From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2021

Essays in Public Economics and Estimation

Lieberman, Carl

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01n870zt94w

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Essays in Public Economics and Estimation

Carl Adam Lieberman

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Economics

Adviser: Alexandre Mas

September 2021

© Copyright by Carl Adam Lieberman, 2021.

All rights reserved.

Abstract

This dissertation has two main themes. The first is using the tools of public and labor economics to estimate how policies and other forces impact the lives of disadvantaged populations so that we may better understand and address both problems faced and their possible solutions. The second applies ideas from econometrics and applied microeconomics to improve how we interpret data. The intersection of these themes in research offers new perspectives and insights into some of society's most pressing problems.

The first chapter studies racial disparities in police use of force, in particular how they vary. Despite efforts to address systemic racial inequities in the criminal justice system, in many areas, our understanding of the problems faced is poor. I combine novel data from the hundreds of independently operated police departments covering New Jersey with empirical Bayes techniques to estimate how racial disparities in police use of force vary across departments. I find significant variation but also note that these disparities are difficult to predict, suggesting a role for more idiosyncratic departmental features.

The second chapter, co-authored with Christina Korting, Jordan Matsudaira, Zhuan Pei, and Yi Shen, examines the econometric role of graphs, focusing on regression discontinuity designs (RDDs). We experimentally evaluate people's abilities to conduct "visual inference" with graphs to determine the presence of discontinuities in RDDs. Moreover, we compare visual and econometric inferences' performances, highlighting the value of graph-based visual inference and the existence of complementarities between them. Visual inference compares favorably with conventional econometric procedures, and we believe that there is an untapped frontier of graph-based econometrics to be explored.

The third chapter studies migratory responses to local minimum wages. Using the universe of tax filers in the US, I find that local minimum wages result in negatively selected out-migration from treated counties. Disemployment effects are one explanation, but I observe some pass-through to home values, and these may temper the positive effects of higher wages and earnings. There are no apparent spillovers to neighboring untreated areas, which may limit spatial distortions from these policies.

Acknowledgements

I would first like to thank Alex Mas, my formal adviser throughout my dissertation process, unofficial adviser before that, and boss before that. His continued guidance, support, and knowledge have been paramount to my growth as a researcher, and I hope to pay the kindness and wisdom he has shown back to the profession. Also invaluable has been Ilyana Kuziemko, a constant source of warmth as a teacher, colleague, and mentor. I have learned so much from Alex and Ilyana not only for their research insights and advice, but also their willingness to advise me as a person, not just as a scientist. My research has benefited tremendously from Princeton's faculty across the applied microeconomics fields, and Leah Boustan, David Lee, and Owen Zidar have been supremely generous with their time and knowledge in providing feedback and advice to myself and other young researchers.

I also thank my coauthors, Christina Korting, Jordan Matsudaira, Zhuan Pei, and Yi Shen. Our work has been the most fun I've had with research in my career, and I'm incredibly proud of what we've put together. I am especially fortunate to have been able to spend a year in the same building as Zhuan for the first time since we were both at Brandeis.

I am indebted to the support of the Industrial Relations Section and Princeton's staff. In addition to financial support, the efforts of Valerie Ching, Jeannie Moore, Patti Tracey, and Lori Mitrano have allowed me to make the most out of the resources available to me. And the indispensable Laura Hedden provided knowledge, advice, therapy, and candy throughout my time in Princeton.

I thank my colleagues and friends at Princeton University. Stephanie Hao, Chris Szerman, Emily Battaglia, Jiwon Choi, Lukas Althoff, Christine Cai, Carolyn Tsao, Dorisz Albrecht, Nastya Karpova, Victoria Larsen, Kwok Hao Lee, Dan McGee, and everyone in the Industrial Relations Section have created the most amazing environment to spend the last few years, and I will truly miss the many lunches, seminars, and hallway conversations we had. I thank, or at least absolve him of his debts for some of his battier proposals, Simon Quach, my pre-pandemic officemate and the other end of a patient partnership of pitching

and panning all manner of research ideas. The Industrial Relations Section has wonderful traditions in bringing in early career economists and ensuring contact between professors, postdocs, graduate students, visitors, and research assistants, and these have resulted in mountains of advice (and one police ride-along) from Jesse Bruhn, Cody Tuttle, and Sara Heller. I also thank the many individuals who supported me on short notice through an impromptu trip through the job market, especially for practice interviews and job talks, including Elisa Jácome and Amanda Eng, whose candidness as they went through the same experiences was hugely appreciated.

Finally, I am grateful for the unending support of my loved ones. Matt, Mom, and Dad have all provided an ear to listen during the most trying times of graduate school. And I would never have come this far without the unconditional love and support of my better half, Kathy.

Try jumping

Contents

Abstract iii

Acknowledgements iv

List of Tables xi

List of Figures xvi

1 Variation in Racial Disparities in Police Use of Force 1

1.1 Introduction 1

1.2 Institutional Background and Data 6

1.2.1 Use of Force and Force Reporting in New Jersey 6

1.2.2 Data 7

1.2.3 Summary Statistics 9

1.3 Empirical Strategy 11

1.3.1 Specifications 11

1.3.2 Identification and Limitations 16

1.4 Results 18

1.4.1 Overall Racial Disparities 18

1.4.2 Variation in Racial Disparities 23

1.4.3 Predicting Departmental Racial Disparities 24

1.5 Conclusion 26

2 Visual Inference And Graphical Representation in Regression Discontinuity

Designs **46**

2.1 Introduction 46

2.2 Conceptual Framework 52

2.3 Description of Experiments and Studies 58

2.3.1 Graphical Parameters Tested 58

2.3.2 Creation of Simulated Datasets and Graphs 62

2.3.3 Non-Expert Experiments 66

2.3.4 Expert Study 68

2.3.5 Eyetracking Study 70

2.4 Results 71

2.4.1 RD Non-Expert Experiment Results 71

2.4.2 RK Non-Expert Experiment Results 73

2.4.3 Expert Study Results 74

2.4.4 Visual vs Econometric Inference 76

2.4.5 Eyetracking Study Results 83

2.5 Conclusion 85

3 Migration and Local Minimum Wages **94**

3.1 Introduction 94

3.2 Data and Institutional Settings 98

3.2.1 Local Minimum Wages 98

3.2.2 IRS Statistics of Income 100

3.2.3 Additional Data 101

3.2.4 Summary Statistics 102

3.3 Empirical Strategy 103

3.3.1 Coarsened Exact Matching 103

3.3.2 Two-Way Fixed Effects and Callaway and Sant'Anna (2020) Difference-in-Differences 104

3.4 Results . . . 107
3.4.1 Match Quality . . . 107
3.4.2 Migration Results . . . 108
3.5 Discussion . . . 111
3.6 Conclusion . . . 114

A Appendix for Chapter 1 **132**
A.1 Data Cleaning . . . 132
A.2 Results with Full Set of Force Outcomes . . . 146
A.3 Results with OLS Estimates of Departmental Disparities . . . 155

