experimental sample can be, which increases the total time burden. However, the importance of discovering discriminatory behavior also scales with the size of the target population (in our case, U.S. principals). So while our experimental sample was (of necessity) large compared to many audit studies, the importance of documenting religious discrimination among American public school principals is also of particular importance for the reasons outlined in our paper. (We are somewhat skeptical of the aggregation of time costs argument as it is unclear to us who is actually bearing this collective cost. That said, if one accepts the aggregation argument, we assert that they should also agree to the aggregation of benefits of a study.)

## Downstream Consequences

Another set of ethical concerns revolves around the idea that audit studies might change the behavior of public officials. The general idea here is that officials could become sensitized to the possibility that they are being studied. Therefore, they could come to doubt the origin and truthfulness of any correspondence they receive, which could in turn lead them to ignore requests or respond to them in less helpful ways.

This outcome strikes us as extremely unlikely. The amount of time it would take officials to determine the true source of an email is likely to be much greater than the time required to simply respond to a request. So long as the total number of auditing emails arriving at each office remains low, it seems unlikely that officials would change their behavior based on assumptions about whether or not they are being audited. We have no reason to expect that the principals in our sample receive numerous communications originating in audit studies. We thus think that this concern is unwarranted.

That being said, based on this (very limited) possibility of negative downstream consequences, we designed our study to not debrief subjects. IRBs often have exemptions for debriefing when deception is involved; indeed, these assert that “debriefing may be inappropriate if debriefing regarding the deception may cause more harm to the participant than the deception itself.”<sup>113</sup> We argue that this holds in our case where the deception is small (and warranted).

All in all, we think a further discussion of the ethical considerations in doing field experiments like audit studies is worthwhile. Many of the concerns listed above are *not* unique to audit studies, but apply equally to other field experimental methodologies (e.g. GOTV studies) as well. In sum, we argue that our study was justified and well within bounds ethically.

---

<sup>113</sup>See “Guidelines for Research involving Deception or Incomplete Disclosure,” Northwestern University IRB.

**Table B.1: Balance**

covariate	<i>p</i> -value
% Asian students	0.97
% Hispanic students	0.28
% Black students	0.57
% White students	0.53
% Male students	0.48
School size	0.35
Pupil/teacher ratio	0.44
% Free or reduced price lunch students	0.57
% GOP (2012 Presidential election)	0.39
% Median household income	0.59
% Bachelor degree	0.40
% Below poverty line	0.75
% Total adherents	0.99
% Black protestant adherents	0.74
% Evangelical protestant adherents	0.52
% Mainline protestant adherents	0.15
% Catholic adherents	0.50
% Muslim adherents	0.56

Note: The table shows exact *p*-values from univariate randomization inference tests of the null hypothesis that balance is as good as one would expect under block random assignment. Exact *p*-values are approximated using 10,000 randomly chosen blocked treatment assignments. The test statistic is the maximum Kolmogorov–Smirnov statistic across all two-way comparisons of treatment groups. The *p*-value is the fraction of test statistics at least as large as the test statistic in our sample. For the religious adherence covariates, the balance tests use the average of five multiply imputed datasets.

**Table B.2: Parameter estimates (probit)**

	est.	se
intercept	0.131***	(0.037)
Male parent	-0.040***	(0.014)
Male child	-0.022*	(0.012)
Protestant, low intensity	0.018	(0.026)
Protestant, medium intensity	-0.137***	(0.026)
Protestant, high intensity	-0.141***	(0.026)
Catholic, low intensity	0.032	(0.026)
Catholic, medium intensity	-0.165***	(0.025)
Catholic, high intensity	-0.169***	(0.027)
Muslim, low intensity	-0.116***	(0.026)
Muslim, medium intensity	-0.220***	(0.027)
Muslim, high intensity	-0.200***	(0.026)
Atheist, low intensity	-0.118***	(0.026)
Atheist, medium intensity	-0.357***	(0.027)
Atheist, high intensity	-0.334***	(0.027)

Note:  $N = 45,710$ . The table shows estimates and robust standard errors clustered at the school district level from a probit model. The model contains email account fixed effects (not shown). Omitted categories are female parent, female child, and no information given.

\* denotes statistical significance at 0.10 level. \*\* denotes statistical significance at 0.05 level. \*\*\* denotes statistical significance at 0.01 level.

**Table B.3: Parameter estimates (probit) controlling for blocks**

	est.	se
intercept	-0.411	(0.253)
Male parent	-0.045***	(0.014)
Male child	-0.024**	(0.012)
Protestant, low intensity	0.017	(0.028)
Protestant, medium intensity	-0.147***	(0.027)
Protestant, high intensity	-0.156***	(0.027)
Catholic, low intensity	0.032	(0.028)
Catholic, medium intensity	-0.180***	(0.027)
Catholic, high intensity	-0.183***	(0.028)
Muslim, low intensity	-0.124***	(0.027)
Muslim, medium intensity	-0.241***	(0.029)
Muslim, high intensity	-0.217***	(0.028)
Atheist, low intensity	-0.127***	(0.027)
Atheist, medium intensity	-0.384***	(0.028)
Atheist, high intensity	-0.362***	(0.028)

Note:  $N = 45,710$ . The table shows estimates and robust standard errors clustered at the school district level from a probit model. The model contains email account fixed effects as well as block fixed effects (not shown). Omitted categories are female parent, female child, and no information given.

\* denotes statistical significance at 0.10 level. \*\* denotes statistical significance at 0.05 level. \*\*\* denotes statistical significance at 0.01 level.

**Table B.4: Parameter estimates (OLS)**

model	(1)		(2)		(3)	
	est.	se	est.	se	est.	se
intercept	0.551***	(0.014)	0.353***	(0.077)	0.670***	(0.157)
Male parent	-0.015***	(0.005)	-0.016***	(0.005)	-0.016***	(0.005)
Male child	-0.009*	(0.004)	-0.009**	(0.004)	-0.009**	(0.004)
Protestant, low intensity	0.007	(0.011)	0.006	(0.010)	0.006	(0.010)
Protestant, medium intensity	-0.054***	(0.010)	-0.053***	(0.010)	-0.053***	(0.010)
Protestant, high intensity	-0.056***	(0.010)	-0.056***	(0.010)	-0.056***	(0.010)
Catholic, low intensity	0.013	(0.010)	0.012	(0.010)	0.012	(0.010)
Catholic, medium intensity	-0.065***	(0.010)	-0.066***	(0.010)	-0.066***	(0.010)
Catholic, high intensity	-0.067***	(0.010)	-0.066***	(0.010)	-0.066***	(0.010)
Muslim, low intensity	-0.046***	(0.010)	-0.045***	(0.010)	-0.046***	(0.010)
Muslim, medium intensity	-0.086***	(0.011)	-0.087***	(0.010)	-0.086***	(0.010)
Muslim, high intensity	-0.079***	(0.010)	-0.079***	(0.010)	-0.079***	(0.010)
Atheist, low intensity	-0.047***	(0.010)	-0.046***	(0.010)	-0.045***	(0.010)
Atheist, medium intensity	-0.138***	(0.010)	-0.136***	(0.010)	-0.135***	(0.010)
Atheist, high intensity	-0.129***	(0.010)	-0.129***	(0.010)	-0.129***	(0.010)
Email account fixed effects	yes		yes		yes	
Block fixed effects	no		yes		yes	
Covariates	no		no		yes	

Note:  $N = 45,710$ . The table shows estimates and robust standard errors clustered at the school district level from three linear probability models. All models contain email account fixed effects (coefficients not shown), model (2) additionally contains block fixed effects (coefficients not shown), and model (3) additionally contains the covariates listed in Table B.1 (coefficients not shown). Omitted categories are female parent, female child, and no information given.

\* denotes statistical significance at 0.10 level. \*\* denotes statistical significance at 0.05 level. \*\*\* denotes statistical significance at 0.01 level.



**Table B.5: Parameter estimates with MA omitted (OLS)**

model	(1)		(2)		(3)	
	est.	se	est.	se	est.	se
intercept	0.555***	(0.015)	0.355***	(0.076)	0.655***	(0.159)
Male parent	-0.014***	(0.005)	-0.015***	(0.005)	-0.015***	(0.005)
Male child	-0.008*	(0.005)	-0.009**	(0.004)	-0.009**	(0.004)
Protestant, low intensity	0.009	(0.011)	0.008	(0.010)	0.008	(0.010)
Protestant, medium intensity	-0.055***	(0.010)	-0.054***	(0.010)	-0.054***	(0.010)
Protestant, high intensity	-0.052***	(0.010)	-0.053***	(0.010)	-0.052***	(0.010)
Catholic, low intensity	0.014	(0.011)	0.013	(0.010)	0.013	(0.010)
Catholic, medium intensity	-0.066***	(0.010)	-0.067***	(0.010)	-0.067***	(0.010)
Catholic, high intensity	-0.069***	(0.011)	-0.068***	(0.010)	-0.068***	(0.010)
Muslim, low intensity	-0.046***	(0.010)	-0.045***	(0.010)	-0.045***	(0.010)
Muslim, medium intensity	-0.087***	(0.011)	-0.088***	(0.011)	-0.087***	(0.011)
Muslim, high intensity	-0.079***	(0.010)	-0.079***	(0.010)	-0.079***	(0.010)
Atheist, low intensity	-0.048***	(0.010)	-0.047***	(0.010)	-0.047***	(0.010)
Atheist, medium intensity	-0.138***	(0.010)	-0.137***	(0.010)	-0.135***	(0.010)
Atheist, high intensity	-0.127***	(0.010)	-0.127***	(0.010)	-0.126***	(0.010)
Email account fixed effects	yes		yes		yes	
Block fixed effects	no		yes		yes	
Covariates	no		no		yes	

Note:  $N = 44,396$ . The state of Massachusetts has been omitted from the sample. The table shows estimates and robust standard errors clustered at the school district level from three linear probability models. All models contain email account fixed effects (coefficients not shown), model (2) additionally contains block fixed effects (coefficients not shown), and model (3) additionally contains the covariates listed in Table B.1 (coefficients not shown). Omitted categories are female parent, female child, and no information given.

\* denotes statistical significance at 0.10 level. \*\* denotes statistical significance at 0.05 level. \*\*\* denotes statistical significance at 0.01 level.

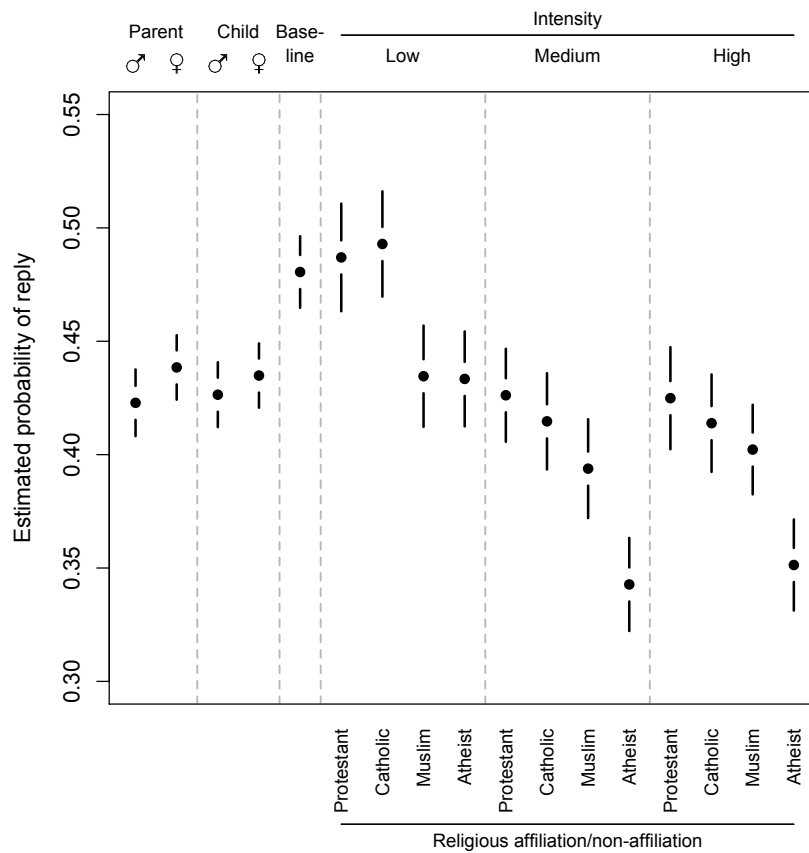
**Table B.6: Parameter estimates from population-weighted sample (WLS)**

	est.	se
intercept	0.554***	(0.017)
Male parent	-0.012**	(0.006)
Male child	-0.009*	(0.005)
Protestant, low fervor	0.014	(0.011)
Protestant, medium fervor	-0.045***	(0.011)
Protestant, high fervor	-0.056***	(0.011)
Catholic, low fervor	0.018	(0.011)
Catholic, medium fervor	-0.063***	(0.011)
Catholic, high fervor	-0.066***	(0.011)
Muslim, low fervor	-0.038***	(0.011)
Muslim, medium fervor	-0.086***	(0.011)
Muslim, high fervor	-0.081***	(0.011)
Atheist, low fervor	-0.046***	(0.011)
Atheist, medium fervor	-0.139***	(0.011)
Atheist, high fervor	-0.124***	(0.012)

Note:  $N = 45,710$ . The table shows estimates and robust standard errors clustered at the school district level from a linear probability model weighted with maximum entropy weights (see text). The model also contains email account fixed effects (coefficient estimates not shown). Omitted categories are female parent, female child, and no information given.

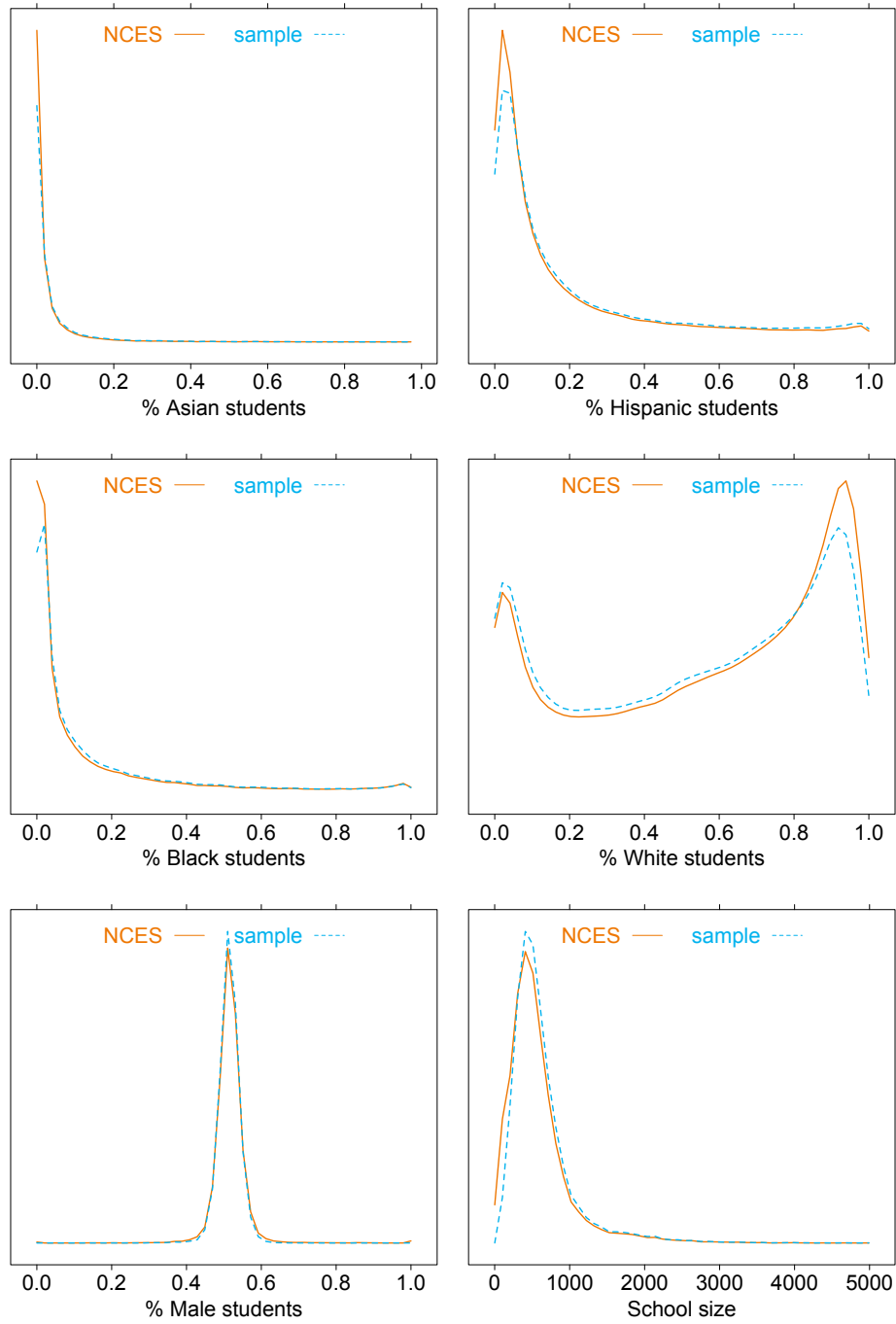
\* denotes statistical significance at 0.10 level. \*\* denotes statistical significance at 0.05 level. \*\*\* denotes statistical significance at 0.01 level.

Figure B.1: Estimated probabilities of reply based on model in Table B.2



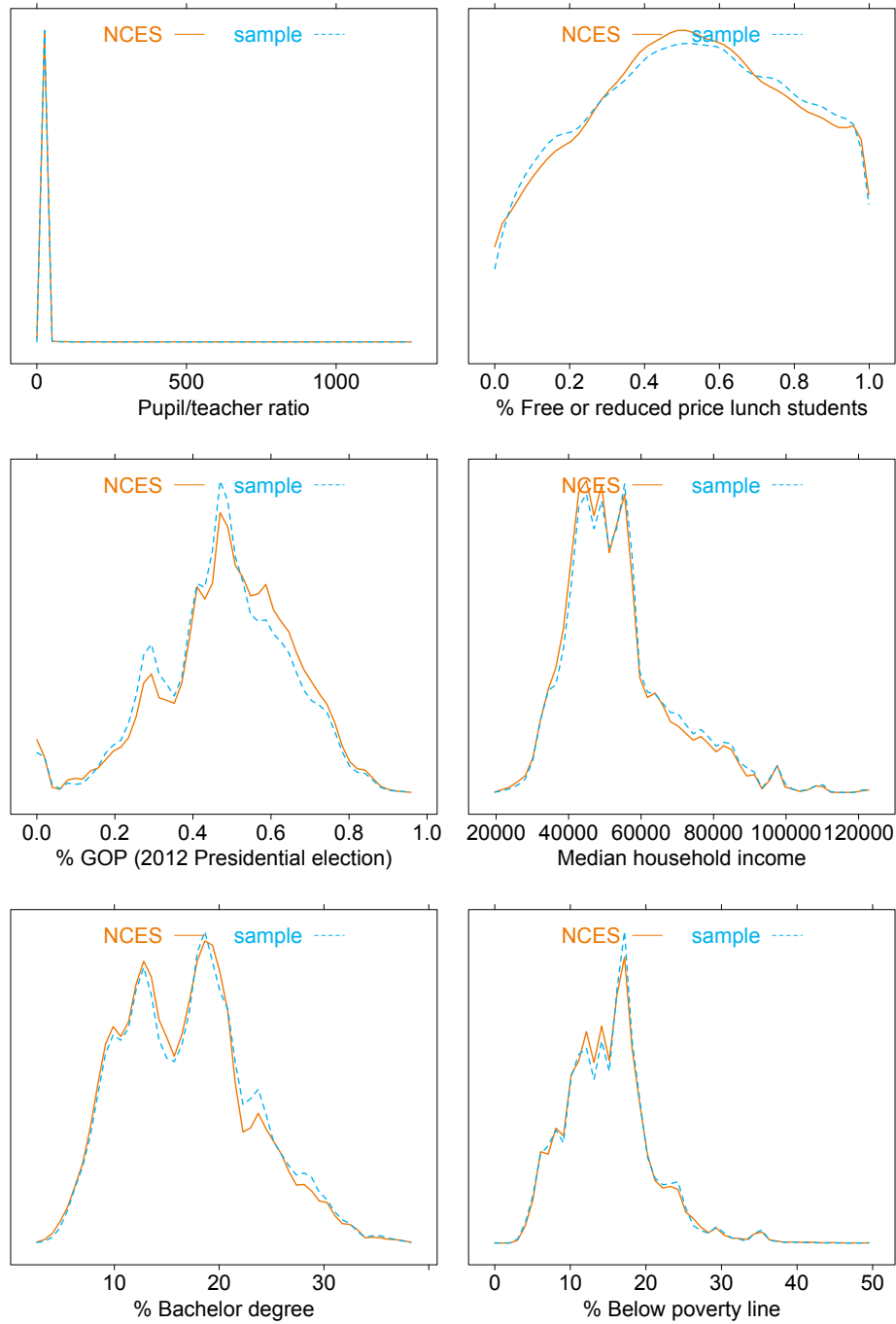
Note: The plot shows estimated probabilities of receiving a reply and 95% confidence intervals based on the probit model in Table B.2. Robust standard errors are clustered at the school district level.

Figure B.2: Comparison between sample and NCES population I



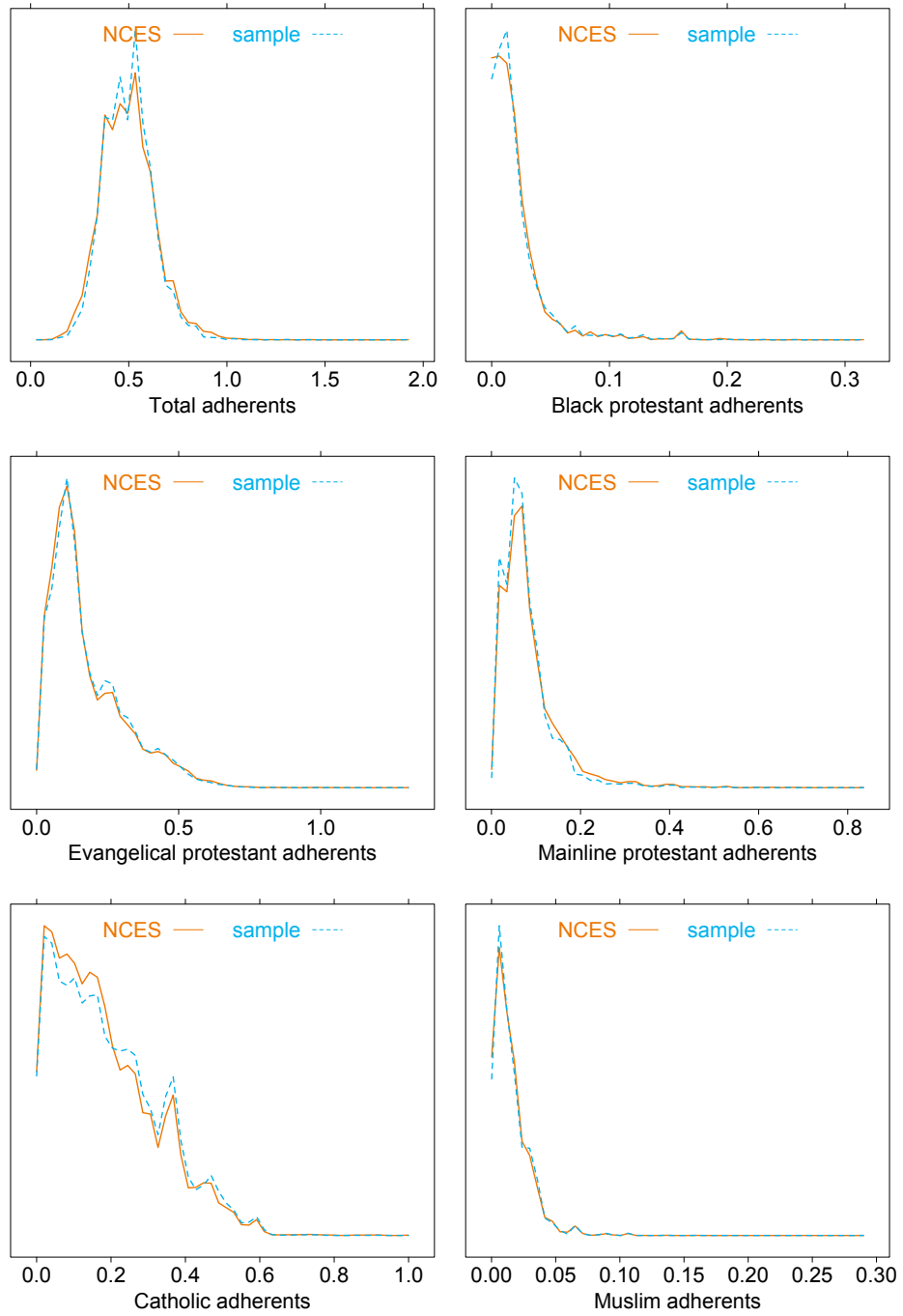
Note: The kernel density plots compare covariate distributions in the experimental sample with the NCES population of 78,348 regular, non-charter public schools without missing data in the 48 contiguous U.S. states.