B Appendix for Chapter 2 **162**
B.1 Theory: Estimator Properties, Bin Selectors, and MSE Decompositions . . . 162
B.1.1 Properties of the Type I and Type II Error Probability Estimators . . . 162
B.1.2 Calonico, Cattaneo, and Titiunik (2015) Bin Selection Algorithms . . . 166
B.1.3 Mean Squared Error Decompositions . . . 167
B.2 Additional Details on DGP Specification and Graph Creation . . . 168
B.3 Design of Experiments and Studies . . . 172
B.3.1 Non-Expert Experiment . . . 172
B.3.2 Expert Study . . . 176
B.3.3 Eyetracking Study . . . 178
B.4 Subject Characteristics and Performance Predictors . . . 179
B.4.1 Demographic Balance and Effects . . . 179
B.4.2 RD DGP Characteristics . . . 180
B.4.3 Dynamic Visual Inference . . . 181
B.5 What t-Statistic Can Eyes Detect? . . . 183
B.6 Combining Visual and Econometric Inferences . . . 184
B.7 Eyetracking Study: Details and Additional Analyses . . . 186

B.8 Appendix Tables 191
B.9 Appendix Figures 209

C Appendix for Chapter 3 **243**

C.1 Additional Estimates 243
C.2 TWFE Estimates 246

Bibliography **257**

List of Tables

1.1 Summary Statistics for Force Reports 29
1.2 Effect of Female Subject on Police Use of Force of At Least Specified Severity, Conditional on Force 30
1.3 Effect of Being in Department with Above-Median Empirical Bayes Racial Disparity, Officers Switching Departments Only 31
1.4 Summary Statistics for Empirical Bayes Department x Black/Hispanic Subject Interactions 32
1.5 Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Variables Normalized by 2x Standard Deviation . . 33
1.6 Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Numeric Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only 34

3.1 Descriptive Statistics for In-Sample County Minimum Wages 117
3.2 Summary Statistics for Treated and Adjacent Counties and Matched Control Counties 118
3.3 Calloway and Sant'Anna Difference-in-Differences Estimates, Total Migration 119
3.4 Calloway and Sant'Anna Difference-in-Differences Estimates, Foreign Migration 120

3.5 Calloway and Sant'Anna Difference-in-Differences Estimates, Same State Migration . . . 121
3.6 Calloway and Sant'Anna Difference-in-Differences Estimates, Different State Migration . . . 122
3.7 Calloway and Sant'Anna Difference-in-Differences Estimates, Neighboring Counties Migration . . . 123
3.8 Additional Calloway and Sant'Anna Difference-in-Differences Estimates . 124

A1 Summary Statistics for Municipalities Represented in Use of Force Data . . 135
A2 Summary Statistics for Police Departments Represented in Use of Force Data 136
A3 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force (No Controls) . . . 137
A4 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force . . . 137
A5 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force . . . 138
A6 Conditional Logit Odds Ratios on Race in Intensity of Force Used, Conditional on Force . . . 138
A7 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject and Subject at Least Physically Threatened/Attacked Officer or Another . . . 139
A8 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject . . . 140
A9 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Subject at Least Physically Threatened/Attacked Officer or Another . . . 140

A10 Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Numeric Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only 141
A11 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force (No Controls) 146
A12 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force 147
A13 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force 148
A14 Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject and Subject at Least Physically Threatened/Attacked Officer or Another 149
A15 Summary Statistics for OLS Estimates of Department x Black/Hispanic Subject Interactions 155
A16 Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation 156
A17 Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation 157
A18 Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only 158

B1 Data-Driven Global Polynomial Orders for Fit Lines 191
B2 Timeline of RDD Experiments 192
B3 Timeline of RKD Experiments 193
B4 Treatment Effects: Phase RDD1 194
B5 Treatment Effects: Phase RDD2 195

B6 Treatment Effects: Phase RDD3 196
B7 Treatment Effects: Phase RDD4 197
B8 Treatment Effects: Phase RKD (Levels) 198
B9 Treatment Effects: Phase RKD (Derivatives) 199
B10 Summary Measures of Eyegaze Data 200
B11 Timeline of Expert Study 201
B12 Covariate Means and Balance across Treatment Arms: Phase RDD1 201
B13 Covariate Means and Balance across Treatment Arms: Phase RDD2 202
B14 Covariate Means and Balance across Treatment Arms: Phase RDD3 202
B15 Covariate Balance across Treatment Arms: Phase RDD4 203
B16 Covariate Means and Balance across Treatment Arms: Phase RKD 203
B17 Correlations of Correct Classification with Subject Covariates for RD . . . 204
B18 Correlations of Correct Classification with Subject Covariates for RK . . . 205
B19 Predictions of Power on DGP Characteristics 206
B20 Predictions of Size on DGP Characteristics 206
B21 Dynamic Discontinuity Classification: AR(1) Regression 207
B22 Fisher's Exact Test of Association: p-values (One-Sided) 208
B23 Timeline of Eyetracking Study 208

C1 Calloway and Sant'Anna Difference-in-Differences Estimates, Domestic Migration 244
C2 Matched Difference-in-Differences Estimates, Total Migration 246
C3 Matched Difference-in-Differences Estimates, Domestic Migration 247
C4 Matched Difference-in-Differences Estimates, Foreign Migration 248
C5 Matched Difference-in-Differences Estimates, Same State Migration 249
C6 Matched Difference-in-Differences Estimates, Different State Migration . . 250
C7 Matched Difference-in-Differences Estimates, Neighboring Counties Migration 251

C8 Additional Matched Difference-in-Differences Estimates 252

List of Figures

1.1 Overall Racial Disparities (No Controls) 35
1.2 Overall Racial Disparities for Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity 36
1.3 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Department Fixed Effects and Clustering) 37
1.4 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Officer Fixed Effects and Clustering) 38
1.5 Conditional Logit Odds Ratios of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity 39
1.6 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject and Subject at Least Physically Threatened/Attacked Officer or Another) 40
1.7 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity by Year 41
1.8 Distribution of Empirical Bayes Estimates of Department-Black/Hispanic Interactions 42

1.9 Within-Department Correlations of Winsorized Empirical Bayes Estimates of Department-Black/Hispanic Interactions Across Force Types 43
1.10 Coefficients from Regression of Empirical Bayes Estimates of Department-Black/Hispanic Interactions on Possible Correlates 44
1.11 Coefficients from Regression of Empirical Bayes Estimates of Department-Black/Hispanic Interactions on Possible Correlates, Departments in 2016 LEMAS Only 45

2.1 RDD Phases: Power Functions 89
2.2 RKD Phase: Power Functions 90
2.3 Expert vs Non-Expert Performance 91
2.4 RDD: Expert Visual vs Econometric Inference 92
2.5 RKD: Expert Visual vs Econometric Inference 93

3.1 Maps of Treated and Neighboring Counties 125
3.2 CS DiD Estimates: Totals 126
3.3 CS DiD Estimates: Foreign 127
3.4 CS DiD Estimates: Same State 128
3.5 CS DiD Estimates: Different State 129
3.6 CS DiD Estimates: Neighboring Counties 130
3.7 Matched DiD Estimates: Log Population, Non-Mover Income, and Home Prices 131

A1 New Jersey Model Use of Force Report 142
A2 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject) 143