Figure B.3: Comparison between sample and NCES population II



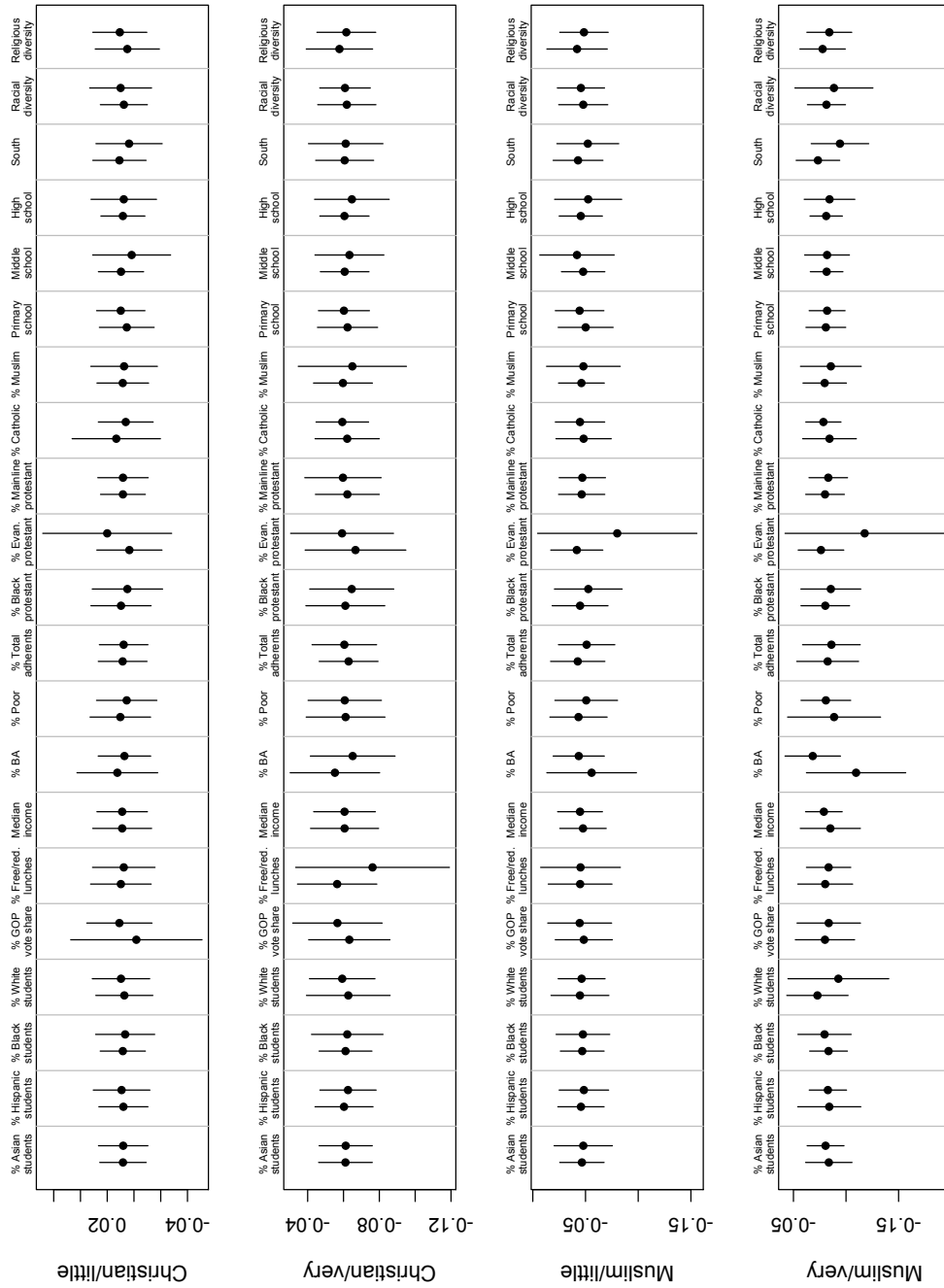
Note: The kernel density plots compare covariate distributions in the experimental sample with the NCES population of 78,348 regular, non-charter public schools without missing data in the 48 contiguous U.S. states.

Figure B.4: Comparison between sample and NCES population III



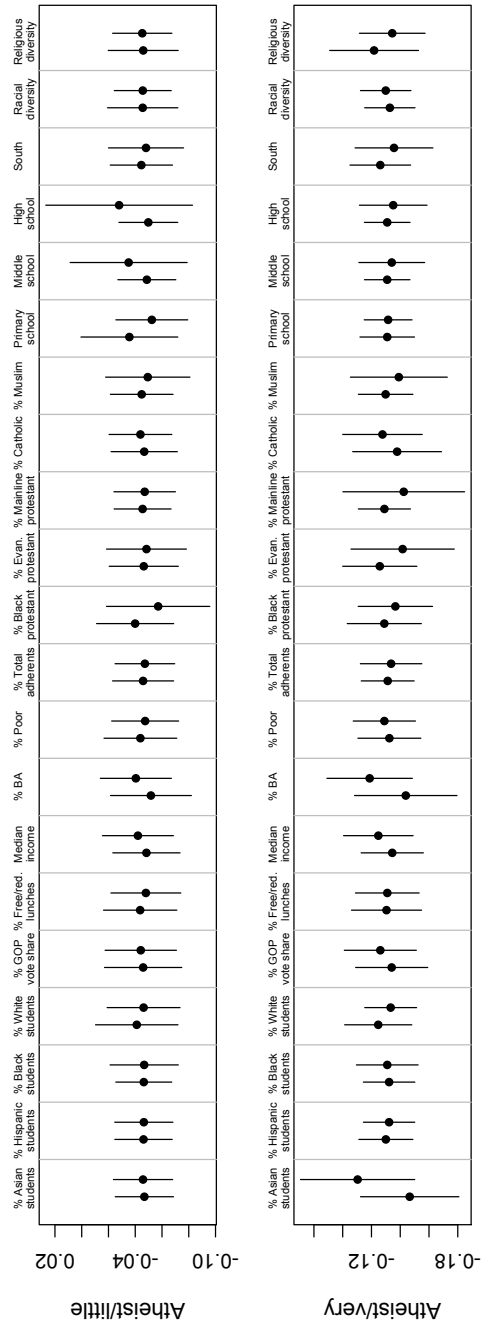
Note: The kernel density plots compare covariate distributions in the experimental sample with the NCES population of 78,348 regular, non-charter public schools without missing data in the 48 contiguous U.S. states.

Figure B.5: Treatment effect heterogeneity I



Note: The plots show treatment effect heterogeneity as a function of covariates from a Bayesian Additive Regression Tree (BART) fit, contrasting treatment effect estimates for when a given covariate is set to the 0.05 sample quantile and for when it is set to the 0.95 sample quantile.

Figure B.6: Treatment effect heterogeneity II



Note: The plots show treatment effect heterogeneity as a function of covariates from a Bayesian Additive Regression Tree (BART) fit, contrasting treatment effect estimates for when a given covariate is set to the 0.05 sample quantile and for when it is set to the 0.95 sample quantile.

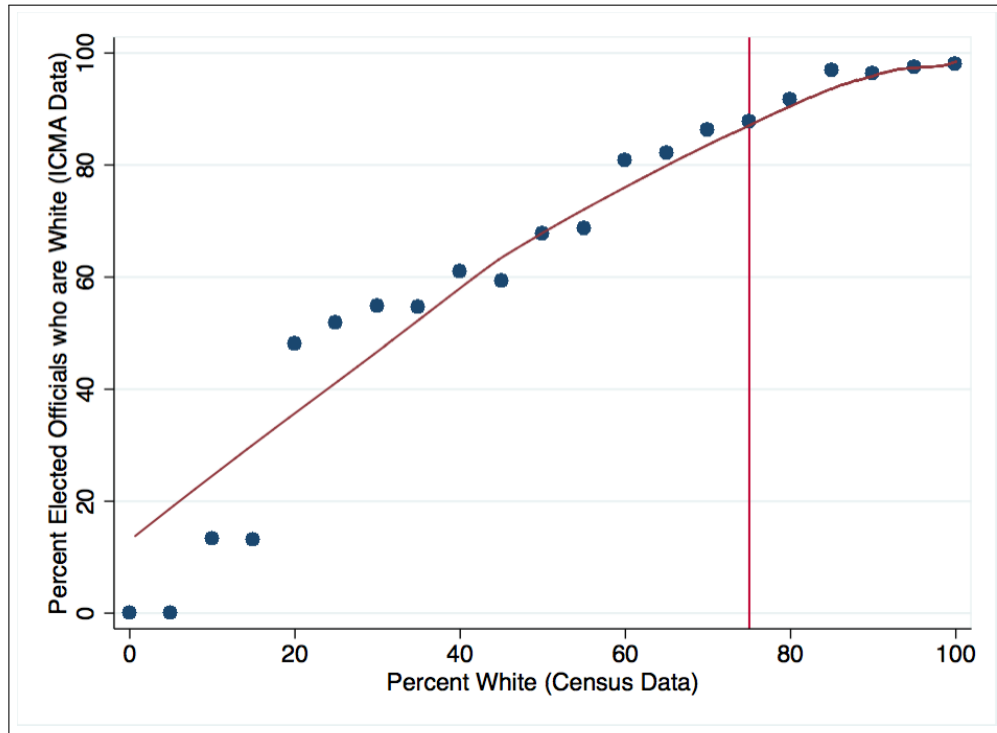


# **Appendix C. Appendix for ‘Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions’**

## **C.1. Census and ICMA Data**

Note: The data on city officials’ race comes from the 2011 International City/County Management Association (ICMA) city survey. The data on the city’s white population comes from the U.S. census. The cities in this Figure are restricted to cities where, according to the Census, less than 15 percent of the population is Latino. The vertical red line shows the cut point we used to determine which cities were included in the sampling frame.

Figure C.1: Race of Officials by Race of Constituents



## C.2. Treatment Email

Note: Identifying information has been redacted for the review.

## Figure C.2: Treatment Email 1

My name is Charles Crabtree. I am a graduate student in the Department of Political Science at The Pennsylvania State University. As part of a larger research project, I want to know what elected officials think about research that suggests that they exhibit bias in their constituent interactions. **Specifically, recent scholarship argues that office holders respond more often and provide better advice to individuals like them. White legislators, for instance, respond at higher rates to inquiries from white constituents than from black constituents.** For a summary of this research, please see the paragraph below. NPR also provides a nice summary of this research in a recent blog post: <http://n.pr/1s8QKrc>. If you are interested, you can also visit <http://sites.psu.edu/bias/research> for more details and a short bibliography of related works.

I want to survey some elected officials about this research. You have been randomly selected as a potential survey participant. Would you please complete a short survey on this topic? It would be very helpful for my research and should only take a few minutes of your time. You can find the survey here: <https://www.surveymonkey.com/s/FGZLPR9>.

Thank you.

All best,

Charles Crabtree  
Graduate student  
Department of Political Science  
The Pennsylvania State University

### Summary of Bias Research

The evidence for racial bias among officials is perhaps the strongest. For instance, some research shows that white legislators respond at higher rates to inquiries from white constituents than from black constituents and that black legislators similarly favor black constituents. Other research has revealed that white and black legislators alike consciously try to represent their own racial groups. A common racial tie can even prompt legislators to assist individuals who are not part of their constituencies.

Copyright © 2014. All rights reserved.  
You are receiving this email because you have been randomly selected as a potential survey respondent.

**My mailing address is:**  
The Pennsylvania State University  
University Park  
State College, Pa 16801

[Add us to your address book](#)

[unsubscribe from this list](#) [update subscription preferences](#)

## C.3. Research Summary

Figure C.3: Site Screenshot - Research Summary Page 1



### Recent Research on Bias

Using both observational methods and experimental studies, scholars have repeatedly found that office holders respond more often and provide better advice to individuals like them. **Research suggests that elected officials provide preferential treatment to those with similar backgrounds.** The evidence for racial bias among officials is perhaps the strongest. For instance, some research shows that white legislators respond at higher rates to inquiries from white constituents than from black constituents and that black legislators similarly favor black constituents. Other research has revealed that white and black legislators alike consciously try to represent their own racial groups. A common racial tie can even prompt legislators to assist individuals who are not part of their constituencies. The select publications below highlight some of the findings on bias.

*Select recent publications:*

- Butler, D. M. (2014). *Representing the Advantaged: How Politicians Reinforce Inequality*. Cambridge University Press.

**Overview:** Political inequality is a major issue in American politics, with racial minorities and low-income voters receiving less favorable representation. Scholars argue that this political inequality stems largely from differences in political participation and that if all citizens participated equally we would achieve political equality. Daniel M. Butler shows that this common view is incorrect. He uses innovative field and survey experiments involving public officials to show that a significant amount of bias in representation traces its roots to the information, opinions, and attitudes that politicians bring to office and suggests that even if all voters participated equally, there would still be significant levels of bias in American politics because of differences in elite participation. Butler's work provides a new theoretical basis for understanding inequality in American politics and insights into what institutional changes can be used to fix the problem.

Note: Identifying information has been redacted for the review.

## Figure C.4: Site Screenshot - Research Summary Page 2

- Harden, J. J. (2013). [Multidimensional Responsiveness: The Determinants of Legislators' Representational Priorities](#). *Legislative Studies Quarterly*, 38(2), 155-184.  
**Abstract:** Scholars of American politics typically conceptualize representation as mass-elite policy congruence, and in doing so have found several factors that hinder that relationship. These findings are at odds with the fact that American legislators often gain enough support to win re-election. I present an explanation for this puzzle by showing that legislators strategically provide four unique dimensions of representation to their constituents: policy, service, allocation, and descriptive. I unify these dimensions in a single theoretical model of legislators' priorities, then test it with data from survey experiments administered to 1,175 state legislators. I posit that legislators systematically emphasize some dimensions over others to further the goal of reelection. Given the constraints of resources and costs, legislators must choose their representational focus based on perceived electoral benefits. I find that institutional, district, and individual-level traits alter these resources, costs, and benefits, thereby driving legislators' strategic representational behavior.
- 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.  
**Abstract:** Why are politicians more likely to advance the interests of those of their race? I present a field experiment demonstrating that black politicians are more intrinsically motivated to advance blacks' interests than are their counterparts. Guided by elite interviews, I emailed 6,928 U.S. state legislators from a putatively black alias asking for help signing up for state unemployment benefits. Crucially, I varied the legislators' political incentive to respond by randomizing whether the sender purported to live within or far from each legislator's district. While nonblack legislators were markedly less likely to respond when their political incentives to do so were diminished, black legislators typically continued to respond even when doing so promised little political reward. Black legislators thus appear substantially more intrinsically motivated to advance blacks' interests. As political decision making is often difficult for voters to observe, intrinsically motivated descriptive representatives play a crucial role in advancing minorities' political interests.

Note: Identifying information has been redacted for the review.

### Figure C.5: Site Screenshot - Research Summary Page 3

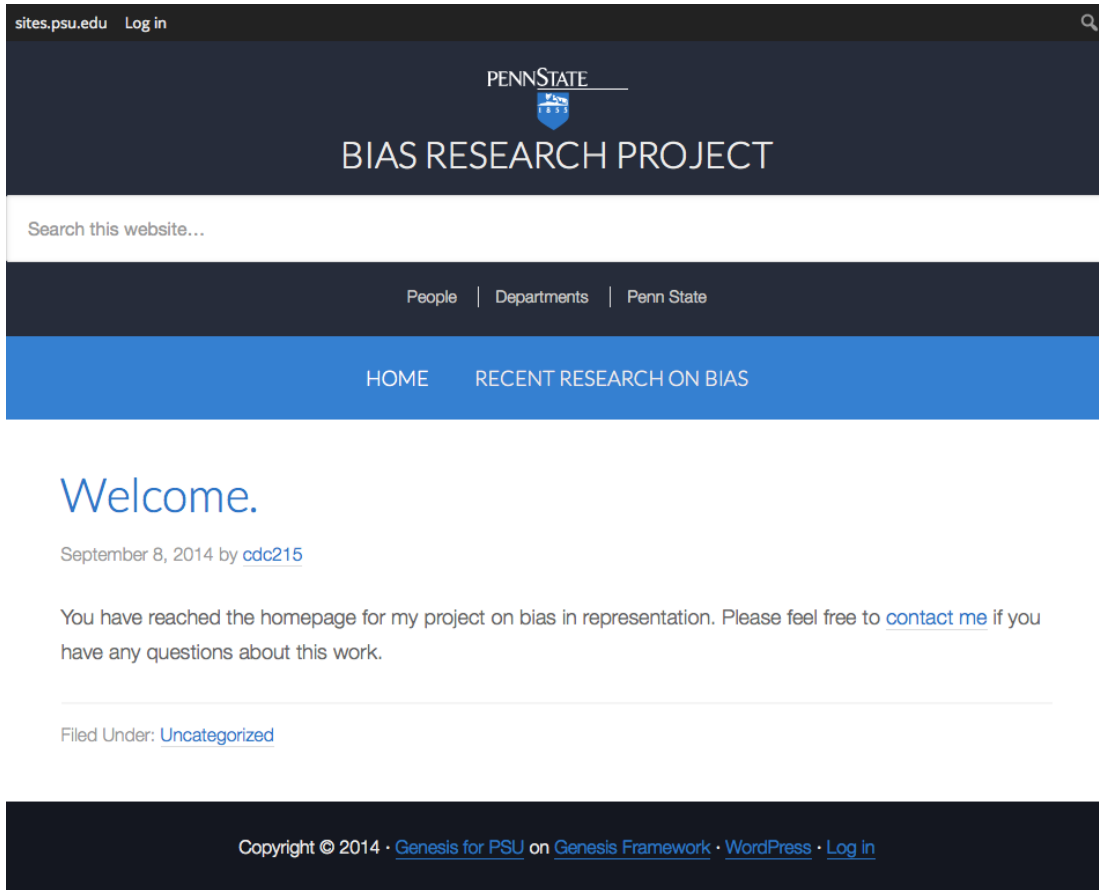
- Butler, D. M., & Broockman, D. E. (2011). [Do politicians racially discriminate against constituents? A field experiment on state legislators](#). *American Journal of Political Science*, 55(3), 463-477.

**Abstract:** We use a field experiment to investigate whether race affects how responsive state legislators are to requests for help with registering to vote. In an email sent to each legislator, we randomized whether a putatively black or white alias was used and whether the email signaled the sender's partisan preference. Overall, we find that putatively black requests receive fewer replies. We explore two potential explanations for this discrimination: strategic partisan behavior and the legislators' own race. We find that the putatively black alias continues to be differentially treated even when the emails signal partisanship, indicating that strategic considerations cannot completely explain the observed differential treatment. Further analysis reveals that white legislators of both parties exhibit similar levels of discrimination against the black alias. Minority legislators do the opposite, responding more frequently to the black alias. Implications for the study of race and politics in the United States are discussed.

Copyright © 2014 · [Genesis for PSU](#) on [Genesis Framework](#) · [WordPress](#) · [Log in](#)

Note: Identifying information has been redacted for the review.

Figure C.6: Site Screenshot - Home Page



Note: Identifying information has been redacted for the review.

# Appendix D: Survey Screenshot

Figure C.7: Survey

PENNSYLVANIA STATE UNIVERSITY  
1855

**Survey about Bias Research**

**1. Prior to today, were you aware of the research about elected officials and bias?**

Yes  
 No

If you answered 'Yes', please proceed to Question 2. If you answered 'No', please proceed to Question 3.

**2. If 'Yes', how accurate do you think the research is?**

Very Accurate  
 Accurate  
 Inaccurate  
 Very Inaccurate

**3. If 'No', please read the short summary of this research at <http://sites.psu.edu/bias/research>. Based on this summary, how accurate do you think the research is?**

Very Accurate  
 Accurate  
 Inaccurate  
 Very Inaccurate

**4. Do you share staff with other city officials?**

Yes  
 No

If you answered 'Yes', please proceed to Question 5. If you answered 'No', please proceed to Question 6.

**5. Do the shared staff read or respond to constituent correspondence?**

Yes  
 No

**6. If you would like to make any comments about either this survey, in particular, or bias research, in general, please write them here.**

**Thank you for your time!**

Done

Powered by **SurveyMonkey**  
Check out our [sample surveys](#) and create your own now!



Note: Identifying information has been redacted for the review.

## C.4. Aliases in Experiment

**Figure C.8: Black-sounding Constituent Names**

1. Alaliyah Booker
2. Alexis Banks
3. Darius Joseph
4. Darnell Banks
5. Tyreke Washington
6. DeAndre Jefferson
7. Deja Jefferson
8. Deja Mosley
9. DeShawn Korsej
10. Dominique Mosley
11. Ebony Mosley
12. Ebony Washington
13. Jada Mosley
14. Jamal Gaines
15. Jamal Rivers
16. Jasmin Jefferson
17. Jasmine Joseph
18. Jazmine Jefferson
19. Jermaine Gaines
20. Keisha Rivers
21. Kiara Jackson
22. Latonya Rivers
23. Latoya Rivers
24. LaShawn Banks
25. LaShawn Washington
26. Precious Washington
27. Rasheed Gaines
28. Raven Korsej
29. Shanice Booker
30. Terrance Booker
31. Tremayne Joseph
32. Trevon Jackson
33. Tyrone Booker
34. Tyrone Joseph
35. Xavier Jackson

**Figure C.9: White-sounding Constituent Names**

1. Allison Nelson
2. Amy Mueller
3. Anne Evans
4. Bradley Schwartz
5. Brett Clark
6. Caitlin Schneider
7. Carly Smith
8. Carrie King
9. Claire Schwartz
10. Cody Anderson
11. Cole Krueger
12. Colin Smith
13. Connor Schwartz
14. Dylan Schwartz
15. Emily Schmidt
16. Garrett Novak
17. Geoffrey Martin
18. Greg Adams
19. Hannah Phillips
20. Heather Martin
21. Holly Schroeder
22. Hunter Miller
23. Jack Evans
24. Jake Clark
25. Jay Allen
26. Jenna Anderson
27. Jill Smith
28. Katherine Adams
29. Kathryn Evans
30. Katie Novak
31. Kristen Clark
32. Logan Allen
33. Madeline Haas
34. Matthew Anderson
35. Maxwell Haas
36. Molly Kruger
37. Sarah Miller
38. Scott King
39. Tanner Smith
40. Todd Mueller
41. Wyatt Smith

## C.5. List of Questions Used in Emails

1) Local Elections I was trying to figure out the election calendar, do you know where I can find out when local elections are scheduled?

2) School Question I have a child who will be starting school soon and I'm wondering what I need to do to enroll them? Thanks for any help you can provide.

3) No subject How do I apply for a marriage license?

4) Do not call list What steps do I take to be added to the do not call list?

5) No subject Do you know who I should talk to if I want to get my name changed?

Thanks in advance for the help,

6) Bldg. Permit Where can I find out more about applying for a permit to do a building project on my home?

7) Question My nephew got a speeding ticket, what does he need to do to pay for it?  
Sincerely,

8) Voting I recently moved and am wondering how long I need to live here before I can register to vote in the next election. Do you know? Thank you,

9) Starting a business I'm looking into possibly trying to start a small business. Is there anything I need to do in the city to apply for that?

10) Community events Is there a place that lists all of the upcoming events in our community? I want to make sure I don't miss anything.

11) Council Meetings Does the city council have any regularly scheduled meetings that the public can attend? Where is the information about those meetings listed?

12) School performance Can I find out how well our schools are doing relative to other schools in the state? Is there a good website with that kind of information? Sincerely,

13) Bulk Trash I recently moved into the area and am trying to figure out what to do about bulk trash. Do you know what I should do with bulk trash? Thanks,

14) Question If I am not happy with something one of my neighbors is building, is there anything I can do about it? Best,

15) No Subject I have a complaint about a local road. Who do I speak to about that?

16) New Dog I just adopted a dog. Are there any city laws about dogs that I know about? Regards,

17) City Budget I would like to see how the city spends its money. Where could I find a copy of the budget?

18) Recycling I just moved here and I would like to know if recycling services are available. Do you know who I should talk to about that? Thanks,

19) No Subject I was just wondering where the website is for our school district in CITYNAME. Thanks!

20) Question about street sign I want to report a problem with a local street sign. Do you know who I should talk to about this? Best,

21) Zoning How do I get a lot of land re-zoned? Thanks,

22) No Subject Is there anyway to find out when a road is going to be repaired? It would just be nice to know the schedule.

23) Question In the last place I lived, I knew when the city collected leaves and cleaned streets. Is there anyway to find out whether the city offers those types of services? Sincerely,

24) No Subject Does the city keep a list of the community organizations (e.g., churches, service clubs, etc)? Sincerely,

25) Crime reports Is there a convenient place to learn about recent crime incidents in our community? I just feel that being informed is really useful.

26) Question Is the list of the city laws available somewhere online? I just wanted to learn more about how things in the city work. Thanks,

27) Local Parks? I'm planning an event and want to find a list of parks in the area. Do

you have any recommendations on where I could find out that information?

# Appendix D. Appendix for ‘How Public Opinion Shapes Discriminatory Policing’

## D.1. Quantitative Analysis of Policing Literature

In what follows, I provide a brief overview of the general contours in the politics of policing literature within political science. Specifically, I focus on the frequency with which the police have been the subject of political science research; I describe the geographic coverage of the politics of policing research; and I identify the central topics that dominate the existing literature.

**Data:** My survey of the literature draws on a large corpus of peer-reviewed publications related to policing. Specifically, I collected the population of journal articles available on JSTOR that include the terms ‘police’, ‘policing’, or ‘security agent’ from 1980 to 2018.<sup>114</sup> This corpus comprises 65,285 texts. After removing articles with incomplete metadata, I

---

<sup>114</sup>I gathered this data from JSTOR’s extremely useful, but relatively underused, Data for Research service.

am left with a cleaned corpus of 58,827 articles across 2,365 journals.<sup>115</sup> These articles obviously appear in the journals of multiple disciplines. Once I limit myself to only those articles that appear in political science journals, I am left with 14,309 articles in 95 journals.<sup>116</sup>

To address the fact that not every article that includes the terms ‘police’, ‘policing’, or ‘security agent’ is focused on the politics of policing, I remove those articles that only infrequently include phrases related to law enforcement. Specifically, I count the number of times that ‘police’ and the many derivatives of this word, such as ‘sheriff’ and ‘cop’, appear in each article. I then calculate the 90<sup>th</sup> percentile of this variable (5), and remove from the corpus any articles that have a lower count than this. This leaves me with 1,439 articles published in political science journals that make frequent mention of domestic security agents. This final corpus contains 10,689,081 words, with an average of 7,164 words per article ( $\sigma = 5,861$ ).