A3 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Subject at Least Physically Threatened/Attacked Officer or Another) 144
A4 Heatmap of of Winsorized Empirical Bayes Estimates of Department-Black/Hispanic Interactions 145
A5 Overall Racial Disparities (No Controls) 150
A6 Overall Racial Disparities for Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity 151
A7 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Department Fixed Effects and Clustering) 152
A8 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Officer Fixed Effects and Clustering) 153
A9 Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject and Subject at Least Physically Threatened/Attacked Officer or Another) 154
A10 Distribution of Unshrunk Estimates of Department-Black/Hispanic Interactions 159
A11 Coefficients from Regression of OLS Estimates of Department-Black/Hispanic Interactions on Possible Correlates 160
A12 Coefficients from Regression of OLS Estimates of Department-Black/Hispanic Interactions on Possible Correlates, Departments in 2016 LEMAS Only . . 161

B1 Same Data with Different Axis Scales for Correlation Detection 209
B2 Original Plot vs Plot with Residualized Y 209
B3 Original Plot vs Plot of Estimated First Derivatives 210

B4 RKD: Even Spacing vs Quantile Spacing Example 210
B5 RDD DGP Conditional Expectation Functions 211
B6 Creation of RDD DGP Conditional Expectation Functions, DGP9 211
B7 RKD DGP Conditional Expectation Functions 212
B8 Lineup Protocol Graph: RD DGP10 213
B9 Lineup Protocol Graph: RD DGP3 214
B10 RDD Phases: Power Functions by DGP 215
B11 RKD Phase: Power Functions By Quintic Type 216
B12 RD: Expert vs Non-Expert Performance Differences 216
B13 RK: Expert vs Non-Expert Performance Differences 217
B14 RD: Expert Preferences and Beliefs regarding Non-Expert Performance . . 218
B15 Phase RDD4: Power Functions for DGP Shown to Experts 219
B16 RK: Expert Preferences and Beliefs regarding Non-Expert Performance . . 220
B17 RD: Expert Visual vs Econometric Inference Power Function Differences . 221
B18 RD: Expert Visual vs CCT Procedure with Inference-Optimal Bandwidth Power Functions 222
B19 RD: Expert Visual vs Size-Adjusted Econometric Inference Power Function Differences 222
B20 RMSE of Point Estimates and MSE Decomposition by DGP 223
B21 RK: Expert Visual vs Econometric Inferences Power Function Differences . 224
B22 Lineup Protocols for DGPs Specified Before and After Adding Noise to the Running Variable 224
B23 Comparison between CEFs from Original Microdata and Adding Noise to Running Variable 225
B24 Video Tutorial Attention Check 225
B25 Example Tasks 226
B26 Example 4 - Feedback and Navigation Buttons 227

B27 Classification Screen . . . 228
B28 Results Screen Example . . . 229
B29 Sequence of Events - RDD Experiments . . . 229
B30 Sequence of Events - RKD Experiments . . . 229
B31 Sequence of Events - Expert Study . . . 230
B32 Expert Survey - Part 1 Introduction . . . 230
B33 Expert Survey - Part 1 Classification Task . . . 231
B34 Expert Survey - Part 2 Instructions . . . 232
B35 Expert Survey - Part 2 Decision Screen . . . 233
B36 Sequence of Events - Eyetracking Study . . . 233
B37 Sequence of Screens for Each Classification Task on Eyetracking Terminal 234
B38 RDD DGP Histograms . . . 235
B39 Comparison between Power Functions for Phase RDD4 by Symmetry of DGP Supports . . . 235
B40 DGP Gini Coefficients and Non-Expert Performance . . . 236
B41 Discontinuity Easier to Detect When Both Slopes at Cutoff are in the Same Direction as the Discontinuity . . . 236
B42 RDD Discontinuity Classifications over 11 Graphs . . . 237
B43 RDD Phases: Power Functions against Asymptotic t-Statistics . . . 238
B44 RDD Phases: Power Functions by DGP against Asymptotic t-Statistics . . . 239
B45 Visual Representation of Eyetracking Data . . . 240
B46 Penalized Logit Coefficient Path Plots . . . 241
B47 Bagging and Random Forest Importance Plots . . . 242

C1 CS DiD Estimates: Domestic . . . 245
C2 Matched DiD Estimates: Totals . . . 253
C3 Matched DiD Estimates: Domestic . . . 253
C4 Matched DiD Estimates: Foreign . . . 254

C5 Matched DiD Estimates: Same State 254
C6 Matched DiD Estimates: Different State 255
C7 Matched DiD Estimates: Neighboring Counties 255
C8 Matched DiD Estimates: Log Population and Non-Mover Income 256

Chapter 1

Variation in Racial Disparities in Police Use of Force[1]

1.1 Introduction

Racial inequality in police use of force is among the most important policy issues in the United States, and calls for reform grow louder with each unarmed person of color killed by police. As society seeks to address these race gaps, it is paramount that we know the extent to which this problem varies. Optimal interventions may not be uniform for different departments or force levels, and if some departments have smaller or no racial differences, they may provide insights on how to improve the others. Unfortunately, data and analytical limitations often require researchers to treat racial disparities as homogeneous in their analyses, limiting our understanding of the problem.

I explore three aspects of how racial disparities in police use of force vary. First, how do overall racial disparities in use of force change along the spectrum of force severity?

[1]I am grateful to Jesse Bruhn, Leah Boustan, Janet Currie, Milena Djourelova, Thomas Fujiwara, Bo Honoré, Ilyana Kuziemko, David Lee, Alex Mas, Jonathan Mummolo, Amy Wickett, seminar participants at Princeton University and the Young Economists Symposium, and members of the Princeton Police Department for helpful comments. The Princeton University Industrial Relations Section provided generous financial support. A preliminary version of this work was presented in summer 2020 at the Young Economists Symposium.

Second, how do these racial differences vary across and within departments? And third, what predicts these department-specific racial disparities?

I answer these questions using new administrative data from New Jersey. The data are unique in their completeness, containing every use of force report by every officer in every municipal police department and the state police over five years. Notably, they include the entire force spectrum, from placing the subject in a compliance hold to discharging a firearm. Further, because the data cover the entire state, they feature substantial variation in locations, with more than 450 police departments serving over 550 municipalities ranging from small, racially homogeneous towns to dense, diverse cities.

For the question of how overall racial disparities vary over the force spectrum, I focus on the intensive margin of force: conditional on force being used and adjusting for event-level observables such as incident type and subject actions, do people of different races have different levels of force used against them? This parameter is empirically and politically relevant, but has not received much attention by researchers. Cases of police officers using force viewed as inordinate against people of color continue to occur, and the unexplained racial variation in police use of force after adjusting for incident factors corresponds with those perceived excesses. I find significant disparities in police use of force across races that are larger when considering more severe levels of force. Compared to outcome means, Blacks and Hispanics are about 3% more likely than Whites and Asians/Pacific Islanders to have force more severe than compliance holds, the lowest level of force, used against them, while they are about 19% more likely when looking at enhanced types of force: pepper spray, batons, canines, stun guns, and firearms. These gaps are driven primarily by large disparities that disfavor Blacks. Additional analyses point to the importance of the department over the identity of the officer in these disparities.