After identifying the corpus of interest, I transform it into a document-term matrix (DTM) for analysis. This is a data frame where the rows are documents, the columns are words, and the cell entries contain the counts of word occurrences. In doing this, I remove a set of stopwords that include the 100 most frequently used words in the English language, some additional words related to errors created in the article digitization process, and all numbers from 1 to 1,000,000. Following standard practice, I also stem words, reducing

---

<sup>115</sup>The metadata that I collected for each article includes the name of the article, the names of its authors and their institutional affiliations, the name of the journal that published it, the year in which it was published, and the included citations.

<sup>116</sup>I used those journals listed by Giles and Garand (2007) to identify political science journals. I updated the Giles and Garand list to include political science journals that have appeared since its publication: the *Journal of Experimental Political Science*, the *Journal of Global Security Studies*, *Political Science Research & Methods*, the *Quarterly Journal of Political Science*, and *Research and Politics*.



them to their base forms.

**How often do political science journals publish articles on policing?** Figure D.1 shows the number of policing articles that have been published in political science journals from 1980 to 2018. The black line refers to the articles in all political science journals, whereas the orange line refers to only those articles published in the the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics*. On average, only about 40 articles per year have been published on policing in political science. This is an incredibly low number given the importance of the police to state governance and the total number of articles published in all of political science. While the overall volume remains low, the upward sloping black line in Figure D.1 does indicate a steady increase in academic research on policing in political science.<sup>117</sup> Interestingly, the increased interest in policing in political science research is not reflected in the publications that appear in the discipline's top journals. Indeed, it is rare to see any articles on policing in these journals in a given year. Overall, the information portrayed in Figure D.1 indicates that despite its importance the police remains a peripheral topic in the political science literature.

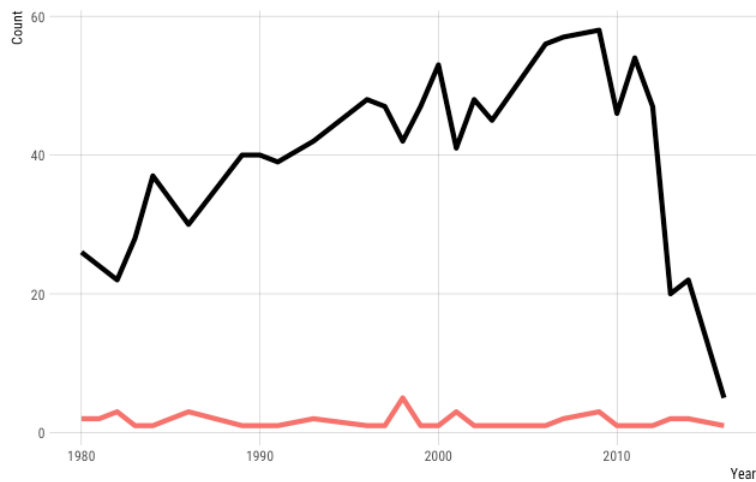
**Where have we studied policing?** Policing-related articles in political science tend to focus on only a handful of countries around the world. Figure D.2 shows the geographic coverage of political science research on policing. Countries shown in darker colors have received more attention in the political science literature.<sup>118</sup> Perhaps unsurprisingly, the

---

<sup>117</sup>While it might appear that there has been a sharp decline in published research on policing since 2012, this is likely an artifact of the embargoes used by JSTOR. In effect, JSTOR embargoes access to the articles published in some political science journals until a certain time period has elapsed. Indeed, the length of this embargo period varies across journals.

<sup>118</sup>To determine the frequency with which different countries are mentioned in the political science literature on policing, I use a dictionary-based approach in which I search my corpus for the 220 country names

**Figure D.1: Number of Policing Articles Published in Political Science Journals**



*Note:* Figure D.1 presents the number of political science articles published on policing over time. The vertical axis indicates the count of articles and the horizontal axis indicates the year. The black line represents the number of articles published in all political science journals, while the orange line represents the number of articles published in the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics*.

United States is by far the most studied country in the policing literature. Specifically, the United States is mentioned in 787 (54.7%) of the policing articles published in political science journals. The next three most frequently studied countries are China (403), India (386), and Russia (321). In contrast, the Global South receives very little attention in the policing literature. Indeed, a number of countries there and elsewhere are not mentioned at all.

On the whole, the map shown in Figure D.2 suggests that we probably know little about the politics of policing in many areas of the world. Institutional and disciplinary incentives naturally encourage researchers to examine law enforcement in the largest and most powerful (economically, militarily, or otherwise) countries. But by focusing on these countries, we are significantly limiting our understanding of police behavior and outcomes to a handful of particular contexts. This narrow scope likely impedes theoretical and conceptual development, and poses questions about the generalizability of our empirical claims.

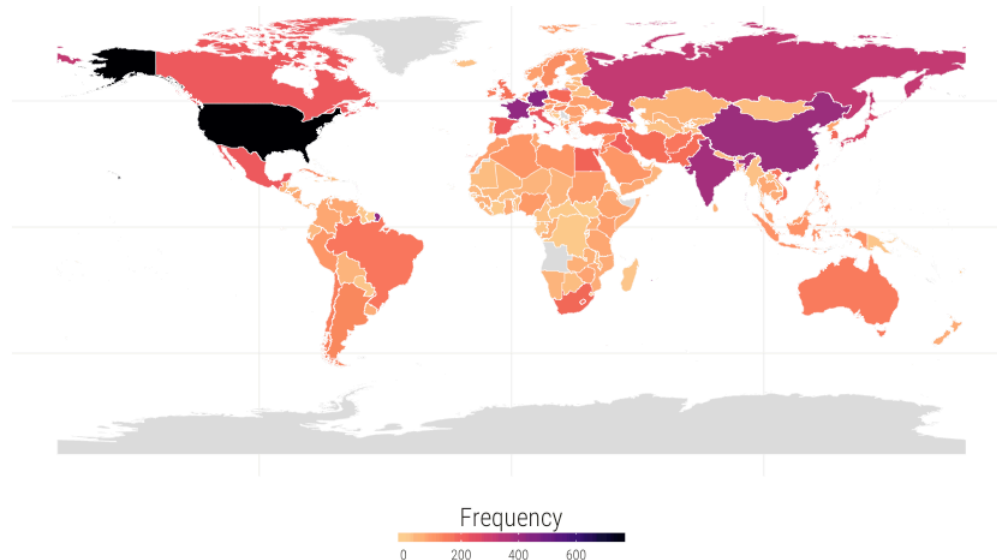
**What are the central topics in the policing literature?** To address this question in an exploratory manner, I estimate a structural topic model using a Latent Dirichlet allocation algorithm (Jockers, 2014; Blei, Ng and Jordan, 2003). This algorithm allows me to identify both the mixture of topics in my corpus of policing articles and the mixture of words in each topic. Using the approach outlined in Griffiths and Steyvers (2004), I am able to identify four distinct topics in the policing literature. To determine the substantive domain of these topics, I examine the most ‘relevant’ terms in each topic.<sup>119</sup> The top ten

---

listed by the Correlates of War Project. I then count how many articles contain each country name. This approach acknowledges that an article can focus on more than one country.

<sup>119</sup>‘Relevance’ is calculated as a weighted average of the probability that a term appears in a topic, and a term’s lift, which is the ratio of a term’s probability of appearing within a topic and its marginal probability of appearing in the corpus.

**Figure D.2: The Geographic Distribution of Political Science Research on Policing**



*Note:* Figure D.2 maps the mentions of country names in policing articles published in political science journals. Darker values represent more frequent country mentions.

most relevant stemmed terms for each topic are listed in Table D.1.

Based on the words listed in Table D.1, we can loosely define the different topics. Topic 1 appears to deal with the domestic security apparatus in communist countries, such as the Soviet Union and the German Democratic Republic. This is indicated by the fact that terms like ‘communist’, ‘union’, ‘class’, ‘soviet’, and ‘worker’ are among the most relevant for this topic. Topic 2 seems to focus on policing in China and the Middle East. Other relevant terms for this topic that are not displayed in Table D.1 include ‘iraq’ and ‘peac’. Topic 3 clusters articles on the political institutions that shape policing practices. This can be seen by the relevance of terms such as ‘court’, ‘legal’, ‘legisl’, and ‘elector’. We can perhaps think of this topic as having to do with ‘who polices the police’. Finally, Topic 4 seems to center on issues dealing with race and ethnicity and policing, particularly in the American context. We can perhaps think of this topic as having to do with the

**Table D.1: Top 10 Most Relevant Stemmed Terms by Topic**

	Topic 1	Topic 2	Topic 3	Topic 4
1.	movement	china	court	app
2.	soviet	militari	feder	black
3.	communist	chines	democraci	counti
4.	class	minist	model	review
5.	union	foreign	vote	american
6.	women	us	elector	white
7.	cultur	arab	legisl	african
8.	worker	isra	reform	inc
9.	german	presid	legal	paper
10.	war	israel	research	citi

*Note:* Table D.1 lists the 10 most relevant stemmed terms for each of the four topics identified by an LDA model.

characteristics and consequences of ‘who polices’.

## D.2. Principal Agent Description

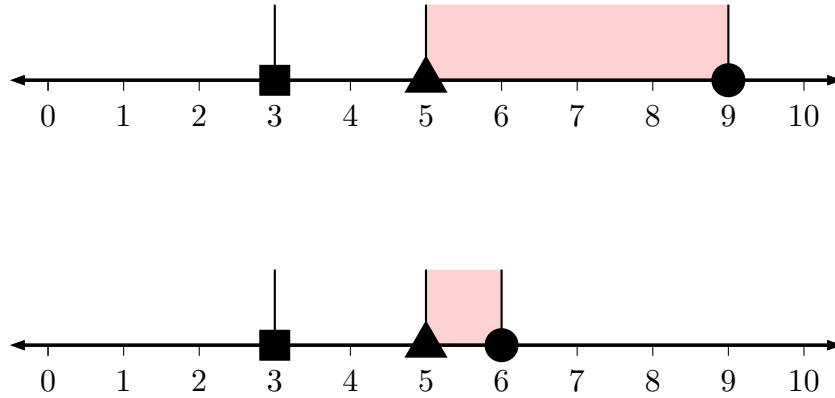
In this section, I review a basic principal agent model. To fix ideas, let us think about the principal as the median citizen in a unit of government (e.g., a city, a county, a state) — and the agent as the primary policing force for that unit.

With these actors defined, we can now think about a couple of principal-agent scenarios that differ in the extent to which the public supports racial equality in policing. Figure D.3 presents these scenarios. The horizontal axis ranges from 0 to 10 with higher values indicating increased levels of racial equality before the police and the plotted points on this line represent different policy positions. ■ denotes the status quo, the agent's ideal point, and ● the principal's ideal point. The shaded blue area marks the areas between the principal's ideal point and the delegation outcome — the agency loss.

The top plot represents a scenario in which the principal is significantly more progressive than the agent and wants considerable reform of policing practices. The status quo is 3, but the principal wants it to be 9. The agent, however, only wants it to be 5. Since the principal is 6 units of racial equality away from the status quo, they will accept any proposal nearer their point than this, including the agent's ideal point. So the agent will propose 5, the principal will accept it, and the status quo will shift to that position. In this situation, the principal gains some of the racial equality they want but still experiences an agency loss of 4. The bottom plot, on the other hand, represents a scenario in which the principal is more moderate. The principal has an ideal point of 6 and will accept any position closer to this than 3. So again the agent proposes 5 and the principal moves the status quo to that position. In this case, the agency loss is only 1.

These scenarios highlight an important point, which is that observed policing practices *on their own* cannot tell us anything about the extent to which agency loss has occurred. To

Figure D.3: Principal-Agent Scenarios



**Note:** ■ denotes the Status Quo, ▲ the agent's ideal point, and ● the principal's ideal point. The shaded blue area marks the areas between the principal's ideal point and the delegation outcome — the agency loss. The top plot represents a scenario in which the principal is significantly more progressive than the agent and there is a large agency loss. The bottom plot represents a scenario in which the principal is more moderate and there is a smaller agency loss.

measure agency loss we also need to know something about the principal's preferences. In other words, we need to consider how the preferences of the police and the public interact.

### D.3. Chain of Delegation

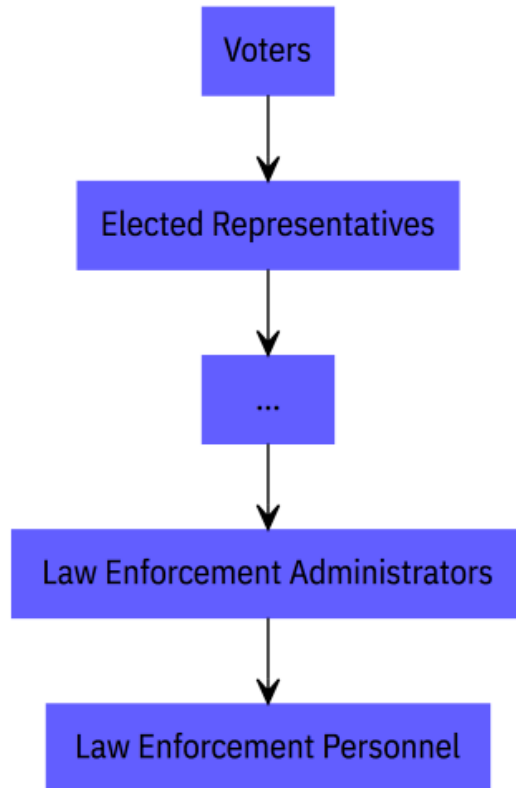
Figure D.4 presents a stylized version of the delegation chain that links these principals and agents in policing.<sup>120</sup> As seen here, voters are the ultimate principal, the first actor to delegate, elected representatives and law enforcement administrators are intermediary principals, actors who both delegate and are delegated to, and law enforcement personnel are the ultimate agent, the last actor in the chain. In this view, any inequalities in how the police use force are often described as an example of agency loss, or the difference between what the police do and what the voters would have wanted them to do (Gailmard, 2012).

---

<sup>120</sup>In some contexts, the public directly delegate to elected law enforcement administrators.



**Figure D.4: Chain of Delegation**



**Note:** Figure D.4 presents a stylized representation of the delegation chain that links voters to law enforcement personnel. ‘...’ indicates the several additional levels of delegation that separate elected officials from law enforcement administrators in some contexts.

## **D.4. Email to Law Enforcement Administrators**

In this appendix, I present the emails used to invite law enforcement administrators and elected officials to complete my survey. The University of Michigan Institutional Review Board required several small changes to the email sent to elected officials. Figure D.5 contains these messages.

Dear Law Enforcement Administrator, **Figure D.5: Email Invitations**

Hello. I'm writing today on behalf of the Annual Law Enforcement Survey (ALES) to ask if you would participate in our annual survey. The survey's purpose is to better understand law enforcement administrators like you and the issues that they face in their communities. Your answers to the questions in this survey are anonymous and will be kept confidential. For every completed survey, ALES will donate \$1 to the National Law Enforcement Officers Memorial Fund. The survey should take only 5-10 minutes to complete. Here is a link to it - (link). While we understand that you are very busy, we would greatly appreciate it if you would participate in this survey. Please feel free to contact us if you have any questions or concerns.

Thank you,

Charles Crabtree  
720.236.0778  
ccrabtr@umich.edu

Dear Elected Official,

Hello. I'm writing today to ask if you would participate in a survey. The survey's purpose is to better understand how political officials like you view law enforcement issues. Your answers to the questions in this survey will be kept confidential. For every completed survey, I will donate \$1 to the National Law Enforcement Officers Memorial Fund. The survey should take about 12 minutes to complete. Here is a link to it - (link). While I understand that you are very busy, I would greatly appreciate it if you would participate in this survey. Please feel free to contact me if you have any questions or concerns.

Thank you,

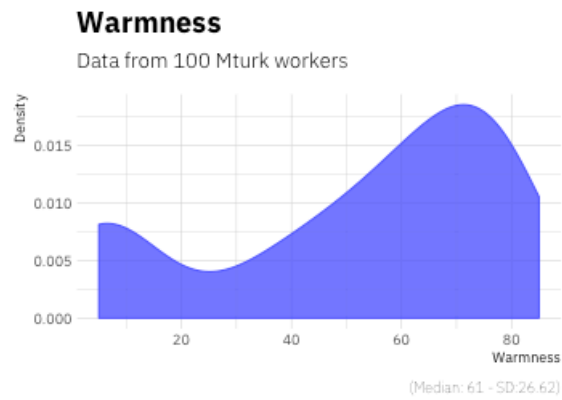
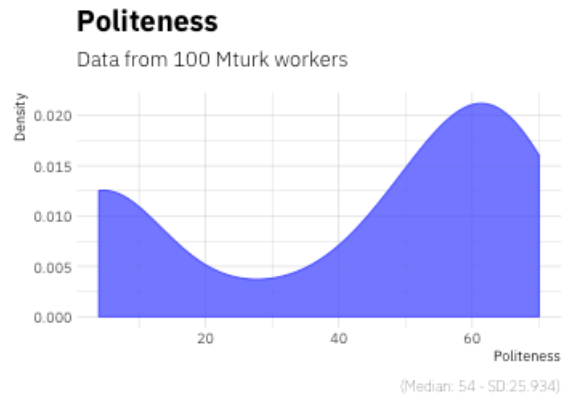
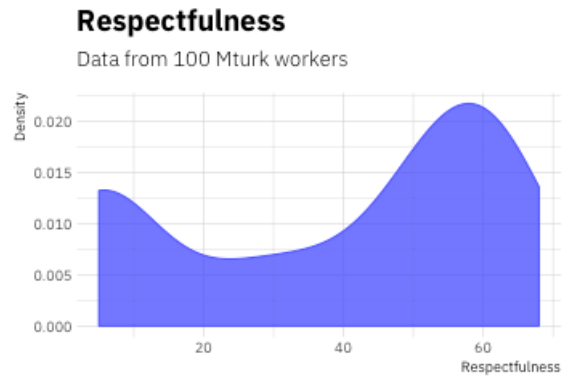
Charles Crabtree  
PhD Candidate  
University of Michigan  
720.236.0778  
ccrabtr@umich.edu

**Note:** Figure D.5 presents the email invitation sent to law enforcement administrators. See text for more details.

## D.5. Email Pre-test

Before sending the email to law enforcement administrators, I asked Amazon Mechanical Turk (MTurk) workers to rate several possible messages based on how respectful, polite, and warm they seemed to be. The text in Figure ?? earned the highest scores across these options. Figure D.6 presents kernel density plots for these scores, based on ratings from 100 MTurk workers. The plots indicate that the message was consistently viewed as respectful, polite, and warm in tone. Since the email I sent to elected officials is very similar to the one I sent law enforcement administrators, I assume that MTurk workers would have rated it similarly.

Figure D.6: Paragraph Ratings

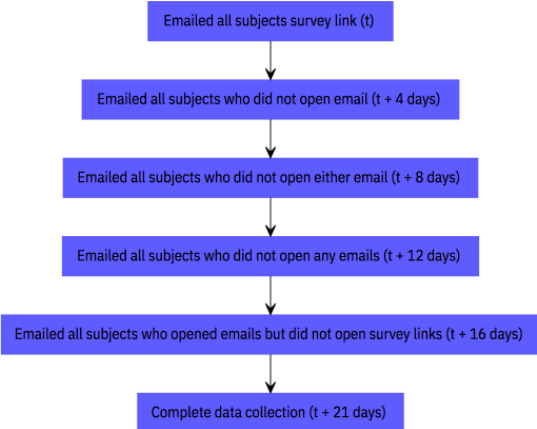


**Note:** Figure D.6 presents kernel density plots for how respectful, polite, and warm my email was considered. These data are based on ratings from 100 MTurk workers.

# D.6. Timeline for Surveys

Figure D.7 presents the timeline for each survey.

**Figure D.7: Survey Timeline**



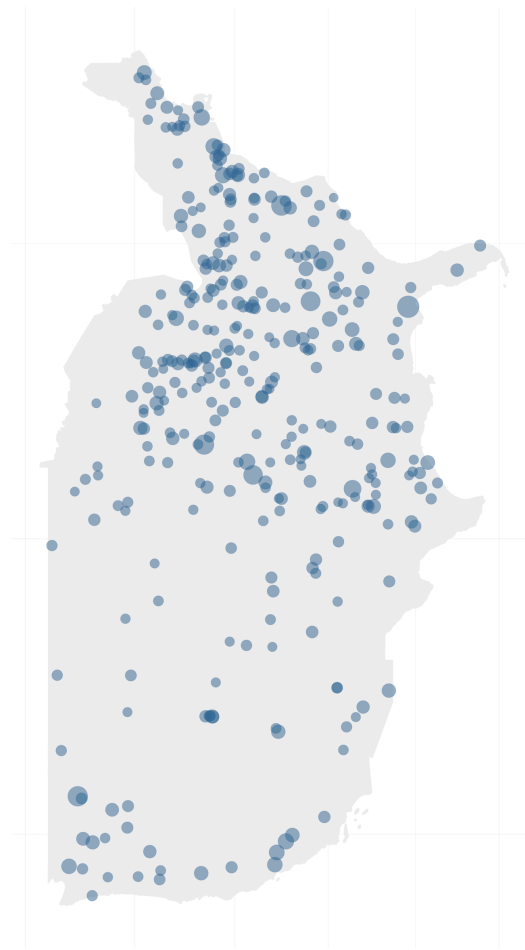
**Note:** Figure D.7 presents the timeline for each survey.

## **D.7. Location of Respondents**

In this appendix, I present a map of the law enforcement administrators who completed my survey. Each plotted point represents a respondent. Points are scaled in size by population served. Not displayed are 2 respondents from Alaska.

The Institutional Review Board prohibited me from collecting locational information for elected official respondents. They also mandated that I turned off Qualtrics' latitude-longitude tracking. As a result, I cannot create a similar map for my elected official respondents. I only know the state in which elected officials serve. The elected officials in my sample come from 47 different states.

Figure D.8: Map of Respondents



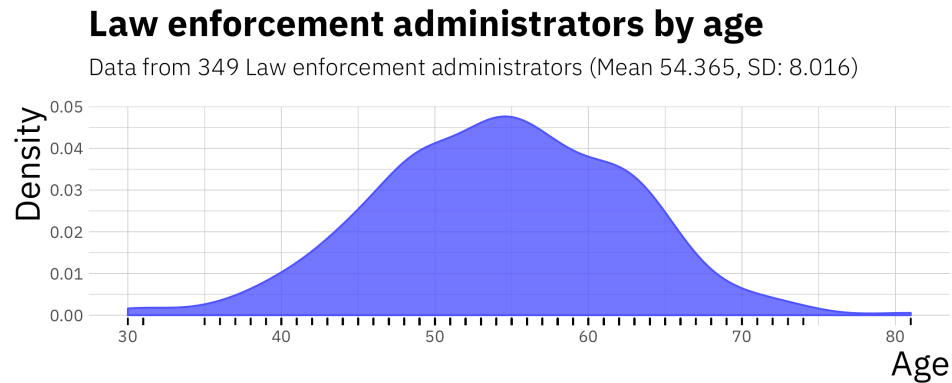
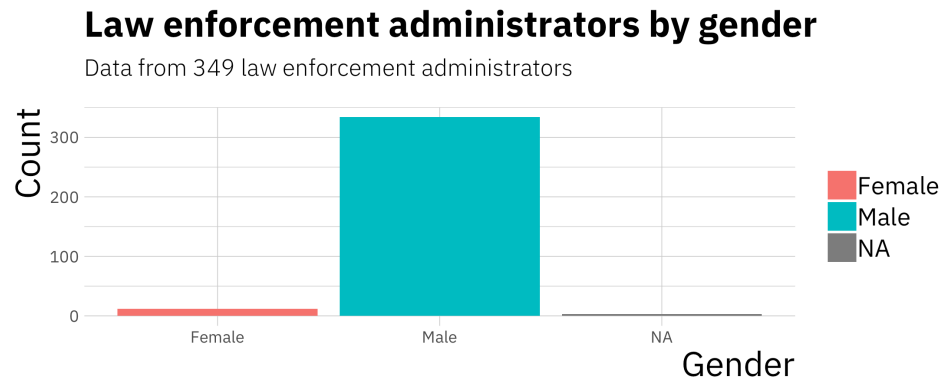
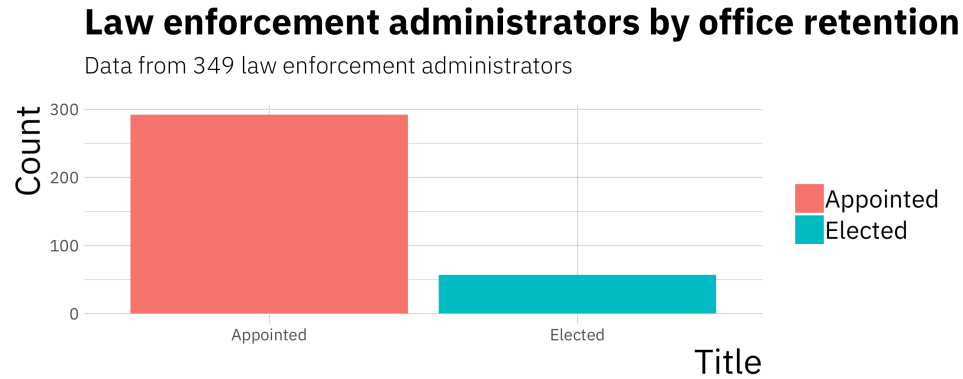
**Note:** Figure D.8 maps survey respondents. Each plotted point represents a respondent. Points are scaled in size by population served. Not displayed are 2 respondents from Alaska.



## **D.8. Law Enforcement Administrator Respondent Details**

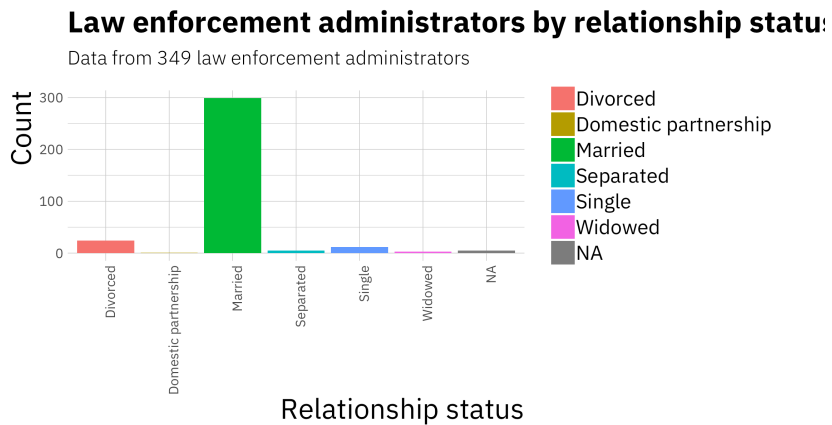
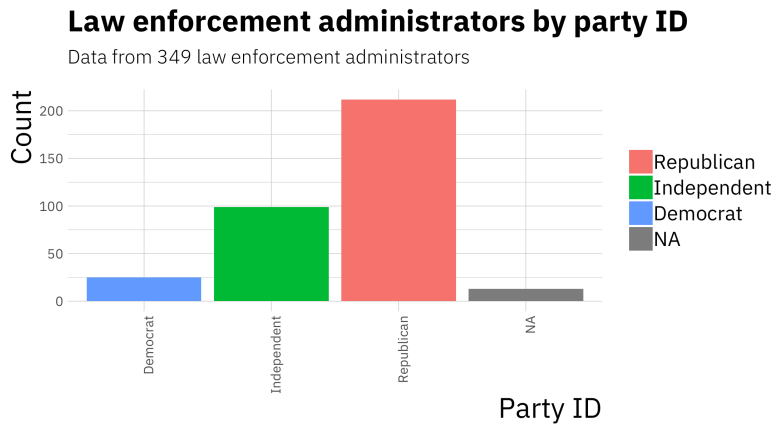
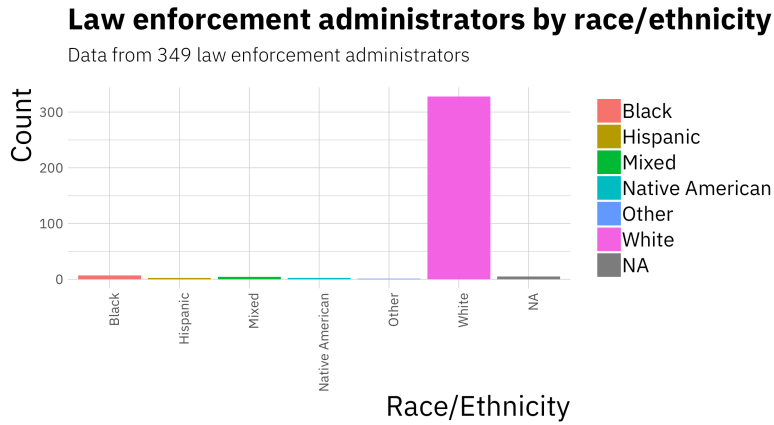
Below I provide information about the individual-level characteristics of the law enforcement administrators who responded to my survey.

Figure D.9: Law Enforcement Administrator Descriptive Statistics



*Note:* Figure D.9 presents additional information about my law enforcement administrator respondents, including descriptive statistics.

**Figure D.10: Law Enforcement Administrator Descriptive Statistics (Continued)**



*Note:* Figure D.10 presents additional information about my law enforcement administrator respondents, including descriptive statistics.

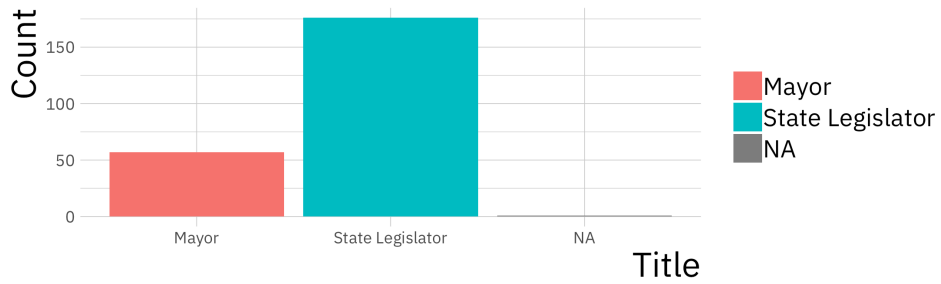
## **D.9. Elected Officials Respondent Details**

Below I provide information about the individual-level characteristics of the elected officials who responded to my survey.

Figure D.11: Elected Official Descriptive Statistics

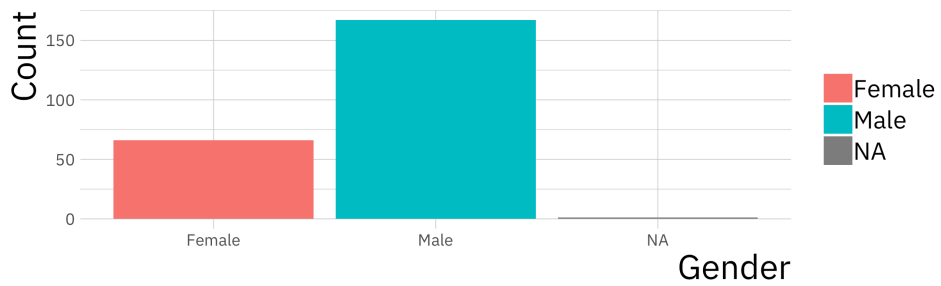
### Elected officials by title

Data from 234 elected officials



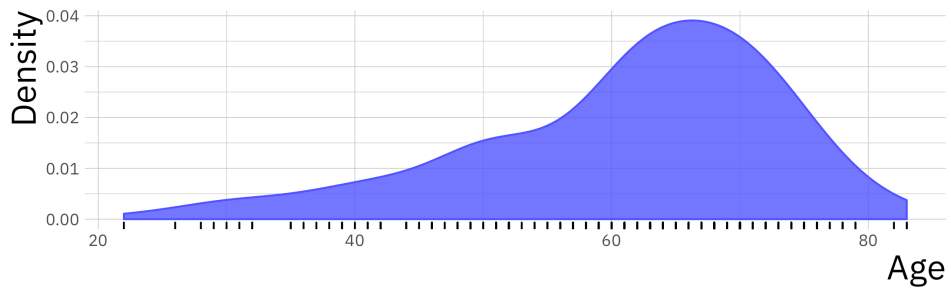
### Elected officials by gender

Data from 234 elected officials



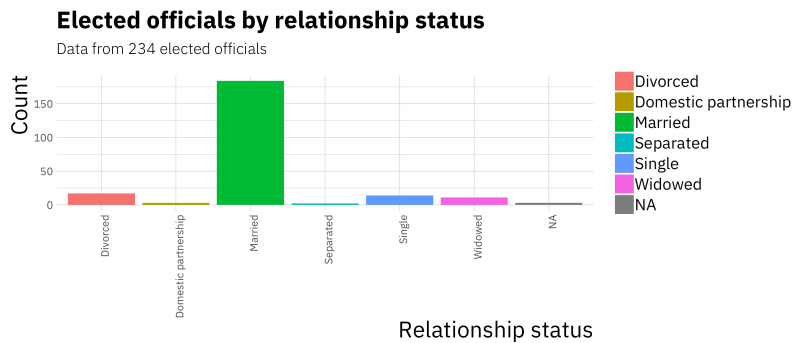
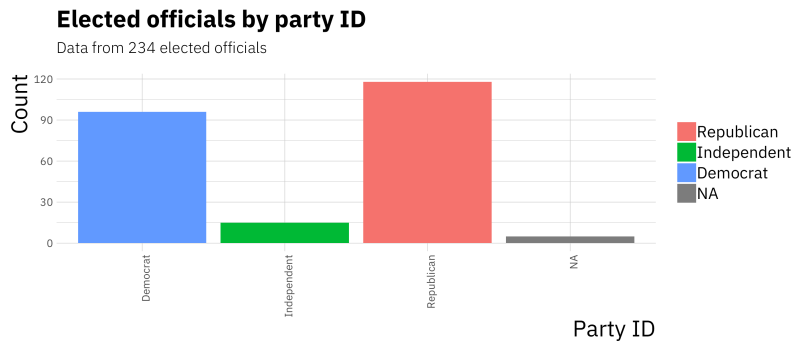
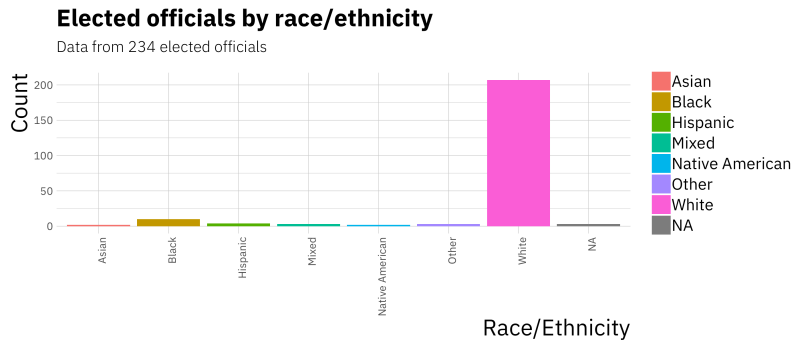
### Elected officials by age

Data from 234 elected officials (Mean 61.1, SD: 12.06)



Note: Figure D.11 presents additional information about my elected official respondents, including descriptive statistics.

**Figure D.12: Elected Official Descriptive Statistics (Continued)**



*Note:* Figure D.12 presents additional information about my elected official respondents, including descriptive statistics.

## D.10. Treatment Pretest

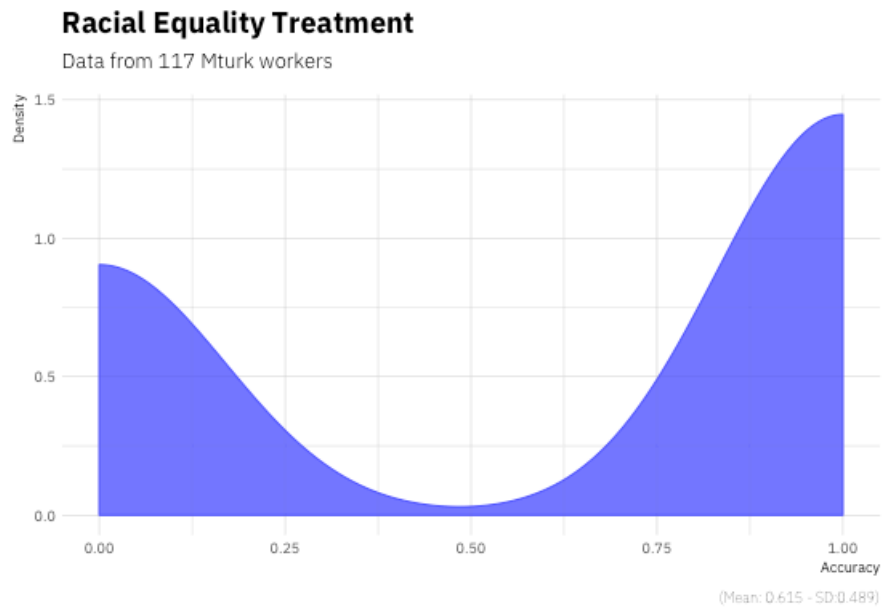
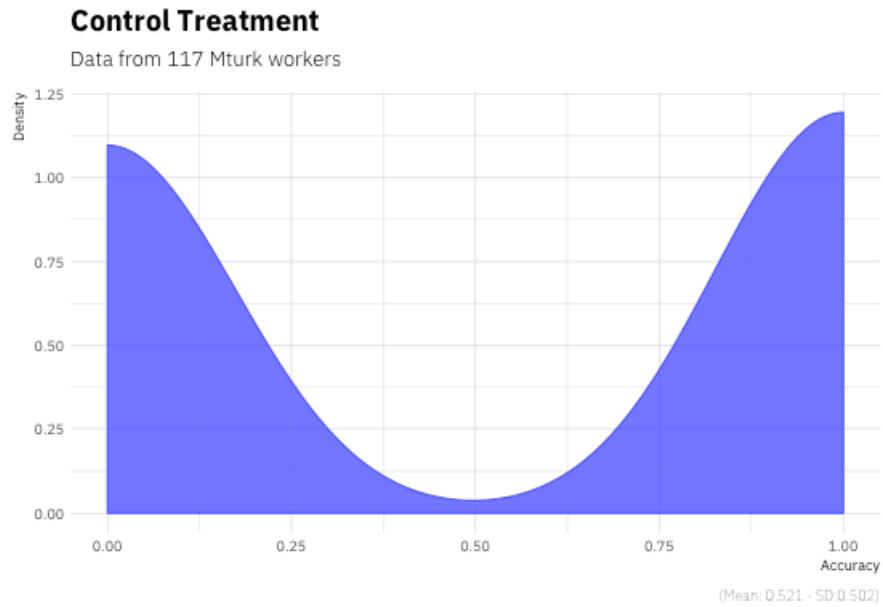
One potential concern with my EQUALITY treatment is that participants might not notice it. The issue here is that it is only 6 words — ‘Increasing racial equality in policing outcomes’ — in a paragraph of about 170. To see if this might be a problem, I checked to see if participants were more likely to remember the two concerns<sup>121</sup> listed in the control condition as opposed to the three conditions<sup>122</sup> listed in the treatment condition. I did this by asking 400 Amazon Mechanical Turk Workers (MTurkers) to complete a basic task. On one page, they read the control or treatment paragraph. 200 different workers were assigned to each condition. On the next page, they were asked the following question ‘According to the vignette that you just read, citizens care about which of the following issues? You can select more than one.’ They could select from four options: ‘crime rate’, ‘racial equality’, ‘police safety’, and ‘prison sentences’. I then created a binary measure, CORRECT, that is coded 1 if MTurkers selected only those issues presented as concerns on the prior page 0 otherwise. Figure D.13 presents the distributions for the CORRECT measure across the two non-race treatment conditions embedded in the law enforcement administrator survey experiment. We can see here that MTurkers assigned to read the EQUALITY cue exhibit higher recall of public concerns than those assigned to read the control condition.

---

<sup>121</sup>‘Lowering the crime rate’ and ‘Making police work less dangerous’.

<sup>122</sup>‘Lowering the crime rate’, ‘Increasing racial equality in policing outcomes’, and ‘Making police work less dangerous’.

**Figure D.13: Treatment Recollection**



**Note:** Figure D.13 presents the distributions for the CORRECT measure across the two non-race treatment conditions embedded in the law enforcement administrator survey experiment.



## **D.11. Public Demand for Equality Treatment**

In this section of the appendix, I present the two paragraphs of instructions in which the EQUALITY factor was embedded. Figure D.14 presents the paragraphs seen by the law enforcement administrators at the top and then the paragraphs seen by the elected officials at the bottom. Words in brackets are randomized with equal probability.

### **Figure D.14: Description of Public Interest**

#### **Text for Law Enforcement Administrators**

*In the last couple of years, the public have become increasingly interested in policing practices. For example, polling data indicate that citizens in your jurisdiction are concerned about:*

*- Lowering the crime rate - <Increasing racial equality in policing, (nothing), Increasing harsher prison sentences> Making police work less dangerous*

*To help us better understand the issues that police face, and the types of decisions that they make, we'd like to know a bit more about how you would react in some fictional situations. In what follows, we'll briefly describe four scenarios and then ask you some questions about each. Each hypothetical situation occurs in a different place and with different people. Since it is important to understand how law enforcement officers like you might react to these scenarios, we'd appreciate it if you would take a moment to reflect on each scenario before answering any question(s).*

*As a reminder, I will keep your replies anonymous and will never share with anyone how any respondent answers these questions.*

#### **Text for Elected Officials**

*In the last couple of years, the public have become increasingly interested in policing practices. Polling data indicate that citizens in your constituency are concerned about:*

*- Lowering the crime rate - <Increasing racial equality in policing, (nothing)> Making police work less dangerous*

*To help us better understand what elected officials think about the issues that police face, and the types of decisions that they make, we'd like to know a bit more about how you would react in some fictional situations. In what follows, we'll briefly describe four scenarios and then ask you some questions about each. Each hypothetical situation occurs in a different place and with different people. Since it is important to understand how elect officials like you might react to these scenarios, we'd appreciate it if you would take a moment to reflect on each scenario before answering any question(s).*

*Once more, we'd like to remind you that your answers to these questions will be kept anonymous.*

## D.12. Name Selection

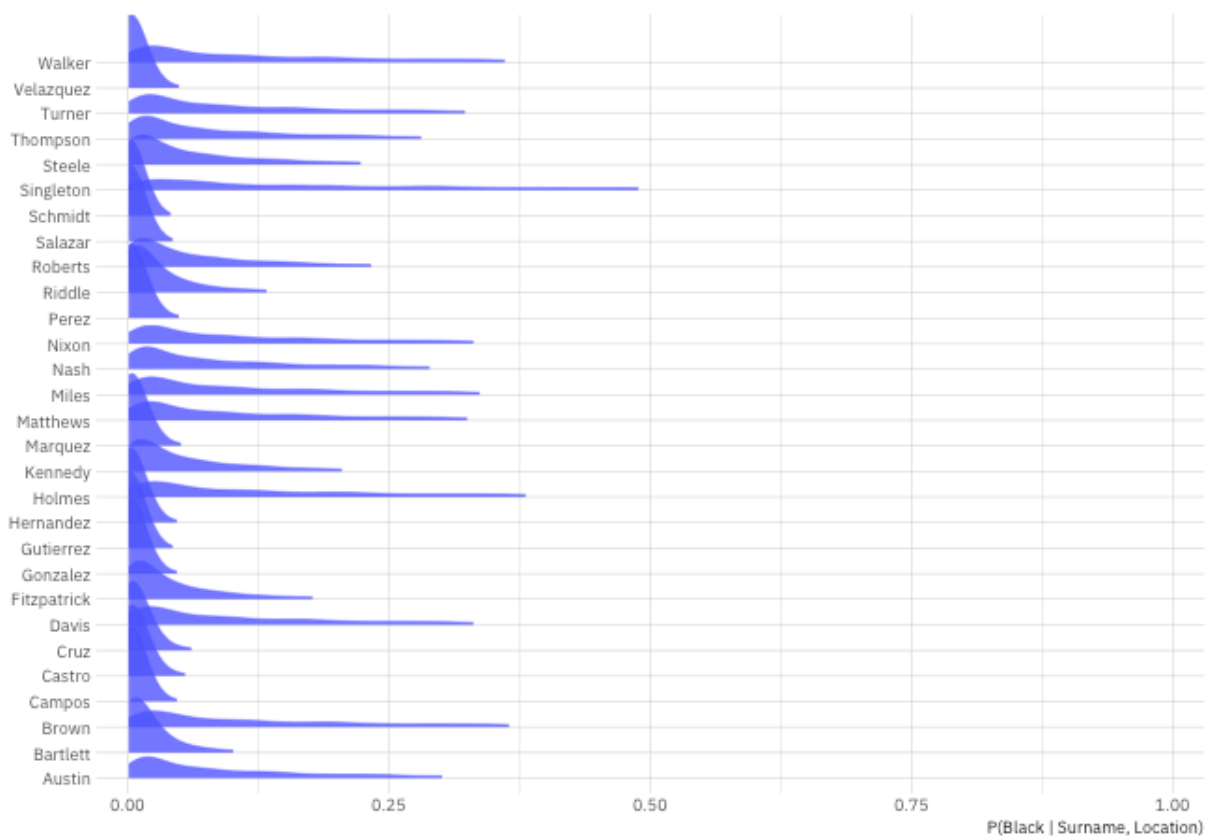
In this appendix, I describe my process for selecting the names of putative citizens described in the survey experiment vignettes. As a starting point, I used data from Hughes et al. (2017) on racial perceptions of Black, Latino, and White names. Hughes et al. (2017) had workers on Amazon’s Mechanical Turk service (MTurkers) estimate their confidence that an individual with a given name belonged to a specific racial or ethnic group. They had MTurkers rate 100 different names for four different group identities: Arabs, Blacks, Latinos, and Whites. I selected the 10 names that MTurkers were most certain belong to Black, Latino, and White individuals as an initial list of potential identities.

As Hughes et al. (2017) show, though, MTurkers were less certain about classifying names as belonging to Black individuals. One possible reason for this is that racial perceptions of some last names typically used by Black individuals can vary across geography (Crabtree and Chykina, 2018). The general idea here is “last names can signal different races in different places” (Crabtree and Chykina, 2018, 23). To assess whether this was a problem here, I use Bayes’ rule to generate the predicted probability of a name signaling a Black identity across all 3,142 United States counties. Figure D.15 displays the distribution of these probabilities across these counties. The probabilities of a name belonging to a race sum to 1.

If a last name consistently denotes a Black identity across the United States, we would expect the distribution of probabilities for that name to be tightly center on 1. Similarly, if a surname does not consistently signal a Black identity across geographic contexts, we would expect the distribution of probabilities to be closely grouped around 0. We can see here that racial perceptions of the Black surnames listed in Table ?? might vary across space.

To strengthen racial perceptions of Black names, I collected from the 2000 United States

Figure D.15: Initial Black Last Names



**Note:** Figure D.15 presents the probability that a last name belongs to a Black individual given local demographics.

census a list of the ten most common last names among Blacks.<sup>123</sup> Table D.2 lists the surnames, the percentage of individuals with this last name that are black, and the percentage of individuals with this last name that are white.

I then take these names and calculate the probability that they belong to a Black individual given local demographics. Figure D.16 presents these results. It shows that the last names that most consistently signal a Black identity are ‘Washington’, ‘Jefferson’, and

<sup>123</sup>These data and a description of them can be found at [goo.gl/UF2axb](http://goo.gl/UF2axb).

**Table D.2: Surnames and Occurrence Across Racial Groups**

Surname	% Black	% White
Washington	89.87%	5.16%
Jefferson	75.24%	18.72%
Booker	65.57%	30.09%
Banks	54.24%	41.3%
Jackson	53.02%	41.93%
Mosley	52.83%	42.69%
Dorsey	51.81%	43.97%
Gaines	50.27%	45.12%
Rivers	50.21%	42.48%
Joseph	48.84%	35.48%

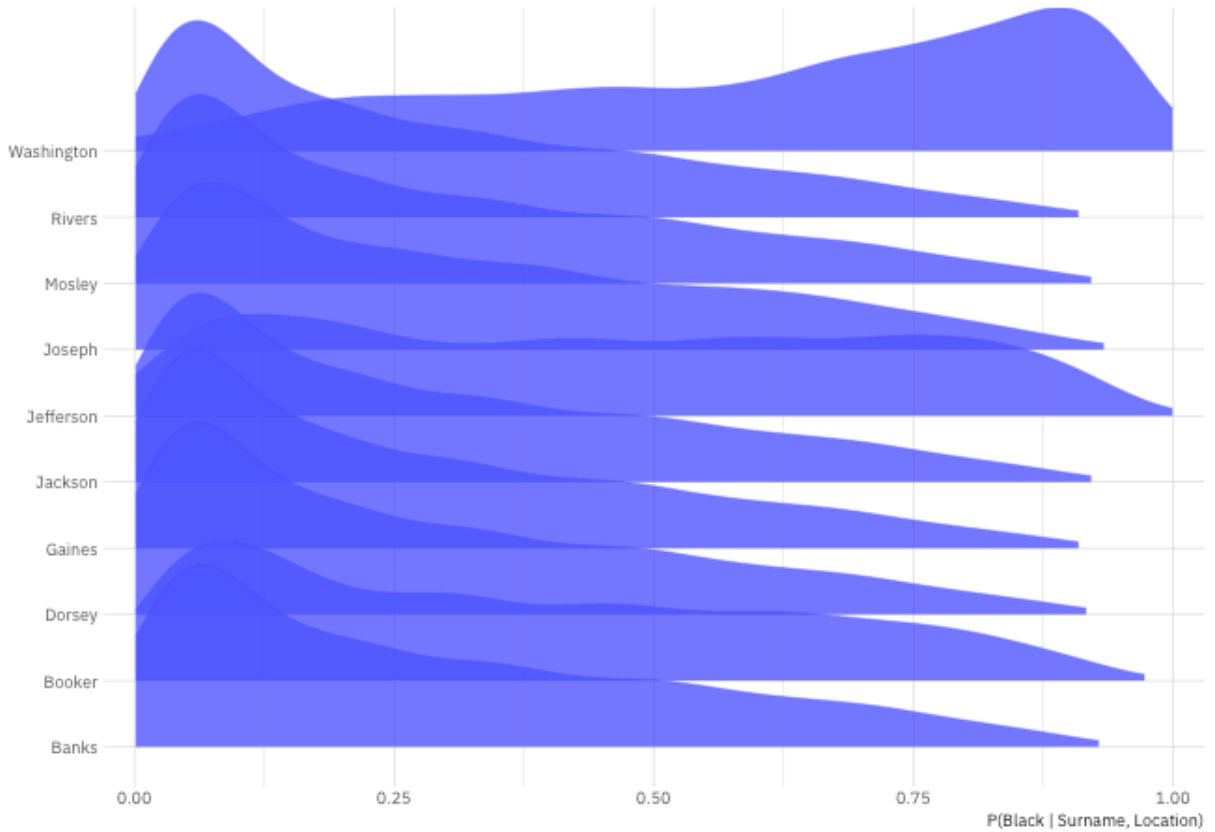
‘Booker’. The mean probabilities that these names denote a Black person across geographical units are 0.62, 0.45, and 0.37, respectively.