To minimize the bias from selection into my data as I address this question, I build on the techniques of Grogger and Ridgeway (2006) and Horrace and Rohlin (2016) in identifying a set of incidents that plausibly reduces the effects of race outside of the decision

of what level of force to use, in particular the officer's decisions of whom to investigate or engage with and whether they should use force at all. These papers are built around the "veil of darkness" hypothesis, arguing that in poor lighting (such as nighttime traffic stops), a subject's race is harder to observe, and hence less likely to factor into the officer's decision to stop a vehicle. I extend this method by restricting the dataset to "always stop" incidents, those in which the subject likely would have been stopped regardless of their race (crimes in progress, disputes, and traffic stops at night), and incidents where the subject at least physically threatened or attacked an officer or another individual, where the outside option of not using any force is less viable. Both of these restrictions should limit the role of race in the officer's decisions whether to engage with a subject or use force at all while preserving interactions between race and the intensity of force used. The results from this analysis are similar to the full-sample estimates, suggesting that selection bias prior to observation may not be a major problem in my setting and bolstering my estimates.

Next, I leverage New Jersey's large number of police departments to explore how racial disparities vary across and within departments. To better conduct this analysis, I adapt empirical Bayes techniques to estimate departmental group-specific differences. Using this new estimator, I find that the overall disparities mask large variation in these disparities across departments that often dwarfs the overall estimated magnitudes. A minority of more than 20% of departments have zero or negative racial differences for a given force level, but there is also a long right tail of departments where Black and Hispanic subjects face disparities significantly larger than the overall estimates. There is substantial variation in disparities within departments across force levels, and within-department correlations are near zero.

Finally, I treat the estimated departmental disparities as the quantities of interest themselves and explore whether departmental or municipality characteristics such as officer diversity, economic inequality, violent crime rates, and political preferences are able to predict these racial gaps through a "horse race" analysis. I find that commonly proposed

factors such as racial diversity do not correlate with the disparities, and these models cannot explain most of the variation in the data. When looking at the smaller subset of departments present in the 2016 Law Enforcement Management and Administrative Statistics (LEMAS) survey, the cultural and policy measures available there are more strongly correlated, but overall explanatory power is still low. Although harder-to-measure factors such as culture may play an important role, the variation across and within departments underlines the difficulty in addressing racial disparities in police use of force: there is no single factor that universally predicts the gaps.

Although there is relatively little work about variation in racial differences in police use of force compared to overall gaps, economists and other social scientists have long grappled with the question of how race interacts with the criminal justice system. Recent efforts have made progress on this methodologically thorny issue using new data or empirical strategies to obtain better estimates. Nix et al. (2017) examine fatal police shootings using novel data collected by *The Washington Post* and find racial differences in whether a shot subject was attacking the officer and whether they were armed. Weisburst (2019) uses data from the Dallas Police Department and finds that Black civilians are disproportionately likely to be involved in an incident involving any level of force, stemming from differences in likelihood of arrest. Fryer (2019) estimates overall disparities by combining several different data sources, including New York City's Stop, Question, and Frisk program and officer-involved shootings from 16 police departments. He finds racial disparities in police use of force for Blacks relative to Whites to be consistent over the nonlethal force spectrum with odds ratios of around 1.2, while Hispanic-White disparities decrease as the force level increases; he also finds no disparities for police shootings. Knox, Lowe, and Mummolo (2020) argue that the estimates from Fryer (2019) understate the causal effect of race by not accounting for racial differences in police-civilian interactions prior to the encounter, such as patrolling habits and the decision of whether to stop a subject.[2] Many other papers

[2]Also see comments by Durlauf and Heckman (2020) and the response by Fryer (2020).

explore related aspects of race and the criminal justice system.[3] This paper differs from prior work by focusing on *variation* in racial differences, specifically by force severity and departments, rather than treating race as having a single, homogeneous effect.

There is a small, highly interdisciplinary literature on variation in racial disparities that asks what can predict or explain the variation, offering different answers. Terrill and Reisig (2003), for example, use data collected on police ride-alongs to explore how "neighborhood context" such as economic status and homicide rate affects police use of force and find that the setting of an incident matters greatly and mediates the overall effect of a subject's race. Ross (2015), in the project most similar to this aspect of my work, conducts an analysis of county-level police shootings with local racial populations as the benchmark, finding significant variation in risk ratios at this level of aggregation associated with "larger metropolitan counties with low median incomes and a sizable portion of Black residents, especially when there is high financial inequality." Hoekstra and Sloan (2020) exploit plausibly exogenous variation from 911 dispatch assignment rules from two cities and observe that White officers increase their use of force more when dispatched to more minority neighborhoods than minority officers do, also finding that racial differences are more severe for firearms than other force types. The variation present in my setting and the new empirical Bayes estimator are key to this paper's value-added, as they allow me to explore how racial differences vary in a unified, expansive setting with detailed microdata on every level of force and hundreds of departments.

In identifying these variations in racial disparities, I move forward our understanding of race and policing. Beyond simply observing heterogeneity in police use of force, the variation present in department-level disparities underscores the difficulties in remedying the differences and how a one-size-fits-all intervention may not be appropriate. Some depart-

[3]Examples include police stops and searches (Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Persico and Todd, 2006; Antonovics and Knight, 2009; Coviello and Persico, 2015; West, 2018), speeding tickets (Goncalves and Mello, 2017), police reform after racially unjust incidents (Heaton, 2010; Devi and Fryer, 2020; Luh, 2020), whether subjects are formally booked (Raphael and Rozo, 2019), court decisions (Rehavi and Starr, 2014; Arnold, Dobbie, and Yang, 2018; Bielen, Marneffe, and Mocan, 2019), and student discipline (Barrett et al., 2019).

ments do not appear to have racial differences against Blacks and Hispanics, but many more do, some very severely, and the identities of these departments change across force levels. Large amounts of unexplained variation in these differences, even with a rich set of covariates on the right-hand side, coupled with other estimates emphasizing the importance of the department, may also suggest that difficult-to-measure factors such as department cultures play an important role in determining how race and police use of force interact.

My results offer new insights to the literature on racial differences in policing, but it is worth noting several caveats. First, I examine the intensive margin of force: whether, conditional on force being used, subjects of certain races have more severe types of force used against them. As in many studies, the extensive margin of whether force is used at all cannot be analyzed without implausibly strong assumptions about how race affects the number and type of police-civilian interactions (such as assuming that it does not). Second, the data I analyze come from the police themselves, and it is conceivable that some reports are withheld or contain misinformation. I discuss this concern in detail when outlining my analytical strategy, noting in particular that crowdsourcing efforts have not indicated widespread suppression of incidents and that incentives for misreporting were hugely diminished by the fact that planned central oversight never materialized. Third, although I use comprehensive data from New Jersey over a five-year period, my results are a result of who is (and is not) in that sample, that is, who has force used against them at all. I take steps to minimize selection bias, but as I discuss later, depending on how cases are selected, one could find different results. Nevertheless, the patterns I do find are clear evidence on both the existence and variation of racial differences in police use of force.