Next, I randomly assign one of these three last names to the ten Black first names taken from Hughes et al. (2017). These new names are listed in Table D.3, along with the 20 names used to indicate Latino and White individuals. I then ask 200 MTurkers to assess the likelihood that an individual with one of these 30 names is Black, Latino, or White individual. Figure D.17, Figure D.18, and Figure D.19 present the distribution of these likelihoods for the names selected to denote Black, Latino, and White individuals, respectively.

**Table D.3: New Names and Racial Conditions**

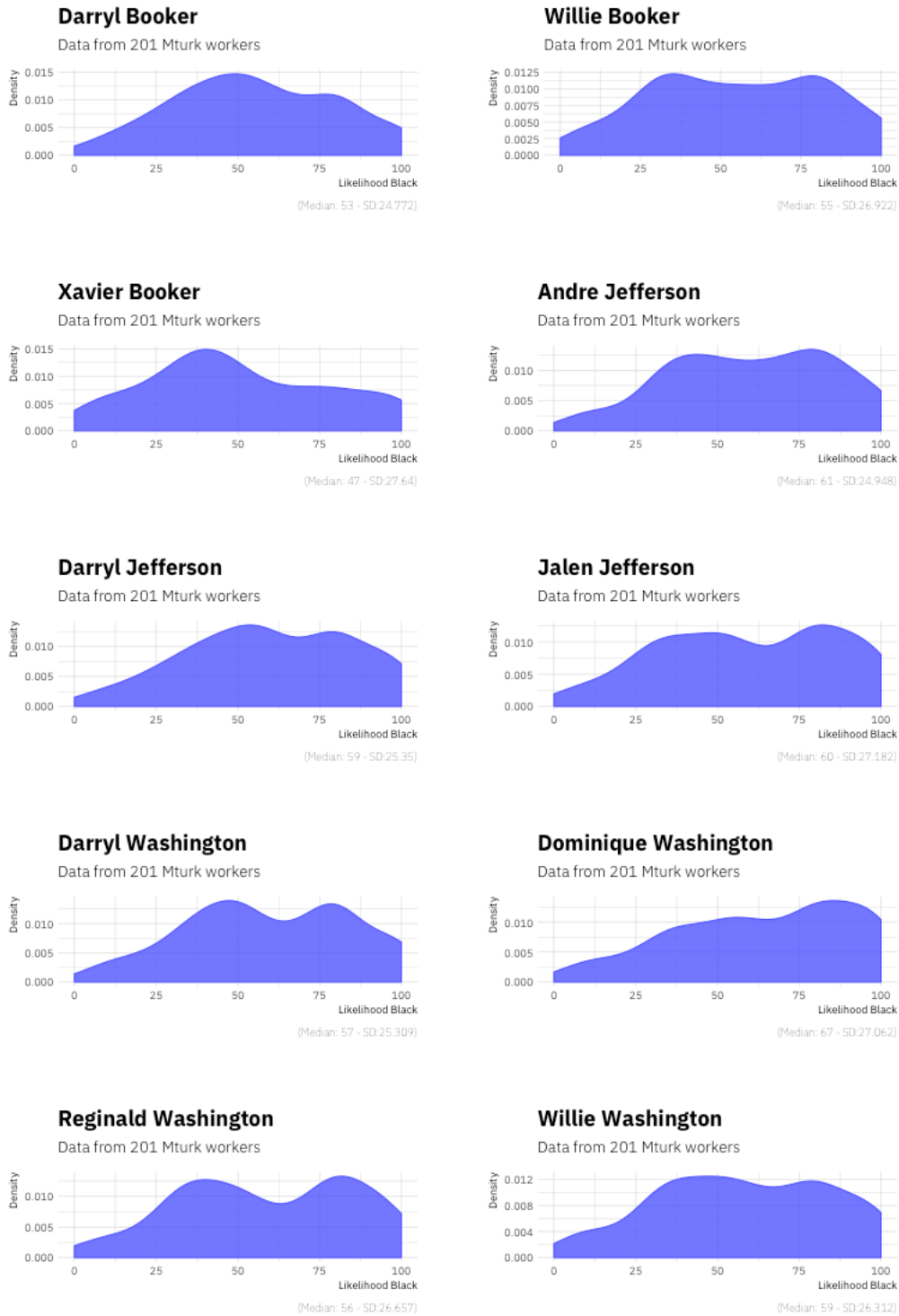
Name	Racial condition
Dominique Washington	Black
Andre Jefferson	Black
Xavier Booker	Black
Darryl Washington	Black
Darryl Jefferson	Black
Willie Booker	Black
Reginald Washington	Black
Jalen Jefferson	Black
Darryl Booker	Black
Willie Washington	Black
Jose Cruz	Latino
Jorge Castro	Latino
Cesar Marquez	Latino
Jose Gutierrez	Latino
Juan Campos	Latino
Saul Gonzalez	Latino
Miguel Salazar	Latino
Jesus Perez	Latino
Diego Velazquez	Latino
Fernando Hernandez	Latino
Daniel Nash	White
Matthew Roberts	White
Alex Steele	White
Nicholas Austin	White
Zachary Fitzpatrick	White
Christopher Schmidt	White
Ryan Thompson	White
Timothy Bartlett	White
Corey Kennedy	White
Garrett Riddle	White

Figure D.16: Black Last Names



**Note:** Figure D.17 presents the probability that last name belongs to a Black individual given local demographics.

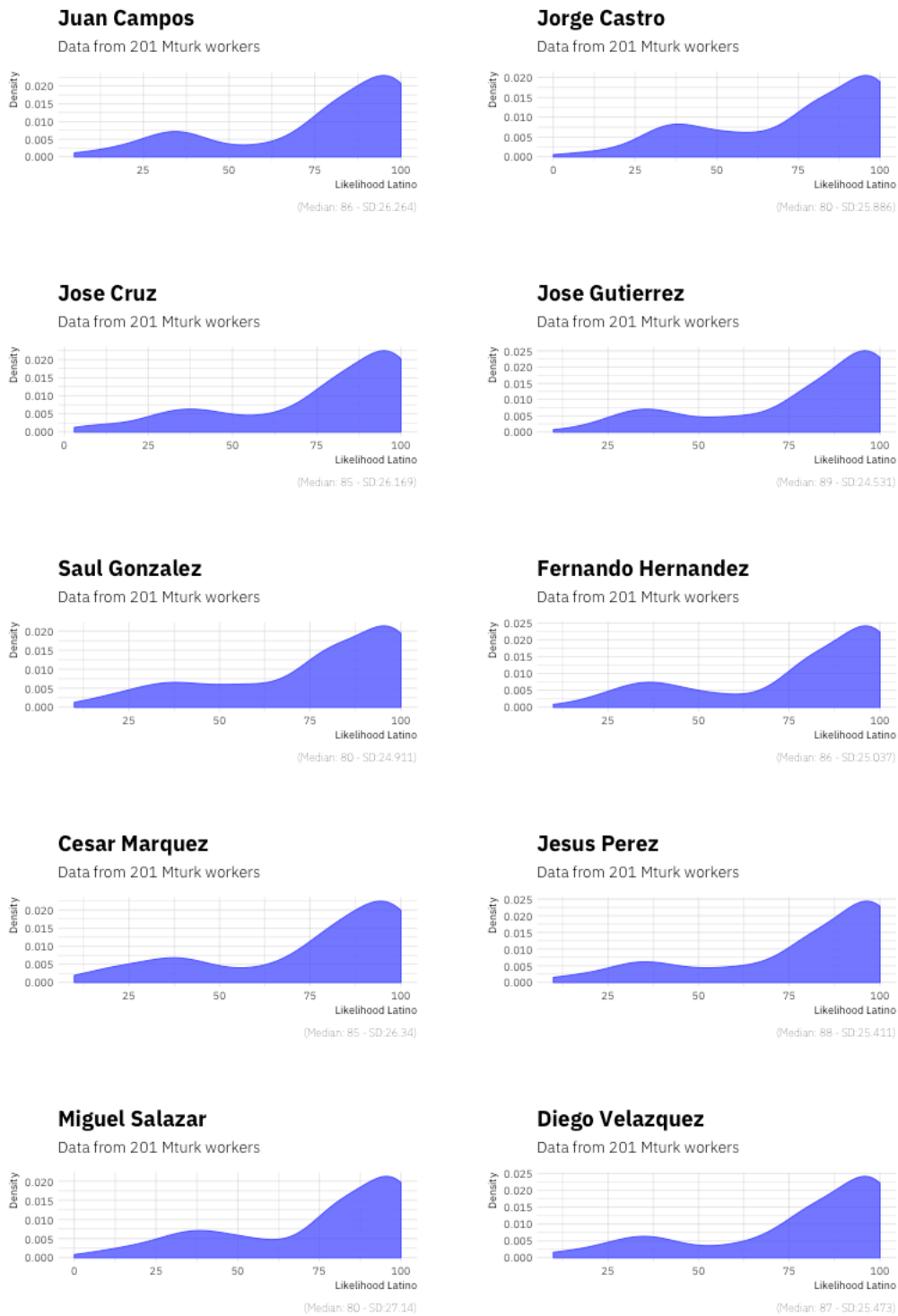
Figure D.17: Black Names



**Note:** Figure D.17 presents the distribution of MTurker likelihoods that the names here denote a Black identity.

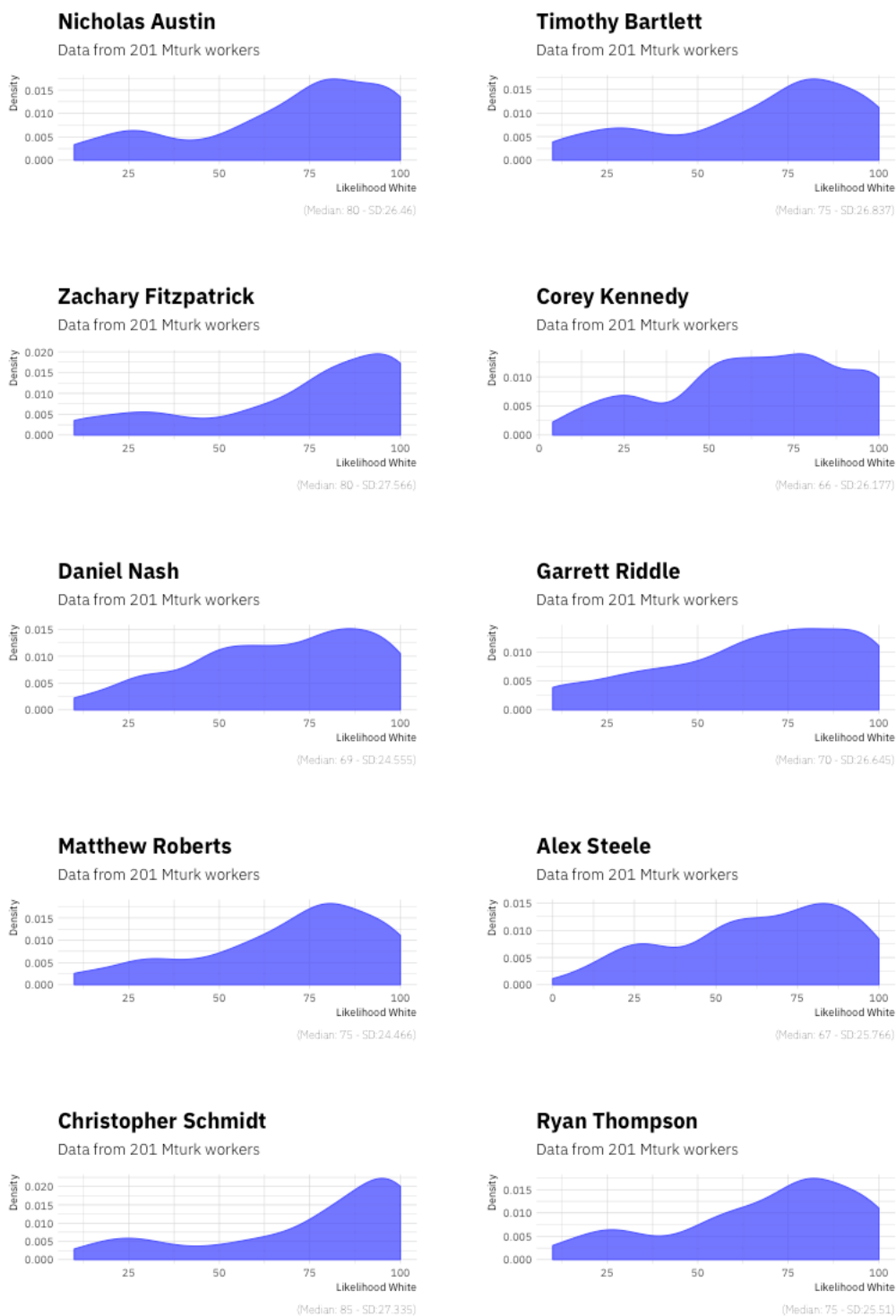


**Figure D.18: Latino Names**



**Note:** Figure D.18 presents the distribution of MTurker likelihoods that the names displayed here denote a Latino identity.

Figure D.19: White Names



**Note:** Figure D.19 presents the distribution of MTurker likelihoods that the names displayed here denote a White identity.

## D.13. Elected Official Vignettes

In this appendix, I present the vignettes and response questions presented to elected officials. Words in brackets were randomly assigned, as was the order in which the vignettes appeared.

[**Use of Force vignette**] *You learn about the following incident in your constituency. Two police officers on foot patrol surprise a 6'2" <Black / Latino / White> man with a large build who appears to be breaking into an apartment. The man flees, shouting at the officers. The officers chase after him for a while, eventually succeeding in tackling him to the ground. After he is under control, both officers punch him a couple of times as punishment for fleeing and resisting.*

**How likely would you be on a scale of 1 to 10 (where 10 is most likely) to investigate this incident?**

[**Traffic Stop vignette**] *A police officer pulls over a silver 2010 Ford Taurus for 40mph in a 30mph zone. When the officer approaches the vehicle, he notices that the driver seems very concerned. The man identifies himself as <Black / Latino / White name> and explains that he is rushing to pick up his child from school. It's mid-afternoon, the school he mentions is nearby, and the driver is very apologetic. The officer decided to issue a ticket.*

**How likely would you be on a scale of 1 to 10 (where 10 is most likely) to approve of this action?**

## D.14. Surveys

In this appendix, I provide links to the survey instruments presented to law enforcement administrators and elected officials.

- Elected Officials Survey - [charlescrabtree.com/market/elected\\_officials\\_survey.pdf](http://charlescrabtree.com/market/elected_officials_survey.pdf)
- Law Enforcement Survey - [charlescrabtree.com/market/law\\_enforcement\\_survey.pdf](http://charlescrabtree.com/market/law_enforcement_survey.pdf)

One noticeable difference across these surveys is that the survey for law enforcement administrators uses the logo of the Annual Law Enforcement Survey and makes several mentions to this organization. The original intention was to also deliver the elected officials survey under the banner of ALES, but a change in the University of Michigan's Institutional Review Board guidelines prohibited this. The IRB required that I use UM branding in the elected officials survey.

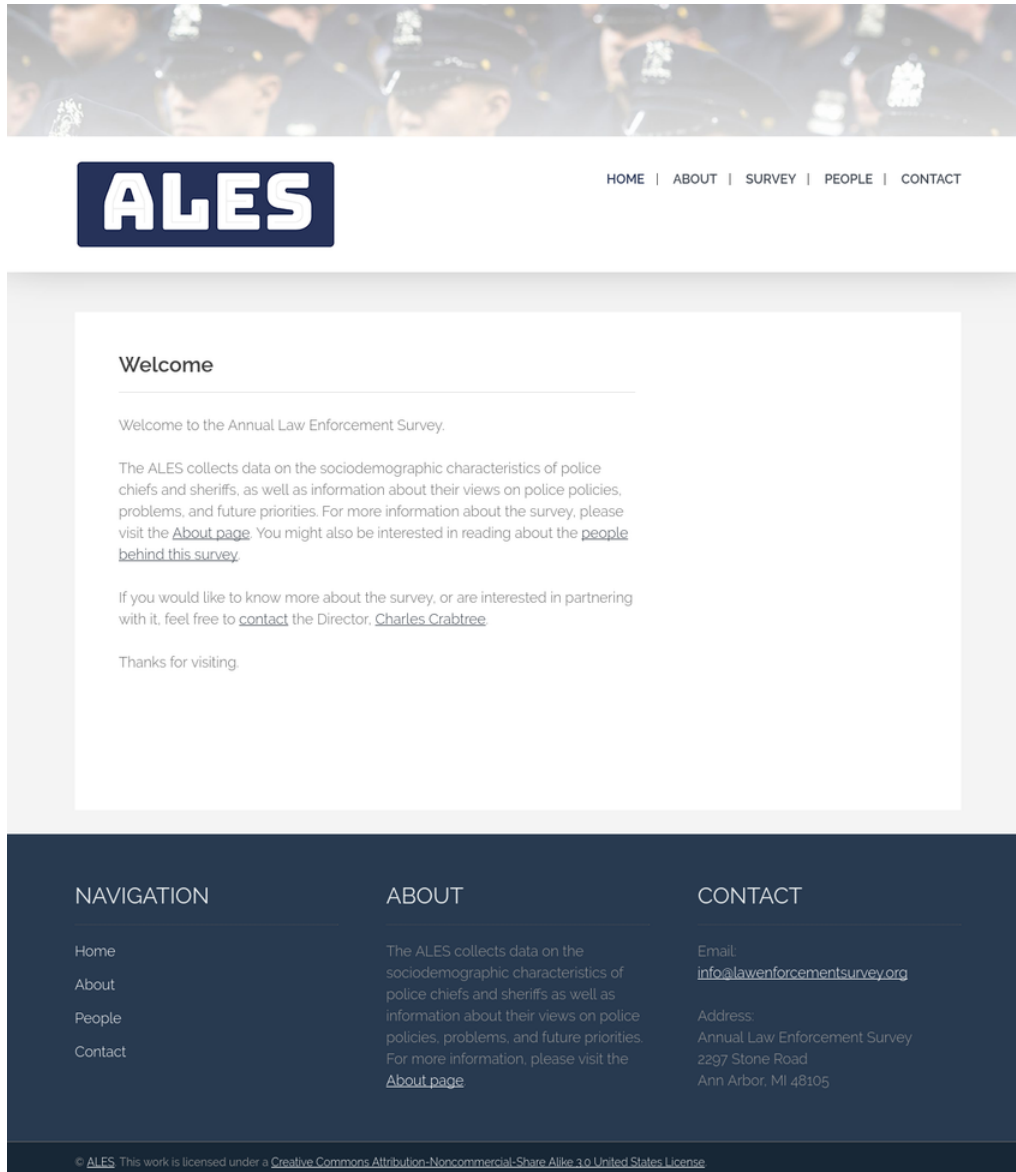
One concern might be that the different logos might induce sponsorship effects. Crabtree, Kern and Pietryka (2018) find, however, that the university banners often included in surveys do not influence respondent answers or effort. While they do not compare university banners to banners of non-government organizations, such as ALES, their results suggest that the different response patterns I observe across surveys are not driven by sponsorship effects.

### D.14.1. ALES Website

In this sub-appendix, I present information about the Annual Law Enforcement Survey (ALES) website, <http://lawenforcementsurvey.org>. Figures D.20, D.21, D.22, D.23, and D.24 contain screenshots of the website's various pages. Before inviting law enforcement administrators to participate in the survey, I asked MTurk workers to evaluate the site. I

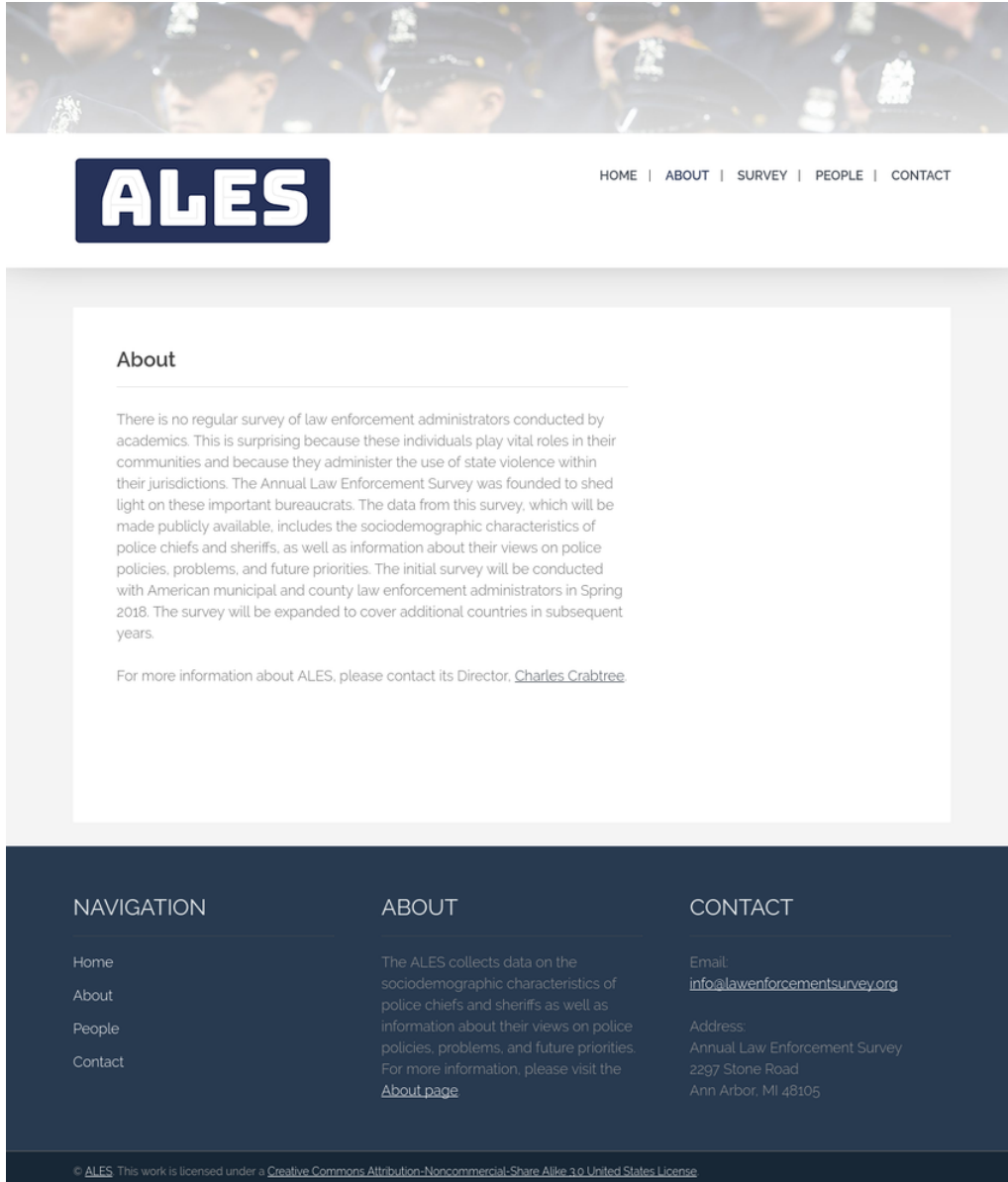
did this in two ways. First, I asked 151 workers to rate the site based on how respectful, polite, and professional it seemed. Figure D.25 presents the data from their evaluations. Second, I asked 600 MTurk workers to describe their impression of the site and its purpose. The open-ended responses suggest that workers viewed the site positively and as belonging to an organization interested in improving policing practices.

Figure D.20: ALES Homepage



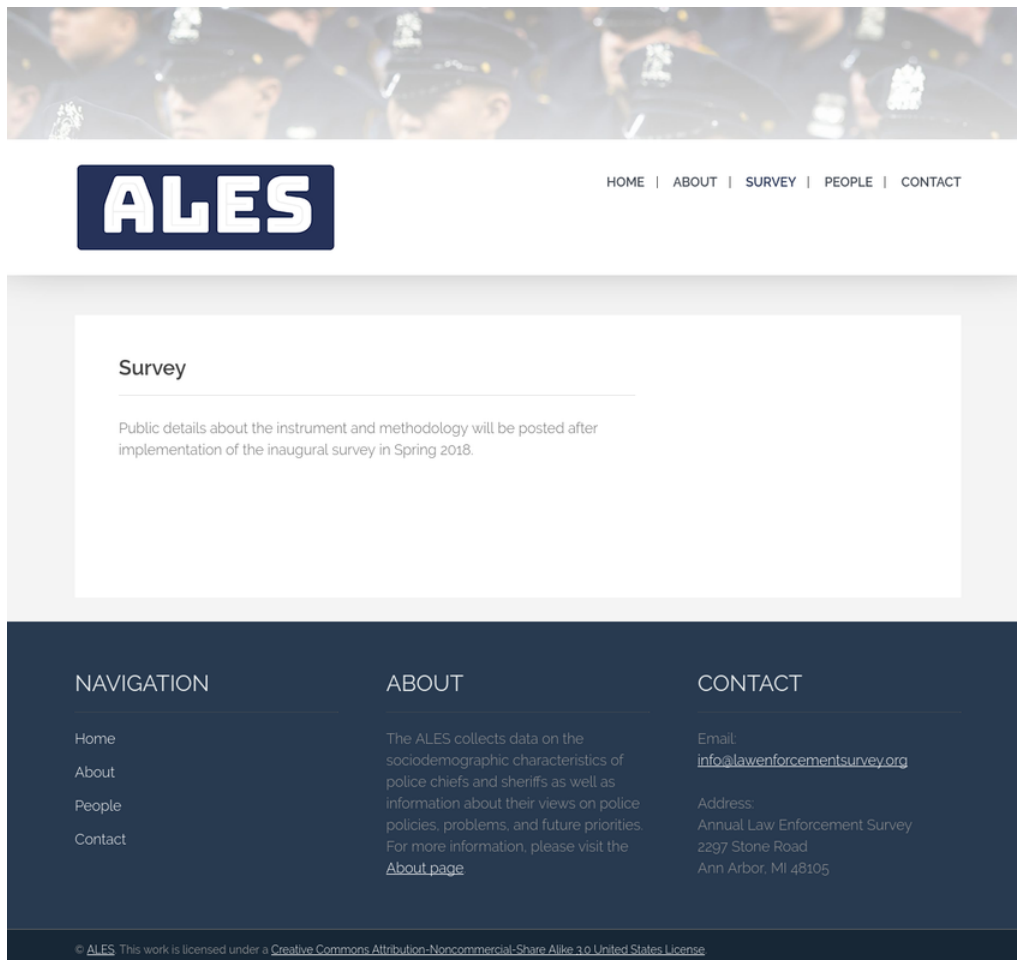
**Note:** Figure D.20 presents the homepage for <http://lawenforcement.org>.