1.2 Institutional Background and Data

1.2.1 Use of Force and Force Reporting in New Jersey

In 2001, the New Jersey Attorney General's Office began to require that police officers

document all uses of force in the line of duty. Figure A1 shows a model form for these force reports. Although not every department seems to follow this template exactly, the information gathered is largely the same across departments. Plans for a centralized system for analysis and oversight by the state never materialized, and most of these reports ended up in storage, unused and inaccessible by the public (NJ Advance Media, 2019).[4]

Police in New Jersey are authorized to use several types of force in the line of duty. These standard force types police use can be ordered along a spectrum of severity. The lowest level of force is compliance holds, such as arm bars and wrist locks. Next are unenhanced types of force: takedowns (forcing a subject to the ground), hands/fists (punches and slaps), and kicks/other leg strikes. Then there are enhanced force outcomes: pepper spray and other chemical agents, baton strikes, canines, stun guns (Tasers), and finally, discharging firearms. Constructive authority, the threat of force without its actual use, such as brandishing a firearm, is permitted, but warning shots are prohibited. Stun guns are mostly absent from the data due to state regulations against their use, though departments began rolling them out after a change in the law in 2016, and canines are also uncommon, as not all departments have canine units. Police may very rarely employ nonlethal firearm rounds, such as beanbag rounds, in settings such as riots to incapacitate subjects, and I drop these incidents, as the situations in which they are used are not representative of typical police-civilian interactions.

1.2.2 Data

This project uses all known force reports from every municipal police department in New Jersey and the New Jersey State Police from 2012 through 2016. ProPublica, a nonprofit newsroom, and NJ Advance Media (NJAM), a news provider, obtained the reports through more than 500 public records requests and several legal threats. Following substantial data entry and cleaning, ProPublica and NJAM made the resulting dataset available for purchase

[4]In the wake of the murder of George Floyd by police in Minneapolis, the New Jersey Attorney General announced that a statewide use of force portal would be ready later in 2020.

in January 2019.

These reports contain rich incident-level information. Among other variables, the reports record the time, date, and location of the incident, the nature of the incident (such as a crime in progress or traffic stop), the officer(s) involved, the subject(s) involved and their actions that led to force being used, officer identities and demographics, the types of force an officer used against the subject, and whether an officer or subject was injured or killed.

Despite the efforts by ProPublica, NJAM, and their partner data entry firm, the final dataset required additional processing before I could use it in my analysis. For example, some types of force used by police are not neatly categorized, and instead consist of irregular descriptions such as "grabbed rock out of her hand" or "in foot pursuit grabbed suspect left hand." Appendix A.1 documents how I clean and process the data. I structure the data so that each observation represents one subject who had force used against them by one officer in an incident. For cases where multiple officers use force against a single subject, I keep the officer who uses the greatest level of force, choosing randomly in the event of ties.[5] I remove 44 subjects whose indicated races do not fall within the categories of White, Black, Hispanic, or Asian/Pacific Islander, such as people marked as "mixed." After cleaning, there are 39,322 incidents that I use in my analysis.

I supplement the force reports with information on New Jersey's police departments and municipalities. Characteristics of the police departments themselves, such as racial diversity, are from ProPublica and NJAM. Local variables come from the 2010 Census, 2012-2016 American Community Survey (ACS) five-year estimates, FBI's Uniform Crime Reporting (UCR) program, the New Jersey Division of Elections, and Project Implicit. Additional data on a smaller sample of departments comes from the 2016 LEMAS survey conducted by the Department of Justice (Bureau of Justice Statistics, 2016).

[5]There are several justifiable ways to structure the data. I choose this format, as it is most natural for my empirical strategy in which the outcome of each incident is a function of the greatest level of force used. When multiple officers use force, it is then sensible to keep the officer who used the greatest level of force, especially because other officers may only able to use lower levels of force because another officer used a more severe type (for example, after Officer A pepper sprays a subject, Officer B is able to place him in a compliance hold).

1.2.3 Summary Statistics

Table 1.1 presents summary statistics for the force reports in the dataset after cleaning. It includes the most extreme type of force used in each incident, the most extreme action the subject took that may have prompted the force, whether an officer was injured, the type of incident, and subject demographics. In 50% of incidents in the sample, officers use only compliance holds, the lowest type of force available. In 25% of incidents, the most severe action officers take is striking subjects with their hands or fists, followed by 11% with pepper spray as their highest force level. All other types of force are relatively uncommon as maximums. Officers discharge firearms in under half a percent of observations, representing over 160 shootings between 2012 and 2016.

Subject actions are concentrated on the lower end of the spectrum. Resisting an officer, which I also use as a catchall for non-missing behaviors that do not fit elsewhere, is the most extreme behavior in 63% of incidents. Physically threatening or attacking an officer, for example punching, kicking, or spitting, follows with 34% of observations. The remaining 3% of the data is divided approximately evenly between blunt weapon, knife, and vehicle threats/attacks, as well as threats with firearms. Only 0.1% of incidents involve a subject actually discharging a firearm. 10% of force reports indicate that an officer was injured.

Incident types are varied. I follow the model force report from the New Jersey Attorney General and include indicators for crimes in progress, domestic disputes, other disputes, suspicious persons, traffic stops, and other incidents. An incident may have multiple types, except for the "other" category, which I reserve for incidents that are not classified as belonging to any other category. Consistent with the sample being only incidents in which officers used force, the most common incident classification is crime in progress, making up more than a quarter of the data. Domestic disputes are the next most frequent at 13% of the sample. Other disputes and suspicious persons are each 11% of the sample, traffic stops are 9%, and 33% of incidents fall into the "other" category.

Subject demographics are not representative of New Jersey's population. 48% of the subjects in the sample are White, 41% are Black, 10% are Hispanic, and 1% are Asian/Pacific Islander. From the ACS five-year estimates for 2012-2016, 57% of the state's population at the time was non-Hispanic White and 9% was non-Hispanic Asian/Pacific Islander, making these two groups underrepresented in the force reports. Blacks are severely overrepresented, with non-Hispanic Blacks comprising only 13% of the population. Due to data limitations, I treat Hispanic status as a distinct racial category; one cannot be Hispanic and another race in the data.[6] Note that officers use their own judgment when recording a subject's race. 19% of the state population was Hispanic of any race, but it is impossible to know whether they are overrepresented or underrepresented in the force data, as the criteria used to classify individuals as Hispanic in the force reports are not the same as those used in the ACS. Only 20% of subjects are female. The average subject is 31 years old.

Table A1 presents summary statistics for the municipalities with their own police departments that are present in the cleaned data (some small municipalities lack dedicated police departments; the New Jersey State Police cover the state's highways and may also serve these towns). For New Jersey's approximately 565 municipalities, 461 police departments are present in the raw data. A handful of small municipalities with a combined 58 force reports do not appear in my subsequent analyses because their reports are missing data in relevant variables.