Figure D.21: ALES About Page



**Note:** Figure D.21 presents the about page for <http://lawenforcement.org>.


Figure D.22: ALES Survey Page



**Note:** Figure D.22 presents the survey page for <http://lawenforcement.org>.




Figure D.23: ALES People Page



HOME | ABOUT | SURVEY | PEOPLE | CONTACT


## People

### Director




Charles Crabtree is a PhD Candidate in the [Department of Political Science](#) at the [University of Michigan](#). His research focuses on various aspects of repression and discrimination in comparative, American, and international politics. Methodologically, he is interested in research design, experiments, and measurement. He has [published work](#) on these topics in journals such as the *British Journal of Political Science*, *International Studies Quarterly*, *Journal of Peace Research*, *Political Research Quarterly*, *Political Analysis*. He has received funding for this research from the [Making Electoral Democracy Work](#) project and the [Swedish Research Council](#) among other sources. More information can be found at his [website](#) and on his [Google scholar profile](#).


### Board of Directors



Christian Davenport is Professor in the [Department of Political Science](#) at the [University of Michigan](#) and Research Professor at the [Peace Research Institute Oslo](#). His research focuses on political conflict and violence. In addition to publishing articles in journals such as the *American Journal of Political Science*, the *American Political Science Review*, the *American Sociological Review*, the *Annual Review of Political Science*, the *Journal of Politics*, and *International Studies Quarterly*, he has also published seven books. His most recent book, *The Peace Continuum: What It Is and How to Study It*, was published by Oxford University Press. His research has been funded by the [Carnegie Foundation](#), the [National Science Foundation](#), and the [Research Council of Norway](#), among other sources. More information can be found at his [website](#) and on his [Google Scholar profile](#).



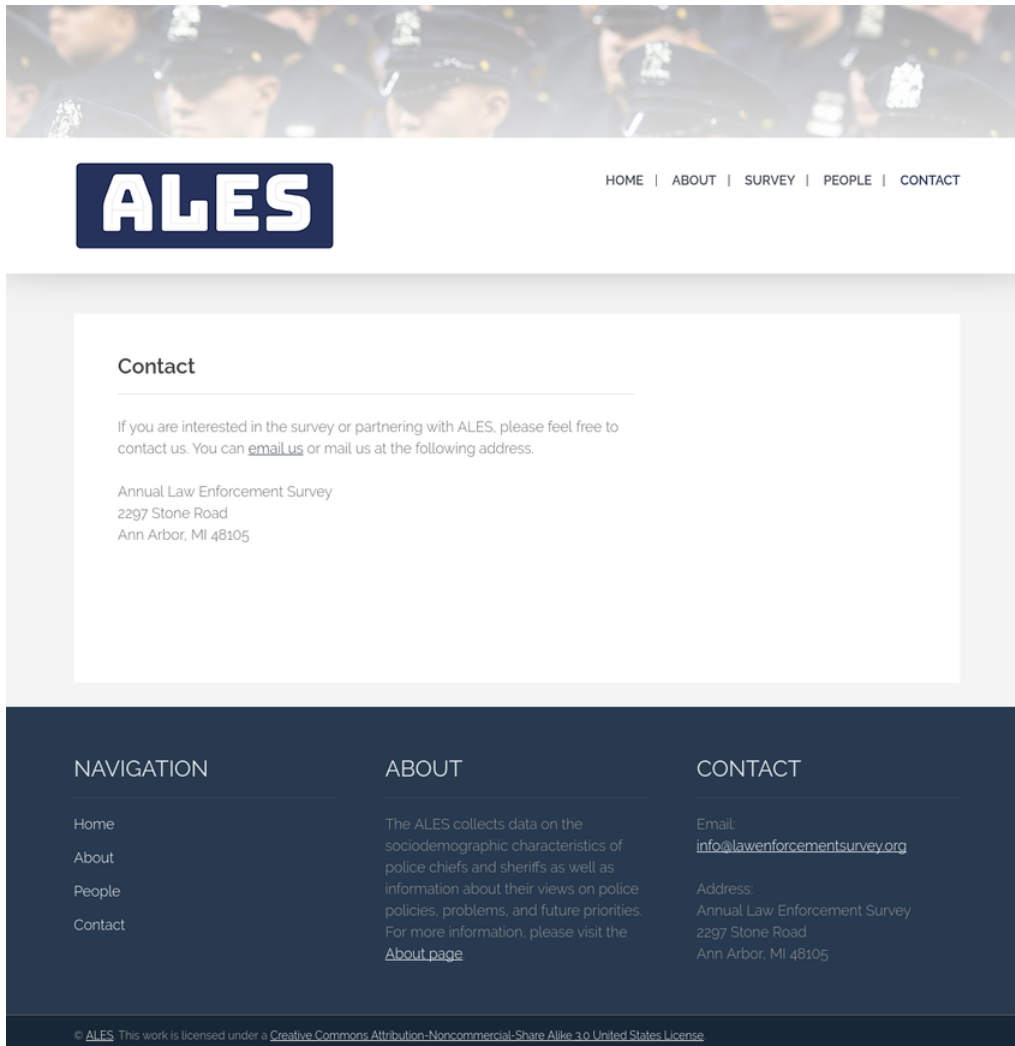
Cassy Dorff is an Assistant Professor in the [Department of Political Science](#) at the [University of New Mexico](#). Her research explores how civilians survive and respond to political violence and political crises. Her current work is focused on civilian victimization in Mexico and the networked nature of civil conflicts across the world. She has published articles in journals such as the *American Journal of Political Science*, the *Journal of Peace Research*, *International Interactions*, *International Studies Review*, *Political Science Research & Methods*, and *Research & Politics*. Her research has been funded by the International Center on Nonviolent Conflict. More information can be found at [her website](#) and on her [Google Scholar profile](#).



Kristine Eck is an Associate Professor in the [Department of Peace and Conflict Research](#) at [Uppsala University](#) and Director of the [Uppsala Conflict Data Program](#). Her research focuses on violence against civilians, conflict dynamics, and rebel recruitment. She has published articles in journals such as *Cooperation and Conflict*, *Human Rights Quarterly*, *International Studies Quarterly*, the *Journal of Conflict Resolution*, the *Journal of Peace Research*, and *Security Studies*. Her research has been funded by the [East Asian Peace Program](#), the [Swedish Research Council](#), and the Norwegian Foreign Ministry. More information can be found on her [Google Scholar profile](#).

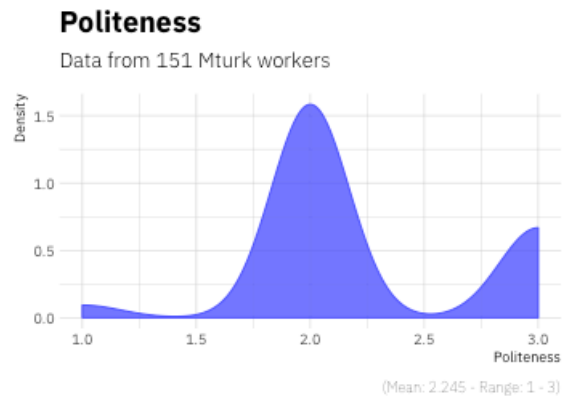
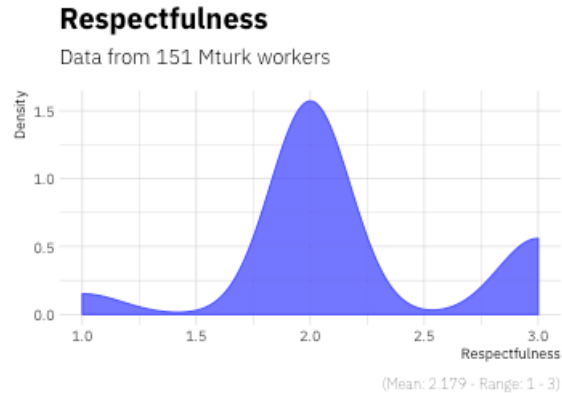
Note: Figure D.23 presents the people page for <http://lawenforcement.org>.

Figure D.24: ALES Contact Page



**Note:** Figure D.24 presents the contact page for <http://lawenforcement.org>.

**Figure D.25: Site Evaluations**



**Note:** Figure D.25 presents the data from 151 MTurk worker evaluations of <http://lawenforcement.org>.

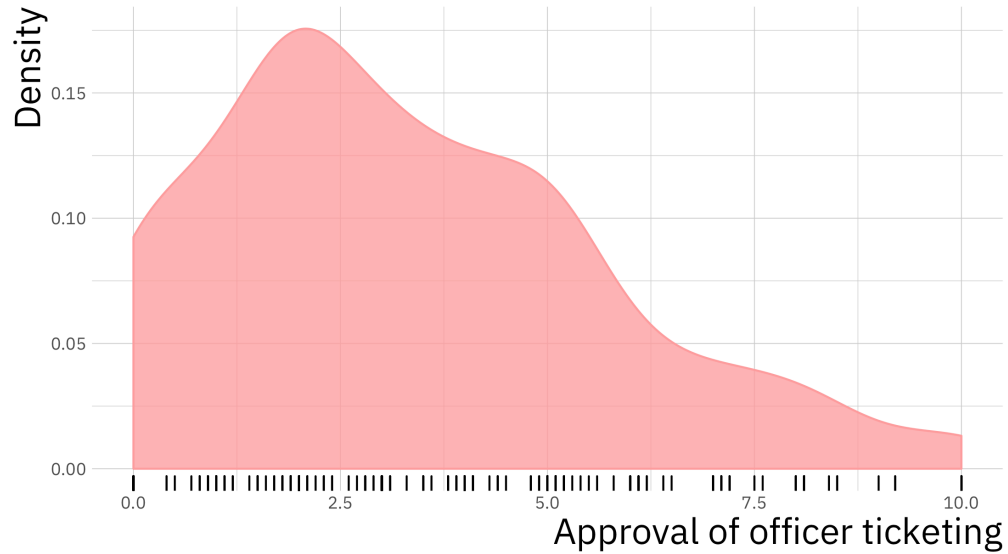
## **D.15. Dependent Variables**

In this appendix, I present the dependent variables used in my models. Figure D.26 presents the values of BEATING or TICKETING for law enforcement administrators. Figure D.27 presents the values of BEATING or TICKETING for elected officials.

Figure D.26: Dependent Variables for Law Enforcement Administrators

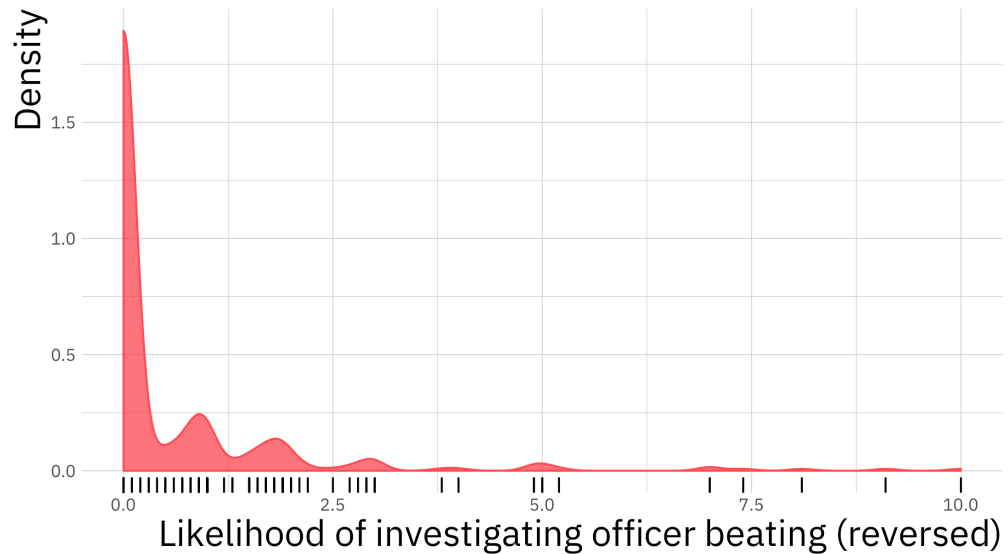
### Approval of officer ticketing

Data from 349 law enforcement administrators (Mean 3.327, SD: 2.387)



### Likelihood of investigating officer beating (reversed)

Data from 349 law enforcement administrators (Mean 0.583, SD: 1.361)

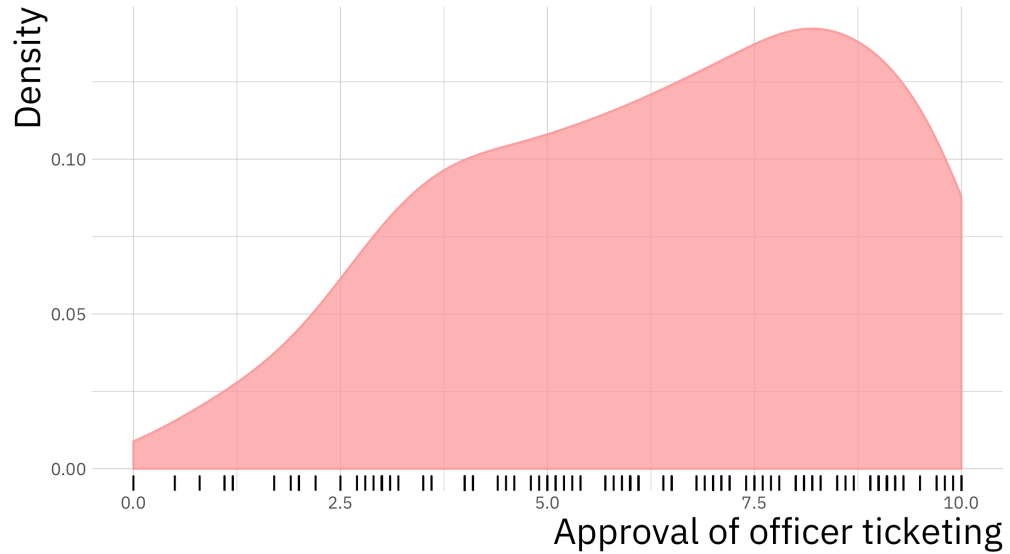


**Note:** Figure D.26 presents the values of BEATING or TICKETING for law enforcement administrators.

Figure D.27: Dependent Variables for Elected Officials

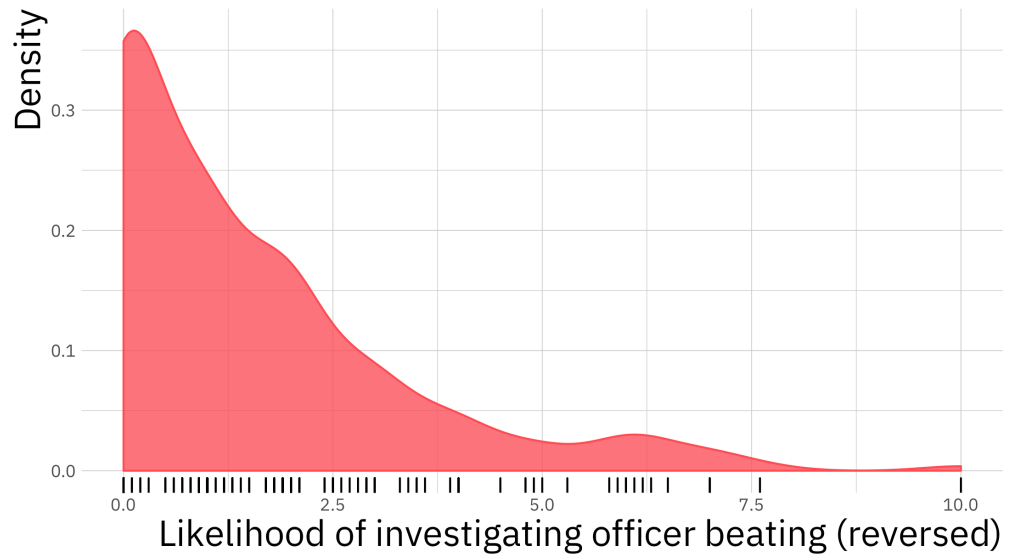
### Approval of officer ticketing

Data from 234 elected officials (Mean 6.383, SD: 2.482)



### Likelihood of investigating officer beating (reversed)

Data from 234 elected officials (Mean 1.498, SD: 1.81)



**Note:** Figure D.27 presents the values of BEATING or TICKETING for elected officials.

## D.16. Results - Elected Officials

Table D.4 presents the results from an OLS model that uses BEATING as the dependent variable (left column) and another that uses TICKETING as the dependent variable (right column). Cell entries contain coefficient estimates and HC2 robust standard errors in parentheses.

**Table D.4: OLS Results - Elected Officials**

	BEATING	TICKETING
BLACK	0.82* (0.46)	-0.27 (0.57)
HISPANIC	0.49* (0.29)	0.12 (0.57)
BLACK × EQUALITY	-1.11** (0.61)	-0.43 (0.77)
HISPANIC × EQUALITY	-1.24** (0.47)	-1.08 (0.83)
EQUALITY	0.91** (0.30)	0.68 (0.57)
N	233	233

\*p < .1; \*\*p < .05; \*\*\*p < .01

**Note:** Table D.4 presents results from two OLS models. Cells contain estimated coefficients. HC2 robust standard errors are in parentheses. The models contain pre-treatment covariates that capture whether the respondent was a state official, whether they are a female, whether they are married, whether they self-identify as a Democrat, and their level of education. Data come from 57 elected mayors and 176 elected state legislators. See text for more details about the data and model.

## D.17. Results - Law Enforcement Administrators

Table D.5 presents the results from a series of OLS models. The first and second columns presents the results from models estimated with data from appointed law enforcement administrators, while the third and fourth columns present the results from models estimated with data from elected law enforcement administrators. The first and third columns presents results from models that use BEATING as the dependent variable, while the second and fourth columns present results from models that use TICKETING as the dependent variable. Cell entries contain coefficient estimates and HC2 robust standard errors in parentheses.

**Table D.5: OLS Results - Law Enforcement Administrators**

	BEATING	TICKETING	BEATING	TICKETING
BLACK	-0.14 (0.23)	-0.56 (0.41)	0.05 (0.42)	-1.09 (1.25)
HISPANIC	0.13 (0.27)	-0.75* (0.40)	-0.20 (0.40)	-0.27 (1.28)
BLACK × EQUALITY	0.39 (0.43)	0.37 (0.69)	-0.37 (0.83)	0.74 (1.70)
HISPANIC × EQUALITY	0.01 (0.38)	0.61 (0.74)	0.75 (0.89)	-0.16 (1.76)
EQUALITY	-0.15 (0.28)	-0.58 (0.51)	0.12 (0.70)	-0.61 (1.15)
N	292	292	57	57

\*p < .1; \*\*p < .05; \*\*\*p < .01

**Note:** Table D.5 presents results from four OLS models. Cells contain estimated coefficients. HC2 robust standard errors are in parentheses. Data come from 292 appointed law enforcement administrators and 57 elected law enforcement administrators. See text for more details about the data and model.



# Bibliography

- Abadie, Alberto and Guido W Imbens. 2006. “Large sample properties of matching estimators for average treatment effects.” *econometrica* 74(1):235–267.
- Abrajano, Marisa A. and Michael M. Alvarez. 2010. *New Faces, New Voices: The Hispanic Electorate in America*. Princeton University Press.
- Adida, Claire L, David D Laitin and Marie-Anne Valfort. 2010. “Identifying barriers to Muslim integration in France.” *Proceedings of the National Academy of Sciences* 107(52):22384–22390.
- Adida, Claire L, David D Laitin and Marie-Anne Valfort. 2014. “Muslims in France: identifying a discriminatory equilibrium.” *Journal of Population Economics* 27(4):1039–1086.
- Ahmed, Ali, Lina Andersson and Mats Hammarstedt. 2013a. “Sexual orientation and full-time monthly earnings, by public and private sector: Evidence from Swedish register data.” *Review of Economics of the Household* 11(1):83–108.
- Ahmed, Ali M, Lina Andersson and Mats Hammarstedt. 2012. “Does age matter for employability? A field experiment on ageism in the Swedish labour market.” *Applied Economics Letters* 19(4):403–406.
- Ahmed, Ali M, Lina Andersson and Mats Hammarstedt. 2013b. “Are gay men and lesbians discriminated against in the hiring process?” *Southern Economic Journal* 79(3):565–585.
- Akerlof, George A. 1978. The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in Economics*. Elsevier pp. 235–251.
- Al-Sayed, Abdulrahman M., Diane S. Lauderdale and Sandro Galea. 2010. “Validation of an Arab Names Algorithm in the Determination of Arab Ancestry for Use in Health Research.” *Ethnicity & Health* 15(6):639–647.
- Alexander, Kern and M David Alexander. 2011. *American public school law*. Cengage Learning.

- Alizade, Jeyhun, Rafaela M Dancygier and Ruth Dittmann. 2018. “National Policies, Local Politics, and Citizenship Acquisition: Field Experiments with Elected Officials in Germany.”
- Arrow, Kenneth. 1972. Some mathematical models of race discrimination in the labor market. In *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal. New York: Lexington Books pp. 187–203.
- Ayres, Ian and Peter Siegelman. 1995. “Race and gender discrimination in bargaining for a new car.” *The American Economic Review* pp. 304–321.
- Baert, Stijn. 2016. “Wage subsidies and hiring chances for the disabled: some causal evidence.” *The European Journal of Health Economics* 17(1):71–86.
- Baert, Stijn and Sunčica Vujić. 2016. “Immigrant volunteering: a way out of labour market discrimination?” *Economics Letters* 146:95–98.
- Baker, Joseph O. 2015. *American secularism: Cultural contours of nonreligious belief systems*. NYU Press.
- Balko, Radley. 2013. *Rise of the warrior cop: The militarization of America’s police forces*. PublicAffairs.
- Baumgartner, Frank R, Derek A Epp and Kelsey Shoub. 2018. *Suspect Citizens: What 20 Million Traffic Stops Tell Us about Policing and Race*. Cambridge University Press.
- Baumgartner, Frank R, Derek A Epp, Kelsey Shoub and Bayard Love. 2017. “Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002–2013.” *Politics, Groups, and Identities* 5(1):107–131.
- Baumgartner, Frank R, Leah Christiani, Derek A Epp, Kevin Roach and Kelsey Shoub. 2017. “Racial Disparities in Traffic Stop Outcomes.” *Duke FL & Soc. Change* 9:21.
- Becker, Gary S. 2010. *The Economics of Discrimination*. Chicago: University of Chicago Press.
- Bellah, Robert N. 1967. “Civil religion in America.” *Daedalus* pp. 1–21.
- Berger, Peter L. 1967. *The social reality of religion*. Faber.
- Berger, Peter L. 2017. Pluralism, protestantization, and the voluntary principle. In *The New Sociology of Knowledge*. Routledge pp. 33–46.
- Berry, Helen L, Anthony Hogan, Jennifer Owen, Debra Rickwood and Lyn Fragar. 2011. “Climate change and farmers’ mental health: risks and responses.” *Asia Pacific Journal of Public Health* 23(2\_suppl):119S–132S.

- Berry, Kate. 2015. *How Judicial Elections Impact Criminal Cases*. Brennan Center for Justice.
- Berry, William D., Jacqueline H. R. DeMeritt and Justin Esarey. 2010. “Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?” *American Journal of Political Science* 54(1):248 – 266.
- Berry, William D, Matt Golder and Daniel Milton. 2012. “Improving tests of theories positing interaction.” *The Journal of Politics* 74(3):653–671.
- Bertrand, M. and E. Duflo. 2017a. Field Experiments on Discrimination. In *Handbook of Field Experiments*, ed. Abhijit Vinayak Banerjee and Esther Duflo. Vol. 1 North-Holland pp. 309 – 393.  
**URL:** <http://www.sciencedirect.com/science/article/pii/S2214658X1630006X>
- Bertrand, Marianne, Dolly Chugh and Sendhil Mullainathan. 2005. “Implicit discrimination.” *American Economic Review* pp. 94–98.
- Bertrand, Marianne and Esther Duflo. 2017b. Field experiments on discrimination. In *Handbook of Economic Field Experiments*. Vol. 1 Elsevier pp. 309–393.
- Bertrand, Marianne and Sendhil Mullainathan. 2004a. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94(4):991–1013.
- Bertrand, Marianne and Sendhil Mullainathan. 2004b. “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination.” *American economic review* 94(4):991–1013.
- Bess, James L and Paul Goldman. 2001. “Leadership ambiguity in universities and K–12 schools and the limits of contemporary leadership theory.” *The Leadership Quarterly* 12(4):419–450.
- Bigo, Didier and Elspeth Guild. 2005. *Controlling Frontiers: Free Movement into and within Europe*. Farnham: Ashgate Publishing.
- Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2016. “Declaring and diagnosing research designs.” *Unpublished manuscript* .
- Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. “Latent dirichlet allocation.” *Journal of Machine Learning Research* 3:993–1022.
- Blommaert, Lieselotte, Marcel Coenders and Frank van Tubergen. 2014. “Ethnic discrimination in recruitment and decision makers’ features: Evidence from laboratory experiment and survey data using a student sample.” *Social indicators research* 116(3):731–754.