New Jersey's hundreds of municipalities offer significant heterogeneity over margins such as race, income, size, and political preferences over which policing might vary. Populations for municipalities in the sample range from hundreds to hundreds of thousands. As of the 2010 Census, New Jersey has seven of the 10 densest incorporated places in the United States, and it is overall the densest state in the country, though this too varies greatly within the state. New Jersey is among the richest states by median income, but there are

[6]Some departments' force reports do distinguish between a subject's race and Hispanic status, but many do not. The cleaned version of the race variable in the dataset made available to researchers treats Hispanic status as a race.

significant areas of poverty and it has high economic inequality as measured by the Gini coefficient relative to other states. The adjusted population figures shown in Table A1 are for the population ages 18 through 65 in each municipality. Although New Jersey is mostly White, many areas have barely any Whites while others are almost exclusively White. Violent crime rates range from 0.03 to 55 crimes per 1000 people. Political preferences, as proxied by Mitt Romney's share of the vote in the 2012 presidential election, are consistent with the state leaning Democratic overall despite the presence of more conservative areas. County-level Black-White implicit association test (IAT) *D*-scores from Project Implicit, where more positive values indicate stronger negative implicit biases against Blacks, show large variation: the standard deviation of 0.05 is approximately equal to the difference in state-level means between Oregon and North Dakota.[7]

Table A2 contains summary statistics for the police departments in the data. The median department has 23 full-time employees, the average is almost 43, and the maximum is over 1,000. Racial diversity on the whole is poor: the median department's officers are more than 90% White, 0% Black, 0% Hispanic, and 0% Asian/Pacific Islander (not all departments' racial breakdowns sum to 100% due to inconsistencies in reporting).

1.3 Empirical Strategy

1.3.1 Specifications

Overall Disparities

For the question of overall racial disparities, my primary econometric specification estimates the following equation via ordinary least squares (OLS):

$$Force_{iopt} = \beta \cdot \mathrm{BH}_i + X'_{iopt}\gamma + \psi_p + \nu_t + \varepsilon_{iopt} \tag{1.1}$$

[7] `https://github.com/lizredford/map-iat`

where subscripts i, o, p, and t denote the incident, officer, department, and year, respectively. *Force* is a binary "at least this severe" measure of the extent of the force used in an incident. I use four levels of force severity: compliance holds, unenhanced forced, enhanced force, and firing a weapon (which is a subset of enhanced force types).[8] For example, if a subject is struck with a baton (a type of enhanced force), the outcome would be 1 for at least compliance holds, unenhanced, and enhanced, but 0 for at fired weapons. I use this measure of force because it is the most intuitive way to interpret outcomes. Alternative parameterizations such as the maximum or minimum force level used incompletely describe an event, and analyzing every force level separately can result in estimates without clear interpretations (Fryer, 2019). The coefficient of interest is β, the difference in the observed probabilities of Black or Hispanic (BH) subjects having more severe types of force used against them conditional on any force relative to the reference group (Whites and Asians/Pacific Islanders). I primarily use a binary race indicator for a subject being Black or Hispanic to improve statistical power, but also estimate models with a full set of race indicators. X is a vector of incident-level characteristics including time, type of incident, officer rank, subject behaviors, subject sex, and a quadratic of the subject's age. Because incidents may have multiple types, I include indicators for each unique combination of types rather than each individual type; the effect of an incident that is a crime in progress and a traffic stop is not the same as the sum of the crime in progress and traffic stop effects. Department fixed effects ψ_p capture time-invariant aspects of each department's propensity to use higher levels of force, such as the severity of crimes on a department's beats. Time fixed effects ν_t adjust for year-month specific changes in overall force usage, capturing trends such as seasonality in crime. I cluster standard errors at the department level.

Although I am able identify many department and officer fixed effects simultaneously due to officers switching departments, including officer fixed effects requires dropping

[8]Results in Appendix A.2 use the full set of outcomes ordered by the NJ Attorney General's form and reporting by McCarthy and Nelson (2019): compliance holds, takedowns, hands/fists, leg strikes, pepper spray, batons, and firearms. I group canines and stun guns are with batons, the next most severe level of force, due to their rarity in the data.

many variables, which will be important for later analyses. Instead, I estimate an additional series of models using officer fixed effects instead of department fixed effects with clustering at the officer level. These estimates allow me to investigate how disparities vary when adjusting for the officer's identity, which may be useful if there is significant heterogeneity in officer-level disparities within departments. Further, comparing these results to the ones using department fixed effects allows me to observe the role of the individual officer as opposed to the department. The officer-based models could hypothetically be sensitive to the data cleaning process in which, for each incident where multiple officers use the most severe level of force in that incident, I keep one randomly chosen force report. However, this is an edge case, and results with a different draw are similar.

As a robustness check, in addition to the OLS estimates from Equation 1.1, I estimate conditional logit models of the following form:

$$\ln\left(\frac{Pr(Force_{iopt}=1)}{1-Pr(Force_{iopt}=1)}\right)=\beta\cdot \mathrm{BH}_i+X'_{iopt}\gamma+\psi_p+\nu_t+\varepsilon_{iopt} \tag{1.2}$$

stratified by department. To prevent separation, I do not include officer fixed effects, use year fixed effects instead of year-month, use event type indicators for each individual event type rather than each combination of event types, and change the officer rank indicators to a dummy for the officer being a superior or not. By exponentiating β, I obtain odds ratios for Black and Hispanic subjects compared to White and Asian/Pacific Islander ones. Logit-based estimators offer several advantages over OLS, in particular probabilities bounded between 0 and 1, useful for rare events like shootings, and conditional logit further addresses issues surrounding the inconsistency of logit with numerous fixed effects. However, as I move up the force spectrum, separation becomes a larger concern. More departments will have only zeros in the outcome, and these observations must be completely dropped from the analysis.

Department-Specific Disparities

Next, I examine variation in racial disparities across and within departments. The simplest approach would be to make a slight modification to Equation 1.1 and estimate equations of the form

$$\text{Force}_{iopt} = \beta_p \cdot \text{BH}_i \times \text{Department}_p + X'_{iopt}\gamma + \psi_p + \nu_t + \varepsilon_{iopt} \tag{1.3}$$

with the interest being in the distribution of β_p. Although these estimates may be unbiased and consistent, OLS tends to generate the most extreme estimates for departments with the fewest observations and does not distinguish between estimates based on departments with many observations and those from much smaller samples.

To improve upon OLS for identifying these hundreds of related disparities, I modify empirical Bayes estimators, such as those commonly used in estimating teacher value-added (see, for example, Kane and Staiger, 2008; Chetty, Friedman, and Rockoff, 2014), to estimate these group-specific differences for each department. Empirical Bayes estimators use the overall distribution of estimates to inform each individual point estimate. Although these general techniques are not new (Morris, 1983), there is almost no work on extending this estimator to a setting where we are interested in estimates of a "treatment" on only a subset of the population, like a department's propensity to use more intense force against Black and Hispanic subjects relative to White and Asian/Pacific Islander ones, as opposed to overall effects (see related work by Kline and Walters, 2019 in the context of audit studies of discrimination).

As is standard practice, I fit a normal distribution for the prior and apply Bayesian updating to obtain a posterior distribution for each department's estimate. Less reliable estimates, such as those from departments with few observations, are shifted towards the population mean, resulting in a "shrinkage" estimator. After updating, I record the centers of the posterior distributions and use them as the estimates.

For each level of force, I begin by estimating the following "pooled" regression:

$$\text{Force}_{iopt} = \beta_0 \cdot \text{BH}_i + X'_{iopt}\gamma + \psi_p + \nu_t + u_{iopt} \tag{1.4}$$

I subsequently use the following distribution as the prior

$$\beta_p \sim N(\hat{\beta}_0, \sigma_p^2)$$

where $\sigma_p^2 \equiv \text{Var}(u_{iopt}^{BH} - \varepsilon_{iopt}^{BH})$, the variance of the difference between residuals between Equations 1.3 and 1.4 using only Black and Hispanic observations rather than the full sample.[9] I then take the estimated racial disparities β_p from Equation 1.3 and compute the empirical Bayes estimates

$$\hat{\beta}_{p,EB} = w_p \cdot \hat{\beta}_p + (1 - w_p) \cdot \hat{\beta}_0 \tag{1.5}$$

where

$$w_p \equiv \frac{\sigma_p^2}{\sigma_p^2 + \text{Var}(\varepsilon_p^{BH})/n_p^{BH}}$$

are the departmental empirical Bayes weights, ε_p^{BH} is the residual for an observation with a Black or Hispanic subject in department p (residuals for other subjects are omitted), and n_p^{BH} is the number of observations with Black or Hispanic subjects for department p.

Predicting Department-Specific Disparities

In this exercise, I treat the estimated department-specific disparities as the outcomes of interest and regress them on variables describing both the departments and the municipalities they serve.

$$\hat{\beta}_{p,EB} = X'_p\alpha + \varepsilon_p \tag{1.6}$$

[9] One could instead estimate σ_p^2 as $\text{Var}(u_{iopt}^{BH}) - \text{Var}(\varepsilon_{iopt}^{BH})$, but this may result in negative estimated variances. It does not in my case, and the results with either method are nearly identical.

X_p is a vector of department/municipality-level characteristics: log median household income, Gini coefficient, log population, log population density, log number of police, the percentage of officers who are not White or Asian/Pacific Islander, the percentage of the population ages 18 to 65 that are not White or Asian/Pacific Islander, the violent crime rate, Mitt Romney's vote share in the 2012 presidential election, and county-level *D*-scores for the Black-White IAT from Project Implicit. I repeat this exercise on the non-random subset of departments in the 2016 LEMAS survey using additional measures of department policies and culture plausibly correlated with racial disparities: asset forfeiture, the use of a written aptitude test in selecting recruits, having a written community policing plan, formally surveying local residents to improve policing, having written policies on stop and frisk, foot pursuits, and unbiased policing, and whether the department's chief executive is female or non-Hispanic White. These regressions provide suggestive results for some of the commonly hypothesized contributors to police violence or bias.

1.3.2 Identification and Limitations

I estimate disparities in the intensity of police use of force experienced by subjects of different races conditional on force and after adjusting for incident characteristics and subject behaviors. Here I discuss this parameter and outline the efforts I take to limit the effects of confounding factors in my analysis.

Every incident in the data in this paper undergoes two "treatments" prior to the officer's decision of what level of force to use. First is the decision to engage with a subject, such as whether to stop someone on the street or pull over a vehicle. Second is the extensive margin of whether to use force at all. Each of these treatments is plausibly affected by race, and there is evidence from similar settings that they may be (Gelman, Fagan, and Kiss, 2007; Fryer, 2019). Because my incidents take place after these decisions, selection bias in who is in the sample, that is, in who has force used against them at all, poses one problem. Analysis conditional on force being used may be biased if an officer's decision to use force

is dependent on a subject's race. Conditional on force, one might find no marginal effect of race, but that neglects to account for varying probabilities across race of interactions with police.

If officers engage with civilians of different races at different rates, this could affect force rates, and depending on the nature of the interactions, also the severity of force used.[10] If officers are overly suspicious of Blacks, there could be many unwarranted stops that end without force or with only compliance holds, of which only the latter would appear in my data. If Whites are then only stopped for committing violent offenses that require high levels of force from officers, my empirical strategy could be biased towards a disparity that disfavors *Whites*. Or it could be that officers patrol in a manner that makes them more likely to engage with Blacks committing violent crimes and Whites committing nonviolent offenses, which could threaten my results in the opposite direction. These problems are complicated by the presence of the second pre-observation treatment in my setting: the extensive margin of whether to use force at all.

Although it is impossible to determine ex post whether a particular stop or decision to use force was motivated by a subject's race, under milder assumptions, I can limit the role of racial disparities prior to the intensive margin of force severity. Using information on incident characteristics and subject actions, I repeat my analysis of overall disparities on a subset of observations where the subject's race was less likely to factor into the decisions whether to investigate and whether to use force. Specifically, I use the intersection of two subsets of the data, building on the ideas of Grogger and Ridgeway (2006) and Horrace and Rohlin (2016). First, I take crimes in progress, disputes, and traffic stops at night (8:00 PM through 5:59 AM), discarding suspicious person incidents, daytime traffic stops, and "other" incidents. Second, I restrict the sample to incidents where the subject at least physically threatened or attacked an officer or another individual, dropping incidents in which subjects only "resist." The former set should contain fewer racial stops and the latter

[10]This concept is related to the more general idea of differences in the marginal subject across races, discussed, for example, by Becker (1957).

should be missing fewer incidents where no force was used, limiting correlations between race and the error term and improving my estimates.

A final concern is that the data I use come from the police themselves, and officers could withhold or misreport information. NJAM has attempted to solve any such problems through a crowdsourcing effort (McCarthy, 2019). This has uncovered minor discrepancies in some reports, often around officer names, and the only missing reports uncovered are 70 from Jersey City (prior to cleaning, there are more than 1100 force reports from Jersey City in the data). Reporting by NJAM's NJ.com and a discussion with an active New Jersey police officer also indicate that there is some disagreement about when officers are required to fill out force reports. Because my coefficient of interest is the racial disparity after adjusting for incident characteristics, errors in the data such as missing reports would only bias my estimates if they are correlated with the subject's race. If officers misreport Black or Hispanic subjects as posing greater threats than they actually do in incidents, this would cause me to underestimate racial disparities. Note that incentives to misreport are greatly diminished by the lack of central oversight over this time period. Force reports existed mostly as physical copies, many only at the departments themselves despite guidelines that all force should be reported to county prosecutors, making external monitoring difficult (Nelson, 2019; McCarthy and Stirling, 2019).

1.4 Results

1.4.1 Overall Racial Disparities

Before estimating the full model, it is helpful to understand the racial disparities that exist without adjusting for any covariates. Figure 1.1 plots the disparities from regressing binary variables for the force used being at least as severe as the stated level on a set of indicators for the subject being Black, Hispanic, or Asian/Pacific Islander on top of the intercept (the mean for Whites). Table A3 contains the corresponding regression results. Black and

Hispanic subjects are much more likely to have higher levels of force used against them. These gaps increase with force severity, and their magnitudes are similar for Blacks and Hispanics within each level of force. Relative to outcome means, disparities for Blacks and Hispanics compared to Whites slightly under 20% for at least unenhanced, and 36% for Blacks and 22% for Hispanics for at least enhanced. Estimates in the fired weapons category are positive and small in absolute terms, but large when considering the rarity of these events. This result contrasts with those from Fryer (2019), who does not find racial disparities against Blacks and Hispanics even in the raw data. Asians/Pacific Islanders are somewhat less likely to have more severe force types used against them than people of other races, though these differences tend to be much smaller in magnitude than the ones for Blacks and Hispanics.