- Bonikowski, Bart and Paul DiMaggio. 2016. "Varieties of American popular nationalism." *American Sociological Review* 81(5):949–980.
- Bonilla-Silva, Eduardo. 2017. *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield.
- Bowling, Ben. 1990. "Conceptual and methodological problems in measuring race differences in delinquency: A reply to Marianne Junger." *British Journal of Criminology* 30:483.
- Brace, Paul and Brent D Boyea. 2008. "State public opinion, the death penalty, and the practice of electing judges." *American Journal of Political Science* 52(2):360–372.
- Brambor, Thomas, William Roberts Clark and Matt Golder. 2006a. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82.
- Brambor, Thomas, William Roberts Clark and Matt Golder. 2006b. "Understanding interaction models: Improving empirical analyses." *Political analysis* 14(1):63–82.
- Braunstein, Ruth. 2019. Muslims as outsiders, enemies, and others: The 2016 presidential election and the politics of religious exclusion. In *Politics of Meaning/Meaning of Politics*. Springer pp. 185–206.
- Bregman, Peter. 2012. "Diversity training doesn't work." *Forbes.com*, March 12.
- Brehm, John O and Scott Gates. 1999. *Working, shirking, and sabotage: Bureaucratic response to a democratic public*. University of Michigan Press.
- Brehm, John and Scott Gates. 1993. "Donut shops and speed traps: Evaluating models of supervision on police behavior." *American Journal of Political Science* pp. 555–581.
- Bright, Stephen B and Patrick J Keenan. 1995. "Judges and the politics of death: Deciding between the Bill of Rights and the next election in capital cases." *BUL rev.* 75:759.
- Broockman, David 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.
- Broockman, David and Evan Soltas. 2017. "A natural experiment on taste-based racial and ethnic discrimination in elections."
- Bruce, Steve. 2002. *God is dead: Secularization in the West*. Blackwell Oxford.
- Buchanan, Larry, Ford Fessenden, KK Rebecca Lai, Haeyoun Park, Alicia Parlapiano, Archie Tse, Tim Wallace, Derek Watkins and Karen Yourish. 2014. "What happened in Ferguson." *The New York Times* .

- Bunzel, John H and Jeffrey KD Au. 1987. "Diversity or Discrimination?-Asian Americans in College." *The Public Interest* 87:49.
- Burgess, Diana, Michelle Van Ryn, John Dovidio and Somnath Saha. 2007. "Reducing racial bias among health care providers: Lessons from social-cognitive psychology." *Journal of general internal medicine* 22(6):882–887.
- Burstein, Paul. 2003. "The impact of public opinion on public policy: A review and an agenda." *Political research quarterly* 56(1):29–40.
- Bushman, Brad J and Angelica M Bonacci. 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, Christopher K, Tali Gluch and Neil J Mitchell. 2007. "Security Forces and Sexual Violence: A Cross-National Analysis of a Principal—Agent Argument." *Journal of Peace Research* 44(6):669–687.
- Butler, Daniel M. 2014. *Representing the Advantaged: How Politicians Reinforce Inequality*. Cambridge University Press.
- Butler, Daniel M and Charles Crabtree. 2017a. "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* .
- Butler, Daniel M and Charles Crabtree. 2017b. "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* 4(1):57–67.
- Butler, Daniel M, Christopher F Karpowitz and Jeremy C Pope. 2012. "A field experiment on legislators' home styles: service versus policy." *The Journal of Politics* 74(02):474–486.
- Butler, Daniel M and David E Broockman. 2009. Who helps DeShawn register to vote? A field experiment on state legislators. In *Unpublished paper*. <https://harrisschool.uchicago.edu/programs/beyond/workshops/ampolpapers/fall09-butler.pdf>.
- Butler, Daniel M and David E Broockman. 2011a. "Do politicians racially discriminate against constituents? A field experiment on state legislators." *American Journal of Political Science* 55(3):463–477.
- Butler, Daniel M and David E Broockman. 2011b. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3):463–477.

- Butler, Daniel M and Jonathan Homola. 2017a. "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments." *Political Analysis* 25(1):122–130.
- Butler, Daniel M and Jonathan Homola. 2017b. "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments." *Political Analysis* 25(1):122–130.
- Butler, Daniel M and Miguel M Pereira. 2018. "Are Donations to Charity an Effective Incentive for Public Officials?" *Journal of Experimental Political Science* 5(1):68–70.
- Butz, Adam M., Brandy A. Kennedy, Nazita Lajevardi and Matthew J. Nanes. 2018. "Race and Representative Bureaucracy in American Policing: New Data, New Opportunities." *Comparative Politics Newsletter* 28:11–18.
- Canes-Wrone, Brandice, Tom S Clark and Jason P Kelly. 2014. "Judicial selection and death penalty decisions." *American Political Science Review* 108(1):23–39.
- Carenen, Caitlin. 2012. *The Fervent Embrace: Liberal Protestants, Evangelicals, and Israel*. NYU Press.
- Carnes, Nicholas and John B Holbein. 2015. Do public officials exhibit social class biases when they handle casework? Evidence from multiple correspondence experiments. Technical report Working paper. <http://people.duke.edu/nwc8/research.html>.
- Cashmore, Ernest and Eugene McLaughlin. 2013. *Out of Order?: Policing Black People*. London: Routledge.
- Center, Pew Research. 2015. "America's changing religious landscape." *Pew Research Center*.
- Chan, Janet BL. 1999. "Governing police practice: limits of the new accountability." *The British journal of sociology* 50(2):251–270.
- Chevigny, Paul and Paul Chevigny. 1995. *Edge of the Knife: Police Violence in the Americas*. New York, NY: New Press.
- Chipman, Hugh A, Edward I George, Robert E McCulloch et al. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4(1):266–298.
- Clark, William Roberts, Matt Golder and Sona Nadenichek Golder. 2017. *Principles of comparative politics*. CQ Press.
- Coe, Kevin and David Domke. 2006. "Petitioners or prophets? Presidential discourse, God, and the ascendancy of religious conservatives." *Journal of Communication* 56(2):309–330.

- Coffman, Lucas C and Muriel Niederle. 2015. "Pre-analysis plans have limited upside, especially where replications are feasible." *The Journal of Economic Perspectives* 29(3):81–97.
- Cohn, Samuel. 2000. *Race and gender discrimination at work*. Westview Press Boulder, CO.
- Conrad, Courtenay R. 2018. "Why do Courts Protect Human Rights? Investigating the Mechanisms by Focusing on the Police." *Comparative Politics Newsletter* 28:19–24.
- Coppock, Alexander. 2019. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6(1):1–4.
- Correll, Joshua, Bernadette Park, Charles M Judd, Bernd Wittenbrink, Melody S Sadler and Tracie Keese. 2007. "Across the thin blue line: police officers and racial bias in the decision to shoot." *Journal of personality and social psychology* 92(6):1006.
- Correll, Joshua, Sean M Hudson, Steffanie Guillermo and Debbie S Ma. 2014. "The police officer's dilemma: A decade of research on racial bias in the decision to shoot." *Social and Personality Psychology Compass* 8(5):201–213.
- Costa, Mia. 2017. "How responsive are political elites? A meta-analysis of experiments on public officials." *Journal of Experimental Political Science* 4(3):241–254.
- Costa, Mia. N.d. "How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials." . Forthcoming.
- Council, National Research et al. 2004. *Measuring racial discrimination*. National Academies Press.
- Crabtree, Charles. 2017. An Introduction to Email Audit Studies. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis. Methodos Series New York, NY: Springer.
- Crabtree, Charles. 2018. "Introduction: The Comparative Politics of Policing." *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 28(1):3–11.
- Crabtree, Charles, Christopher Fariss and Holger Kern. 2015. Truth Replaced by Silence: Private Censorship in Russia.
- Crabtree, Charles and Christopher J Fariss. 2016. "Stylized Facts and Experimentation." *Sociological Science* 3:910–914.
- Crabtree, Charles, Holger L Kern and Matthew T Pietryka. 2018. "Sponsorship Effects in Online Surveys."

- Crabtree, Charles, John Holbein, L. Holger Kern and Steve Pfaff. 2018. Religious Discrimination among K-12 Principals: Evidence from a Large-Scale Field Experiment. Technical report Working Paper.
- Crabtree, Charles, Matt Golder, Thomas Gschwend and Indriði H. Indriðason. N.d. Campaign Sentiment in European Party Manifestos. Technical report Working Paper.
- Crabtree, Charles and Volha Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5:21–28.
- Crabtree, Charles, Yue Hou and Chuyu Liu. 2018. Anti-Muslim Discrimination in the Chinese Labor Market. Technical report Working Paper.
- Cragun, Ryan T, Barry Kosmin, Ariela Keysar, Joseph H Hammer and Michael Nielsen. 2012. "On the receiving end: Discrimination toward the non-religious in the United States." *Journal of Contemporary Religion* 27(1):105–127.
- Crosson, Jesse, Zander Furnas and Geoffrey Lorenz. 2018. "Estimating interest group ideal points using public position-taking data." *Working Paper* .
- Da Silva, Chantal. 2018. "NYPD will launch implicit bias training for police officers three years after Eric Garner's death." Newsweek.
- Dahl, Robert Alan. 1989. *Democracy and its Critics*. Yale University Press.
- Dancygier, Rafaela M and David D Laitin. 2014. "Immigration into Europe: Economic discrimination, violence, and public policy." *Annual Review of Political Science* 17:43–64.
- Daniel, William Wentworth. 1968. *Racial discrimination in England: based on the PEP report*. Vol. 257 Penguin.
- Davenport, Christian. 2005. "Understanding Covert Repressive Action: The Case of the U.S. Government against the Republic of New Africa." *Journal of Conflict Resolution* 49(1):120 – 140.
- Davidson, James D and Ralph E Pyle. 2011. *Ranking faiths: Religious stratification in America*. Rowman & Littlefield Publishers.
- Davis, Angela J. 2017. *Policing the Black Man: Arrest, Prosecution, and Imprisonment*. New York: Pantheon.
- Davis, Angela Y. 2016. *Freedom is a constant struggle: Ferguson, Palestine, and the foundations of a movement*. Haymarket Books.
- Desposato, Scott. 2018. "Subjects and Scholars' Views on the Ethics of Political Science Field Experiments." *Perspectives on Politics* 16(3):739–750.



- Devine, Patricia G. 1989. "Stereotypes and prejudice: Their automatic and controlled components." *Journal of personality and social psychology* 56(1):5.
- Devine, Patricia G and Margo J Monteith. 1999. "Automaticity and control in stereotyping."
- Dharmapala, Dhammika, Nuno Garoupa and Richard H McAdams. 2016. "Punitive police? Agency costs, law enforcement, and criminal procedure." *The Journal of Legal Studies* 45(1):105–141.
- Dietrich, Nicholas and Charles Crabtree. 2018. Domestic Demand for Human Rights: Free Speech and the Freedom–Security Trade-Off. Technical report Working Paper.
- Distelhorst, Greg and Yue Hou. 2017. "Constituency service under nondemocratic rule: evidence from China." *The Journal of Politics* 79(3):1024–1040.
- Douthat, Ross. 2013. *Bad religion: How we became a nation of heretics*. Simon and Schuster.
- Dovidio, John F, Kerry Kawakami and Samuel L Gaertner. 2002. "Implicit and explicit prejudice and interracial interaction." *Journal of personality and social psychology* 82(1):62.
- Dovidio, John F and Samuel L Gaertner. 2004. "Aversive racism." *Advances in experimental social psychology* 36:4–56.
- Driscoll, Jesse. 2015. "Prison States & Games of Chicken." *Ethics in Comparative Politics Experiments* .
- Dropp, Kyle and Zachary Peskowitz. 2012. "Electoral security and the provision of constituency service." *The Journal of Politics* 74(1):220–234.
- Drydakis, Nick. 2014. "Sexual orientation discrimination in the Cypriot labour market. Distastes or uncertainty?" *International Journal of Manpower* 35(5):720–744.
- Dryzek, John S. 1996. "Political inclusion and the dynamics of democratization." *American Political Science Review* 90(3):475–487.
- Earle, James H. 1988. "Law Enforcement Administration: Yesterday-Today-Tomorrow." *FBI L. Enforcement Bull.* 57:2.
- Eck, Kristine and Charles Crabtree. 2018. Gender Differences in the Prosecution of Police Brutality: Evidence from a Natural Experiment. Technical report Working Paper.
- Eckhouse, Laurel. 2016. "Descriptive Representation and Political Power: Explaining Racial Inequalities in Policing."

- Edgell, Penny, Douglas Hartmann, Evan Stewart and Joseph Gerteis. 2016. "Atheists and other cultural outsiders: Moral boundaries and the non-religious in the United States." *Social Forces* 95(2):607–638.
- Edgell, Penny, Joseph Gerteis and Douglas Hartmann. 2006. "Atheists as "other": Moral boundaries and cultural membership in American society." *American Sociological Review* 71(2):211–234.
- Edwards, F, MH Esposito and H Lee. 2018. "Risk of Police-Involved Death by Race/Ethnicity and Place, United States, 2012-2018." *American journal of public health* p. e1.
- Einstein, Katherine Levine and David M. Glick. 2017a. "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." *American Journal of Political Science* 61:100–116.
- Einstein, Katherine Levine and David M Glick. 2017b. "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.
- Eith, Christine and Matthew R. Durose. 2011. "Contacts between police and the public, 2008." Special Report, Bureau of Justice Statistics, U.S. Department of Justice.
- Ekins, Emily E. 2016. "Policing in America: understanding public attitudes toward the police. Results from a national survey."
- Elliott, Marc N, Allen Fremont, Peter A Morrison, Philip Pantoja and Nicole Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." *Health services research* 43(5p1):1722–1736.
- Elliott, Marc N, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja and Nicole Lurie. 2009. "Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities." *Health Services and Outcomes Research Methodology* 9(2):69.
- Engel, Robin S and Kristin Swartz. 2013. "Race, crime, and policing." *The Oxford handbook of ethnicity, crime, and immigration* pp. 135–65.
- Enns, Peter K. 2014. "The public's increasing punitiveness and its influence on mass incarceration in the United States." *American Journal of Political Science* 58(4):857–872.
- Enns, Peter K. 2016. *Incarceration nation*. Cambridge University Press.

- Epp, Charles R, Steven Maynard-Moody and Donald Haider-Markel. 2017. "Beyond profiling: The institutional sources of racial disparities in policing." *Public Administration Review* 77(2):168–178.
- Epp, Charles R, Steven Maynard-Moody and Donald P Haider-Markel. 2014. *Pulled over: How police stops define race and citizenship*. University of Chicago Press.
- Erikson, Robert S, Gerald C Wright and John P McIver. 1989. "Political parties, public opinion, and state policy in the United States." *American Political Science Review* 83(3):729–750.
- Essex, Nathan L. 2002. *School law and the public schools: A practical guide for educational leaders*. ERIC.
- Fagan, Jeffrey. 2017. "Recent Evidence And Controversies In "The New Policing"." *Journal of Policy Analysis and Management* 36(3):690–700.
- Findley, Michael, Daniel Nielson and Scott Desposato. 2016. "Obligated to Deceive? Aliases, Confederates, and the Common Rule in International Field Experiments." *Ethics and Experiments. Problems and Solutions for Social Scientists and Policy Professionals* pp. 151–70.
- Findley, Michael G, Daniel L Nielson and Jason Campbell Sharman. 2014. *Global shell games: Experiments in transnational relations, crime, and terrorism*. Vol. 128 Cambridge University Press.
- Findley, Michael G, Daniel L Nielson and JC Sharman. 2015. "Causes of noncompliance with international law: A field experiment on anonymous incorporation." *American Journal of Political Science* 59(1):146–161.
- Finke, Roger and Rodney Stark. 1992. "The Churching of America, 1776-1990. New Brunswick."
- Fiscella, Kevin and Allen M Fremont. 2006. "Use of geocoding and surname analysis to estimate race and ethnicity." *Health services research* 41(4p1):1482–1500.
- Foa, Roberto Stefan and Yascha Mounk. 2016. "The democratic disconnect." *Journal of Democracy* 27(3):5–17.
- Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.
- Freedman, David, Robert Pisani and Roger Purves. 2007. "Statistics (international student edition)." *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* .

- Fridell, Lorie A. 2016. "Racial aspects of police shootings: Reducing both bias and counter bias." *Criminology & Public Policy* 15(2):481–489.
- Fryer, Roland G. 2016. "An empirical analysis of racial differences in police use of force." National Bureau of Economic Research Working Paper 22399.
- Fryer, Roland and S. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119(3):767–805.
- Furstenberg, Frank F and Charles F Wellford. 1973. "Calling the police: The evaluation of police service." *Law & Society Review* 7(3):393–406.
- Gaddis, S Michael. 2014. "Discrimination in the credential society: An audit study of race and college selectivity in the labor market." *Social Forces* p. sou111.
- Gaddis, S. Michael. 2017a. *The Ethics of Audit Studies: Best Practices and Academic Cooperation*. Vol. Audit Studies: Behind the Scenes with Theory, Method, and Nuance Springer.
- Gaddis, S Michael. 2017b. "How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies." *Sociological Science* 4:469–489.
- Gaddis, S Michael. 2017c. "How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–489.
- Gaddis, S. Michael. 2017d. An Introduction to Audit Studies in the Social Sciences. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis. Methodos Series New York, NY: Springer.
- Gaddis, S Michael. 2017e. "Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." *Socius: Sociological Research for a Dynamic World* .
- Gaddis, S Michael. 2018a. *Audit studies: Behind the scenes with theory, method, and nuance*. Vol. 14 Springer.
- Gaddis, S Michael. 2018b. An introduction to audit studies in the social sciences. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer pp. 3–44.
- Gaddis, S Michael and Raj Ghoshal. 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.
- Gailmard, Sean. 2012. "Accountability and principal-agent models." *Chapter prepared for the Oxford Handbook of Public Accountability* .
- García-Bedolla, Lisa and Melissa R. Michelson. 2012. *Mobilizing Inclusion: Transforming the Electorate Through Get-out-the-Vote Campaigns*. Yale University Press.

- Ge, Yanbo, Christopher R. Knittel, Don MacKenzie and Stephen Zoepf. 2018. Racial and Gender Discrimination in Transportation Network Companies. Technical report NBER Working Paper.
- Gell-Redman, Micah, Neil Visalvanich, Charles Crabtree and Christopher Fariss. 2018a. “It’s all about race: How state legislators respond to immigrant constituents.” *Political Research Quarterly* .
- Gell-Redman, Micah, Neil Visalvanich, Charles Crabtree and Christopher J Fariss. 2018b. “It’s All about Race: How State Legislators Respond to Immigrant Constituents.” *Political Research Quarterly* .
- Gelman, Andrew, Jeffrey Fagan and Alex Kiss. 2007. “An analysis of the New York City police department’s “stop-and-frisk” policy in the context of claims of racial bias.” *Journal of the American Statistical Association* 102(479):813–823.
- Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press.
- Gerber, Alan S and Donald P Green. 2008. Field experiments and natural experiments. In *The Oxford handbook of political science*.
- Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.
- Gerteis, Joseph. 2011. Civil religion and the politics of belonging. In *Rethinking Obama*. Emerald Group Publishing Limited pp. 215–223.
- Gervais, Will M. 2014. “Everything is permitted? People intuitively judge immorality as representative of atheists.” *PloS one* 9(4):e92302.
- Ghoshal, Raj and S Michael Gaddis. 2015. “Finding a roommate on craigslist: Racial discrimination and residential segregation.”
- Giles, Micheal W. and James C. Garand. 2007. “Ranking political science journals: Reputational and citational approaches.” *PS: Political Science & Politics* 40(4):741–751.
- Giulietti, Corrado, Mirco Tonin and Michael Vlassopoulos. 2015. “Racial Discrimination in Local Public Services: A Field Experiment in the US.”
- Golder, Matt and Benjamin Ferland. 2017. “Electoral rules and citizen-elite ideological congruence.”
- Golder, Matt and Jacek Stramski. 2010. “Ideological congruence and electoral institutions.” *American Journal of Political Science* 54(1):90–106.

- Green, Alexander R, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni and Mahzarin R Banaji. 2007. "Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients." *Journal of general internal medicine* 22(9):1231–1238.
- Green, Donald P, Shang E Ha and John G Bullock. 2010. "Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose." *The Annals of the American Academy of Political and Social Science* 628(1):200–208.
- Griffin, John D and Brian Newman. 2008. *Minority report: Evaluating political equality in America*. University of Chicago Press.
- Griffiths, Thomas L. and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101(suppl. 1):5228–5235.
- Grimmer, Justin, Eitan Hersh, Marc Meredith, Jonathan Mummolo and Clayton Nall. 2018. "Obstacles to estimating voter ID laws' effect on turnout." *Journal of Politics* 80(3).
- Grose, Christian R. 2011. *Congress in black and white: Race and representation in Washington and at home*. Cambridge University Press.
- Grose, Christian R. 2014. "Field Experimental Work on Political Institutions." *Annual Review of Political Science* 17.
- Grzymała-Busse, Anna. 2015. *Nations under God: How churches use moral authority to influence policy*. Princeton University Press.
- Guryan, Jonathan and Kerwin Kofi Charles. 2013. "Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots." *The Economic Journal* 123(572):F417–F432.
- Hadar, Ilana and John R Snortum. 1975. "The eye of the beholder: Differential perceptions of police by the police and the public." *Correctional Psychologist* 2(1):37–54.
- Hadden, Sally E. 2001. *Slave Patrols: Law and Violence in Virginia and the Carolinas*. London, UK: Harvard University Press.
- Hafner-Burton, Emilie M. 2008. "Sticks and stones: Naming and shaming the human rights enforcement problem." *International Organization* 62(4):689–716.
- Hahn, Harlan. 1971. "Ghetto assessments of police protection and authority." *Law & Society Review* 6(2):183–194.
- Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* 20(1):25–46.

- Hainmueller, Jens and Daniel J Hopkins. 2015. "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science* 59(3):529–548.
- Hainmueller, Jens and Dominik Hangartner. 2013. "Who gets a Swiss passport? A natural experiment in immigrant discrimination." *American political science review* 107(1):159–187.
- Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2016. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* pp. 1–30.
- Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2018. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* In Press.
- Hajnal, Zoltan and Marisa Abrajano. 2015. *White Backlash: Immigration, Race, and American Politics*. Princeton University Press.
- Hajnal, Zoltan, Nazita Lajevardi and Lindsay Nielson. 2017. "Voter Identification Laws and the Suppression of Minority Votes." *The Journal of Politics* 79(2):363–379.
- Hajnal, Zoltan and T. Lee. 2011. *Why Americans Don't Join the Party: Race, Immigration, and the Failure (of Political Parties) to Engage the Electorate*. Princeton University Press.
- Hall, Melinda Gann. 1995. "Justices as representatives: Elections and judicial politics in the American states." *American Politics Quarterly* 23(4):485–503.
- Hanson, Andrew and Michael Santas. 2014. "Field experiment tests for discrimination against Hispanics in the US rental housing market." *Southern Economic Journal* 81(1):135–167.
- Hartmann, Douglas, Daniel Winchester, Penny Edgell and Joseph Gerteis. 2011. "How Americans understand racial and religious differences: a test of parallel items from a national survey." *The Sociological Quarterly* 52(3):323–345.
- Hartmann, Douglas, Xuefeng Zhang and William Wischstadt. 2005. "One (Multicultural) Nation Under God? Changing Uses and Meanings of the Term "Judeo-Christian" in the American Media." *Journal of Media and Religion* 4(4):207–234.
- Heckman, James J. 1998a. "Detecting discrimination." *The Journal of Economic Perspectives* 12(2):101–116.
- Heckman, James J. 1998b. "Detecting discrimination." *The Journal of Economic Perspectives* 12(2):101–116.

- Hehman, Eric, Jessica K Flake and Jimmy Calanchini. 2017. “Disproportionate use of lethal force in policing is associated with regional racial biases of residents.” *Social Psychological and Personality Science* p. 1948550617711229.
- Hemker, Johannes and Anselm Rink. 2017. “Multiple dimensions of bureaucratic discrimination: Evidence from German welfare offices.” *American Journal of Political Science* 61(4):786–803.
- Herberg, Will. 1983. *Protestant–Catholic–Jew: An Essay in American Religious Sociology*. University of Chicago Press.
- Hill, Jennifer L. 2011. “Bayesian nonparametric modeling for causal inference.” *Journal of Computational and Graphical Statistics* 20(1):217–240.
- Hirschman, Charles. 2004. “The Role of Religion in the Origins and Adaptation of Immigrant Groups in the United States 1.” *International Migration Review* 38(3):1206–1233.
- Hjorth, Frederik. 2017. “The Influence of Local Ethnic Diversity on Group-Centric Crime Attitudes.” *British Journal of Political Science* pp. 1–23.
- Hogan, Bernie and Brent Berry. 2011. “Racial and ethnic biases in rental housing: An audit study of online apartment listings.” *City & Community* 10(4):351–372.
- Holbein, John. 2016. “Left behind? Citizen responsiveness to government performance information.” *American Political Science Review* 110(2):353–368.
- Hölmstrom, Bengt. 1979. “Moral hazard and observability.” *The Bell journal of economics* pp. 74–91.
- Honaker, James, Gary King and Matthew Blackwell. 2012. “Version 1.2–15.” *Amelia program*. *Gking.harvard.edu/Amelia* .
- Hout, Michael and Claude S Fischer. 2002. “Why more Americans have no religious preference: Politics and generations.” *American Sociological Review* pp. 165–190.
- Hughes, D. Alex, Micah Gell-Redman and Charles Crabtree. 2016. “Who gets to vote?” Evidence in Government and Politics, EGAP ID: 20161001AA.  
**URL:** <http://egap.org/registration/2183>
- Hughes, D Alex, Micah Gell-Redman, Charles Crabtree, Natarajan Krishnaswami, Diana Rodenberger and Guillermo Monge. 2017. “Continuing Evidence of Discrimination Among Local Election Officials.”
- Hughes, D Alex, Micah Gell-Redman, Charles Crabtree, Natarajan Krishnaswami, Diana Rodenberger and Guillermo Monge. 2019. “Persistent Bias Among Local Election Officials.” *Journal of Experimental Political Science* .



- Imai, Kosuke and Kabir Khanna. 2016. "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis* 24(2):263–272.
- Inglehart, Ronald and Pippa Norris. 2004. *Sacred and Secular: Religion and Politics Worldwide*. *Cambridge Studies in Social Theory, Religion, and Politics*. Cambridge University Press Cambridge.
- Jamal, Amaney and Nadine Naber. 2007. *Race and Arab Americans Before and After 9/11: From Invisible Citizens to Visible Subjects*. Syracuse University Press.
- James, Lois, Bryan Vila and Kenn Daratha. 2013. "Results from experimental trials testing participant responses to White, Hispanic and Black suspects in high-fidelity deadly force judgment and decision-making simulations." *Journal of Experimental Criminology* 9(2):189–212.
- James, Tom. 2017. "Can cops unlearn their unconscious biases?" *The Atlantic*.
- Jauregui, Beatric. 2016. *Provisional Authority: Police, Order, and Security in India*. Chicago, IL: The University of Chicago Press.
- Jee-Lyn García, Jennifer and Mienah Zulfacar Sharif. 2015. "Black lives matter: a commentary on racism and public health." *American journal of public health* 105(8):e27–e30.
- Jockers, Matthew Lee. 2014. *Text Analysis with R for Students of Literature*. Cham: Springer.
- Johnson, Daniel A, Richard J Porter and Patricia L Mateljan. 1971. "Racial Discrimination in Apartment Rentals 1." *Journal of Applied Social Psychology* 1(4):364–377.
- Johnson, Monica Kirkpatrick, Robert Crosnoe and Glen H Elder Jr. 2001. "Students' attachment and academic engagement: The role of race and ethnicity." *Sociology of education* pp. 318–340.
- Justice, Benjamin and Colin Macleod. 2016. *Have a little faith: Religion, democracy, and the American public school*. University of Chicago Press.
- Kao, Grace and Jennifer S Thompson. 2003. "Racial and ethnic stratification in educational achievement and attainment." *Annual review of sociology* 29(1):417–442.
- Kauff, Mathias, Ralf Wölfer and Miles Hewstone. 2017. "Impact of discrimination on health among adolescent immigrant minorities in Europe: The role of perceived discrimination by police and security personnel." *Journal of Social Issues* 73(4):831–851.
- Kelly, Nathan J and Peter K Enns. 2010. "Inequality and the dynamics of public opinion: The self-reinforcing link between economic inequality and mass preferences." *American Journal of Political Science* 54(4):855–870.

- Kennedy, Brandy A., Adam M. Butz, Nazita Lajevardi and Matthew J. Nanes. 2017a. *Race and Representative Bureaucracy in American Policing*. London: Palgrave.
- Kennedy, Brandy A, Adam M Butz, Nazita Lajevardi and Matthew J Nanes. 2017b. *Race and Representative Bureaucracy in American Policing*. Springer.
- King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722.
- King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.
- Knight, Amy W. 1990. *The KGB: Police and Politics in the Soviet Union*. London, UK: Unwin Hyman.
- Knowles, John, Nicola Persico and Petra Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* 109(1):203–229.
- Kuyper, Jonathan W. 2016. "Systemic representation: democracy, deliberation, and non-electoral representatives." *American Political Science Review* 110(2):308–324.
- Lahey, Joanna N and Ryan A Beasley. 2009. "Computerizing audit studies." *Journal of economic behavior & organization* 70(3):508–514.
- Lax, Jeffrey R and Justin H Phillips. 2009. "Gay rights in the states: Public opinion and policy responsiveness." *American Political Science Review* 103(3):367–386.
- Lebrecht, Sophie, Lara J Pierce, Michael J Tarr and James W Tanaka. 2009. "Perceptual other-race training reduces implicit racial bias." *PloS one* 4(1):e4215.
- Lin, Winston and Donald P Green. 2015. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *Berkeley*. Retrieved from [www. stat. berkeley. edu/~winston/sop-safety-net. pdf](http://www.stat.berkeley.edu/~winston/sop-safety-net.pdf) (2014). *Promoting transparency in social science research*. *Science (New York, NY)* 343(6166):30–1.
- Lindsay, D Michael. 2007. *Faith in the halls of power: How evangelicals joined the American elite*. Oxford University Press on Demand.
- Lippy, Charles H. 2006. *Faith in America: Changes, challenges, new directions*. Vol. 1 Greenwood Publishing Group.
- Lipsky, Michael. 1971. "Street-level bureaucracy and the analysis of urban reform." *Urban Affairs Quarterly* 6(4):391–409.
- Lipsky, Michael. 1980. *Street-level Bureaucracy: Dilemmas of the Individual in Public Services*. Russell Sage.

- Lipsky, Michael. 2010. *Street-level Bureaucracy: Dilemmas of the Individual in Public Services*. New York, NY: Russell Sage Foundation.
- Lohman, Joseph D and Dietrich C Reitzes. 1952. "Note on race relations in mass society." *American Journal of Sociology* 58(3):240–246.
- Lohr, Sharon. 2009. *Sampling: design and analysis*. Nelson Education.
- Macpherson, David A and Barry T Hirsch. 1995. "Wages and gender composition: why do women's jobs pay less?" *Journal of Labor Economics* 13(3):426–471.
- Mak, Aaron. 2018. "What can Starbucks accomplish?" Slate.
- Manning, Christel. 2015. *Losing our religion: How unaffiliated parents are raising their children*. NYU Press.
- Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*. pp. 55–60.
- Manning, Peter K. 2005. "The study of policing." *Police quarterly* 8(1):23–43.
- Mansbridge, Jane. 2003. "Rethinking representation." *American political science review* 97(4):515–528.
- Marcus, Nancy C. 2016. "From Edward to Eric Garner and beyond: the importance of constitutional limitations on lethal use of force in police reform." *Duke J. Const. L. & Pub. Pol'y* 12:53.
- Marler, Penny Long and C Kirk Hadaway. 2002. "'Being religious' or 'being spiritual' in America: A zero-sum proposition?" *Journal for the Scientific Study of Religion* 41(2):289–300.
- Mas, Alexandre. 2006. "Pay, reference points, and police performance." *The Quarterly Journal of Economics* 121(3):783–821.
- Matzke, Nicholas J. 2016. "The evolution of antievolution policies after Kitzmiller versus Dover." *Science* 351(6268):28–30.
- Mauer, Marc. 2006. *Race to incarcerate*. New Press, The.
- Maurer-Fazio, Margaret. 2012. "Ethnic discrimination in China's internet job board labor market." *IZA Journal of Migration* 1(1):12.
- Maynard, Robyn. 2017. "Policing black lives: State violence in Canada from slavery to the present." *Vancouver, BC: Fernwood Publishing. Google Scholar* .

- Mayrl, Damon. 2016. *Secular conversions: Political institutions and religious education in the United States and Australia, 1800–2000*. Cambridge University Press.
- McAdams, Richard H, Dhammika Dharmapala and Nuno Garoupa. 2015. “The law of police.” *The University of Chicago Law Review* pp. 135–158.
- McClendon, Gwyneth. 2012. Race, Responsiveness, and Electoral Strategy: A Field Experiment with South African Politicians. Technical report Working Paper, Princeton University.
- McDonnell, Lorraine M. 2013. “Educational accountability and policy feedback.” *Educational Policy* 27(2):170–189.
- McNulty, John E., Conor M. Dowling and Margaret H. Ariotti. 2009. “Driving Saints to Sin: How Increasing the Difficulty of Voting Dissuades Even the Most Motivated Voters.” *Political Analysis* 17(4):435–455.
- Michelitch, Kristin. 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.
- Milkman, Katherine L, Modupe Akinola and Dolly Chugh. 2012. “Temporal Distance and Discrimination An Audit Study in Academia.” *Psychological Science* 23(7):710–717.
- Milkman, Katherine L, Modupe Akinola and Dolly Chugh. 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.
- Mitchell, Michael J and Charles H Wood. 1999. “Ironies of citizenship: skin color, police brutality, and the challenge to democracy in Brazil.” *Social Forces* 77(3):1001–1020.
- Monkkonen, Eric H. 1981. *Police in Urban America: 1860-1920*. Cambridge, UK: Cambridge University Press.
- Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. “How conditioning on posttreatment variables can ruin your experiment and what to do about it.” *American Journal of Political Science* 62(3):760–775.
- Moore, Ryan T and Keith Schnakenberg. 2012. “blockTools: Blocking, assignment, and diagnosing interference in randomized experiments.” *R package version 0.5-7*. [http://rtm.wustl.edu/software.blockTools.htm] .
- Morgan, Stephen L and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

- Morgan, Stephen L and David J Harding. 2006. "Matching estimators of causal effects: Prospects and pitfalls in theory and practice." *Sociological methods & research* 35(1):3–60.
- Muir, William Ker. 1977. *Police: Streetcorner Politicians*. Chicago, IL: The University of Chicago Press.
- Näsström, Sofia. 2015. "Democratic representation beyond election." *Constellations* 22(1):1–12.
- Nelder, John Ashworth and Robert WM Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society: Series A (General)* 135(3):370–384.
- Neuhaus, Richard John. 1984. "The naked public sphere: Religion and democracy in America." *Grand Rapids, Mich: Eerdmans* .
- Neumark, David. 2012. "Detecting discrimination in audit and correspondence studies." *Journal of Human Resources* 47(4):1128–1157.
- Neumark, David, Roy J Bank and Kyle D Van Nort. 1995. Sex discrimination in restaurant hiring: an audit study. Technical report National Bureau of Economic Research.
- Nickerson, David W. 2005. "Scalable protocols offer efficient design for field experiments." *Political Analysis* 13(3):233–252.
- Nix, Justin, Bradley A Campbell, Edward H Byers and Geoffrey P Alpert. 2017. "A bird's eye view of civilians killed by police in 2015: Further evidence of implicit bias." *Criminology & Public Policy* 16(1):309–340.
- Noll, Mark A. 2002. *America's God: From Jonathan Edwards to Abraham Lincoln*. Oxford University Press on Demand.
- Northup, Temple. 2010. "Is Everyone a Little Bit Racist? Exploring Cultivation Using Implicit and Explicit Measures." *Southwestern Mass Communication Journal* 26(1).
- Oh, Sun Jung and John Yinger. 2015. "What Have We Learned From Paired Testing in Housing Markets?" *Cityscape* 17(3):15.
- Olken, Benjamin A. 2015. "Promises and perils of pre-analysis plans." *The Journal of Economic Perspectives* 29(3):61–80.
- Ondrich, Jan, Alex Stricker and John Yinger. 1999. "Do landlords discriminate? The incidence and causes of racial discrimination in rental housing markets." *Journal of Housing Economics* 8(3):185–204.
- Pager, Devah. 2003. "The mark of a criminal record." *American journal of sociology* 108(5):937–975.

- Pager, Devah. 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(1):104–133.
- Pager, Devah. 2007b. "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(1):104–133.
- Pager, Devah, Bart Bonikowski and Bruce Western. 2009. "Discrimination in a low-wage labor market: A field experiment." *American sociological review* 74(5):777–799.
- Pager, Devah and Hana Shepherd. 2008. "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets." *Annual review of sociology* 34:181.
- Pager, Devah and Lincoln Quillian. 2005. "Walking the talk? What employers say versus what they do." *American Sociological Review* 70(3):355–380.
- Panagopoulos, Costas. 2006. "The Polls-Trends: Arab and Muslim Americans and Islam in the Aftermath of 9/11." *Public Opinion Quarterly* 70(4):608–624.
- Pateman, Carole. 1970. *Participation and democratic theory*. Cambridge University Press.
- Pedulla, David S. 2016. "Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories." *American sociological review* 81(2):262–289.
- Peek, Lori. 2011. *Behind the backlash: Muslim Americans after 9/11*. Temple University Press.
- Pennebaker, James W. 2015. "LIWC: How it works (<http://liwc.wpengine.com/how-it-works/>).".
- Persico, Nicola. 2009. "Racial profiling? Detecting bias using statistical evidence." *Annu. Rev. Econ.* 1(1):229–254.
- Peruche, B Michelle and E Ashby Plant. 2006. "The correlates of law enforcement officers' automatic and controlled race-based responses to criminal suspects." *Basic and Applied Social Psychology* 28(2):193–199.
- Pfaff, Steven. 2008. "The religious divide: Why religion seems to be thriving in the United States and waning in Europe." *Growing Apart: America and Europe in the Twenty-First Century*. Ed. Jeffrey Kopstein and Sven Steinmo. New York: Cambridge UP pp. 24–52.
- Pfaff, Steven, Charles Crabtree, Holger L Kern and John B Holbein. 2018. "Does religious bias shape access to public services? A large-scale audit experiment among street-level bureaucrats.".

- Pierné, Guillaume. 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(1):4.
- Pitkin, Hanna F. 1967. *The concept of representation*. Univ of California Press.
- Pope, Devin G, Joseph Price and Justin Wolfers. 2013. Awareness reduces racial bias. Technical report National Bureau of Economic Research.
- Prottas, Jeffrey Manditch. 1979. *People processing: the street-level bureaucrat in public service bureaucracies*. Lexington Books.
- Puhani, Patrick. 2000. "The Heckman correction for sample selection and its critique." *Journal of economic surveys* 14(1):53–68.
- Putnam, Robert D, Robert Leonardi and Raffaella Y Nanetti. 1994. *Making democracy work: Civic traditions in modern Italy*. Princeton university press.
- Putnam, Robert and David Campbell. 2010. "American grace: How religion is reshaping our civic and political lives."
- Quillian, Lincoln, Devah Pager, Ole Hexel and Arnfinn H Midtbøen. 2017. "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time." *Proceedings of the National Academy of Sciences* 114(41):10870–10875.
- Radicati, Sara and Quoc Hoang. 2011. "Email statistics report, 2011-2015." Retrieved May 25:2011.
- Rafky, David M. 1975. "Racial discrimination in urban police departments." *Crime & Delinquency* 21(3):233–242.
- Ray, Victor, Kasim Ortiz and Jacob Nash. 2018. "Who is policing the community? A comprehensive review of discrimination in police departments." *Sociology compass* 12(1):e12539.
- Rebhun, Uzi. 2016. *Jews and the American religious landscape*. Columbia University Press.
- Reese, William J. 2011. *America's Public Schools: From the Common School to "No Child Left Behind"*. JHU Press.
- Rehfeld, Andrew. 2006. "Towards a general theory of political representation." *Journal of Politics* 68(1):1–21.
- Reiner, Robert. 2010. *The politics of the police*. Oxford University Press.
- Reiss, Albert J. 1973. *The Police and the Public*. Vol. 39 New Haven, CT: Yale University Press.

- Riach, Peter A and Judith Rich. 2002. "Field experiments of discrimination in the market place." *The economic journal* 112(483):F480–F518.
- Riach, Peter A and Judith Rich. 2004. "Deceptive field experiments of discrimination: are they ethical?".
- Risse-Kappen, Thomas. 1991. "Public opinion, domestic structure, and foreign policy in liberal democracies." *World Politics* 43(4):479–512.
- Rivera, Lauren A and András Tilcsik. 2016. "Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market." *American Sociological Review* 81(6):1097–1131.
- Rocque, Michael. 2011. "Racial disparities in the criminal justice system and perceptions of legitimacy: A theoretical linkage." *Race and justice* 1(3):292–315.
- Roscigno, Vincent J. 1998. "Race and the reproduction of educational disadvantage." *Social Forces* 76(3):1033–1061.
- Ross, Cody T. 2015. "A multi-level Bayesian analysis of racial bias in police shootings at the county-level in the United States, 2011–2014." *PLoS One* 10(11):e0141854.
- Ross, Stephen L and John Yinger. 2002. *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. MIT Press.
- Ross, Stephen L and Margery Austin Turner. 2005. "Housing discrimination in metropolitan America: Explaining changes between 1989 and 2000." *Social Problems* 52(2):152–180.
- Rudman, Laurie A, Richard D Ashmore and Melvin L Gary. 2001. "'Unlearning' automatic biases: the malleability of implicit prejudice and stereotypes." *Journal of personality and social psychology* 81(5):856.
- Salisbury, Robert H and Kenneth A Shepsle. 1981. "US Congressman as enterprise." *Legislative Studies Quarterly* pp. 559–576.
- Salter, Michael. 2016. *Crime, justice and social media*. Routledge.
- Sarna, Jonathan D. 2004. *American Judaism: a history*. Yale University Press.
- Saward, Michael. 2008. "Representation and democracy: revisions and possibilities." *Sociology compass* 2(3):1000–1013.
- Schmitt, Michael T, Nyla R Branscombe and Tom Postmes. 2003. "Women's emotional responses to the pervasiveness of gender discrimination." *European Journal of Social Psychology* 33(3):297–312.



- Sen, Maya and Omar Wasow. 2016. "Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics." *Annual Review of Political Science* 19:499–522.
- Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing.
- Sharman, Jason 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.
- Sherkat, Darren E. 2014. *Changing faith: The dynamics and consequences of Americans' shifting religious identities*. NYU Press.
- Sherkat, Darren E and Christopher G Ellison. 1999. "Recent developments and current controversies in the sociology of religion." *Annual review of sociology* 25(1):363–394.
- Shusta, Robert M, Deena R Levine, Philip R Harris and Herbert Z Wong. 2002. *Multicultural law enforcement: Strategies for peacekeeping in a diverse society*. Prentice Hall Upper Saddle River, NJ.
- Sides, John, Michael Tesler and Lynn Vavreck. 2018. *Identity crisis: The 2016 presidential campaign and the battle for the meaning of America*. Princeton University Press.
- Sim, Jessica J, Joshua Correll and Melody S Sadler. 2013. "Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot." *Personality and social psychology bulletin* 39(3):291–304.
- Simpson, Ain and Kimberly Rios. 2017. "The moral contents of anti-atheist prejudice (and why atheists should care about it)." *European Journal of Social Psychology* 47(4):501–508.
- Sinclair, Betsy, Margaret McConnell and Donald P Green. 2012. "Detecting spillover effects: Design and analysis of multilevel experiments." *American Journal of Political Science* 56(4):1055–1069.
- Skolnick, Jerome H and James J Fyfe. 1993. *Above the law: Police and the Excessive Use of Force*. Free Press New York.
- Smith, Bruce. 1940. *Police systems in the United States*. Harper & brothers.
- Smith, Janet L. 2015a. Public housing transformation: Evolving national policy. In *Where are poor people to live?: transforming public housing communities*. Routledge pp. 31–52.
- Smith, Michael R and Geoffrey P Alpert. 2007. "Explaining police bias: A theory of social conditioning and illusory correlation." *Criminal justice and behavior* 34(10):1262–1283.

- Smith, Robert J. 2015b. “Reducing Racially Disparate Policing Outcomes: Is Implicit Bias Training the Answer.” *U. Haw. L. Rev.* 37:295.
- Sniderman, Paul M, JN Druckman, DP Green, JH Kuklinski and A Lupia. 2011. “The logic and design of the survey experiment.” *Cambridge handbook of experimental political science* pp. 102–114.
- Soroka, Stuart N and Christopher Wlezien. 2010. *Degrees of democracy: Politics, public opinion, and policy*. Cambridge University Press.
- Soss, Joe. 1999. “Lessons of welfare: Policy design, political learning, and political action.” *American Political Science Review* 93(2):363–380.
- Soss, Joe and Sanford F Schram. 2007. “A public transformed? Welfare reform as policy feedback.” *American Political Science Review* 101(1):111–127.
- Soss, Joe and Vesla Weaver. 2017. “Police are our government: Politics, political science, and the policing of race-class subjugated communities.” *Annual Review of Political Science* 20:565–591.
- Spencer, Katherine B, Amanda K Charbonneau and Jack Glaser. 2016. “Implicit bias and policing.” *Social and Personality Psychology Compass* 10(1):50–63.
- Staats, Cheryl, Kelly Capatosto, Robin A Wright and Danya Contractor. 2015. *State of the science: Implicit bias review 2015*. Vol. 3 Kirwan Institute for the Study of Race and Ethnicity, The Ohio State University.
- Staubli, Silvia. 2017. “Trusting the police: Comparisons across Eastern and Western Europe.” Swiss National Science Foundation.
- Stephens-Davidowitz, Seth. 2014. “The cost of racial animus on a black candidate: Evidence using Google search data.” *Journal of Public Economics* 118:26–40.
- Stocksdale, Michael. 2013. “E-mail: Not dead, evolving.”.
- Streb, Matthew J. 2007. *Running for judge: The rising political, financial, and legal stakes of judicial elections*. NYU Press.
- Subramanian, Kadayam Suryanarayanan. 2007. *Political violence and the police in India*. SAGE Publications India.
- Sullivan, Christopher M and Zachary P O’Keeffe. 2017. “Evidence that curtailing proactive policing can reduce major crime.” *Nature Human Behaviour* 1(10):730.
- Sun, Ivan Y., Yuning Wu and Rong Hu. 2013. “Public assessments of the police in rural and urban China: A theoretical extension and empirical investigation.” *British Journal of Criminology* 53(4):643–664.

- Surette, Raymond. 1985. "Crimes, arrests, and elections: Predicting winners and losers." *Journal of Criminal Justice* 13(4):321–327.
- Sutton, Matthew Avery. 2014. *American apocalypse: A history of modern evangelicalism*. Harvard University Press.
- Swatos, William H and Kevin J Christiano. 1999. "Introduction—Secularization theory: The course of a concept." *Sociology of Religion* 60(3):209–228.
- Teele, Dawn, Joshua L Kalla and Frances Rosenbluth. 2017. Faces of Bias in Politics: Evidence from Elite and Voter Conjoint Experiments on Gender. Technical report Working Paper.
- Terechshenko, Zhanna, Charles Crabtree, Kristine Eck and Christopher Fariss. 2019. "Evaluating the Influence of International Norms and Sanctioning on State Respect for Rights: A Field Experiment with Foreign Embassies." *International Interactions* .
- Tomaskovic-Devey, Donald, Marcinda Mason and Matthew Zingraff. 2004. "Looking for the driving while black phenomena: Conceptualizing racial bias processes and their associated distributions." *Police Quarterly* 7(1):3–29.
- Tonry, Michael H. 2011. *Punishing race: A continuing American dilemma*. Oxford University Press.
- Turner, Margery A, Stephen Ross, George C Galster and John Yinger. 2002. Discrimination in metropolitan housing markets: national results from phase 1 of the housing discrimination study (HDS). Technical report.
- Turney, Kristin and Grace Kao. 2009. "Barriers to school involvement: Are immigrant parents disadvantaged?" *The Journal of Educational Research* 102(4):257–271.
- Vedantam, Shankar. 2008. "Most diversity training ineffective, study finds." *The Washington Post* .
- Voas, David and Mark Chaves. 2016. "Is the United States a counterexample to the secularization thesis?" *American Journal of Sociology* 121(5):1517–1556.
- von Spakovsky, Hans A. 2018. "Racial Discrimination at Harvard University and America's Elite" Institutions. Legal Memorandum. No. 236." *Heritage Foundation* .
- Vuolo, Mike, Christopher Uggen and Sarah Lageson. 2016. "Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes." *Sociological Methods & Research* 45(2):260–303.
- Wald, Kenneth D and Clyde Wilcox. 2006. "Getting religion: has political science rediscovered the faith factor?" *American Political Science Review* 100(4):523–529.

- Wallace, Michael, Bradley RE Wright and Allen Hyde. 2014. "Religious Affiliation and Hiring Discrimination in the American South A Field Experiment." *Social Currents* 1(2):189–207.
- Warren, Patricia, DONALD Tomaskovic-Devey, William Smith, Matthew Zingraff and Marcinda Mason. 2006. "Driving while black: Bias processes and racial disparity in police stops." *Criminology* 44(3):709–738.
- Waterman, Richard W and Kenneth J Meier. 1998. "Principal-agent models: an expansion?" *Journal of public administration research and theory* 8(2):173–202.
- Weber, Max. 1965. "Politics as a Vocation."
- Weber, Max. 2013. *The Protestant ethic and the spirit of capitalism*. Routledge.
- Wetherell, Geoffrey A, Mark J Brandt and Christine Reyna. 2013. "Discrimination across the ideological divide: The role of value violations and abstract values in discrimination by liberals and conservatives." *Social Psychological and Personality Science* 4(6):658–667.
- White, Ariel R., Noah L. Nathan and Julie K. Faller. 2015a. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* 109(1):129–142.
- White, Ariel R, Noah L Nathan and Julie K Faller. 2015b. "What do I need to vote? Bureaucratic discretion and discrimination by local election officials." *American Political Science Review* 109(1):129–142.
- Wienk, Ronald E et al. 1979. "Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey."
- Wilcox, Clyde. 2018. *Onward Christian soldiers?: the religious right in American politics*. Routledge.
- Williams, Rowan. 2012. *Faith in the public square*. Bloomsbury Publishing.
- Wilson, James Q. 1978. *Varieties of Police Behavior: The Management of Law and Order in Eight Communities, With a New Preface by the Author*. Harvard University Press.
- Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York, NY: Basic Books.
- Wolfers, Justin. 2006. "Diagnosing discrimination: Stock returns and CEO gender." *Journal of the European Economic Association* 4(2-3):531–541.
- Word, David L, Charles D Coleman, Robert Nunziata and Robert Kominski. 2008. "Demographic aspects of surnames from census 2000." *Technical Report for the U.S. Census Bureau* .

- Wright, Bradley RE, Michael Wallace, John Bailey and Allen Hyde. 2013. "Religious affiliation and hiring discrimination in New England: A field experiment." *Research in Social Stratification and Mobility* 34:111–126.
- Yanow, Dvora and Peregrine Schwartz-Shea. 2016. "Encountering Your IRB 2.0: What Political Scientists Need to Know." *PS: Political Science & Politics* 49(02):277–286.
- Zimring, Franklin E. 2017. *When Police Kill*. Cambridge: Harvard University Press.