Next, I estimate racial disparities in the full model defined by Equation 1.1. Both sets of models treat race as a binary variable where the reference group is Whites and Asians/Pacific Islanders and the alternative group contains Blacks and Hispanics. Figure 1.2 plots the estimated racial disparities over outcome means, with the corresponding regression results in Table A4.

Examining these racial disparities reveals several patterns. First, the racial differences are diminished compared to the previous models that do not adjust for incident observables, though they are still present. For the models using department fixed effects and clustering, the disparities estimated for Black or Hispanic individuals for at least unenhanced, enhanced, and fired weapons are approximately 1.5, 2.8, and 0.1 percentage points, respectively, with all except the firearms level statistically significant at the 5% level. While the positive disparities for firing weapons have economically significant point estimates considering the rarity of the events, that rarity also greatly limits statistical power. Second, relative to the baseline level of each force type, racial disparities increase with the severity of the force type. In percentage terms, the department fixed effects-based disparities for Black or Hispanic subjects are 3.0%, 19.2%, and 25.0%, and this monotonicity is also

present when using the full set of force levels in Appendix A.2. Finally, the results are similar for models using department and officer fixed effects and clustering. Estimates from the latter tend to be slightly larger, with marginally less precision. The similarities of both models provide suggestive evidence that the department or municipality is a more important factor than the identify of the officer in the incident, a hypothesis I will explore more later in this paper.

Figures 1.3 and 1.4 and Table A5 present the estimated racial differences from Equation 1.1 with indicators for each race. Breaking apart the Black/Hispanic dummy, it becomes clear that the previous disparities are largely driven by Blacks. The coefficient on being Black in each regression is similar to but slightly larger than the Black or Hispanic estimates from the previous models. In contrast, I find much smaller disparities for Hispanics. For Asians/Pacific Islanders, point estimates are small and mostly negative. The group's point estimate of -0.5 percentage points for having a weapon fired at them with department fixed effects is enormous considering that only 0.4% of all incidents involve police shootings.

Figure 1.5 and Table A6 show the odds ratios from the conditional logit models in Equation 1.2. The results are qualitatively similar to the OLS ones, suggesting that the results are not dependent on the usage of linear probability models. Point estimates are positive and small at the lower levels of force, and they increase with the severity of the force, with the odds ratios for at least unenhanced, enhanced, and fired weapons being 1.05, 1.25, and 1.44, respectively.

As described in Section 1.3.2, to limit the role of racial differences in the unobserved decisions to stop a subject and use force against them, I estimate Equation 1.1 on the intersection of two subsets of the data. This new sample consists of incidents where a subject's race was less likely to affect the officer's decision whether to engage with the subject and incidents where the outside option of not using force was less feasible and contains about 7,800 incidents.

Figure 1.6 and Table A7 contain the results from regressions with this new sample. Note

that although the point estimates are larger in magnitude than their full sample analogs, outcome means are naturally higher here than in the full sample, as these incidents tend to be more severe than the dropped ones, necessitating higher levels of force. Relative to the new outcome means, the estimated racial disparities using department fixed effects represent increases of about 6.9% and 23.7% for at least unenhanced and enhanced force. These figures are similar to those using the full sample, though slightly larger. The point estimate for fired weapons is approximately 0. Standard errors increase severalfold due to the decreased sample size. The estimates using officer fixed effects are fairly similar to the ones based on department fixed effects, but their standard errors are about twice as large. These estimates may also be more fragile from having to estimate officer fixed effects on such a reduced sample. Both sets of estimates largely exhibit the same pattern from the full sample where the relative disparities increase with the severity of the force (including the results with full set of outcomes in Appendix A.2), excepting firing weapons. Figures A2 and A3 and Tables A8 and A9 contain estimates from each subset individually instead of their intersection, which remain similar. Overall, the similarities of these estimates using the subsetted data to their full data counterparts support the use of the entire dataset for subsequent exercises and may indicate that racial differences in events leading to the officer's decision about the intensive margin of force do not have large effects in this setting.

There may be reasons to expect disparities to shrink over the course of the data, such as increased scrutiny towards police (the shooting of Michael Brown in Ferguson, Missouri occurred in 2014) incentivizing misreporting or departments improving through learning and reducing excessive force against Black and Hispanic subjects. Figure 1.7 presents point estimates from the full model for the subject being Black/Hispanic interacted with the year using 2012 as a reference to test this idea. There is no compelling evidence of changes in racial disparities over time, casting doubt on these hypotheses.

Although it is exceedingly difficult to identify causal mechanisms with observational police data, we may wonder whether these disparities are the result of taste-based or statis-

tical discrimination, as the policy implications for treating each may be very different. Note that racially discriminatory policing is illegal in New Jersey and should not be considered justified, regardless of motivation. Statistical discrimination might manifest here due to a belief that, all else equal, Black or Hispanic subjects pose greater threats than subjects of other races, for example from aggressiveness or physical strength, prompting higher levels of force (these beliefs need not be accurate; see Bohren et al., 2020). Under the assumption that officers should not have this belief for female subjects, who tend to have lower levels of force used against them, interacting Black/Hispanic status with the female subject indicator can help to differentiate between channels of discrimination. A positive coefficient for the subject being Black/Hispanic is consistent with both statistical and taste-based discrimination. A negative coefficient on the interaction between Black/Hispanic and female would then suggest statistical discrimination, while a zero coefficient would be more consistent with taste-based discrimination. Table 1.2 shows that coefficients from this augmented full model on the subject being Black/Hispanic, the subject being female, and their interaction are positive, negative, and approximately zero, respectively, for both the at least unenhanced and enhanced models, supporting the taste-based discrimination hypothesis (I do not include the fired weapons outcome, as only five women are shot in my sample).

Given the similar results from models based on department and officer fixed effects, I now explore the relative importance of the department and officer more closely. On a set of 1041 incidents by 174 manually identified officers who switch departments, I fit Equation 1.1 with officer fixed effects and an additional indicator for the department in which an officer served having an above-median racial disparity for the relevant levels of force as estimated in Section 1.4.2.[11] Table 1.3 contains the results. For at least unenhanced force, the interaction between a subject being Black or Hispanic and the department having an above-median disparity is positive, large, and significant, supporting the idea that the department in which an officer serves is a larger determinant of racial disparities than the

[11]I do not analyze firearms here, as no switching officers use that level of force in my data. Additionally, I include year fixed effects instead of month-year due to the greatly reduced sample size.

identity of the officer. For at least enhanced force, the interaction is 0, possibly because more severe force is less discretionary, leaving less room for the new department to affect an officer's force intensity.

1.4.2 Variation in Racial Disparities

In this section, I focus on how racial disparities vary with departments. Overall racial disparities treat racial differences as monolithic, while they may vary both across and within departments. Furthermore, overall estimates place the most weight on the places with the most incidents, typically large, urban areas. As explained in Section 1.3, my preferred estimates of department-level racial disparities β_p come from the new empirical Bayes estimator in Equation 1.5. I leave results from the unshrunk OLS estimates to Appendix A.3 for conciseness.

Figure 1.8 presents kernel density estimates of the empirical Bayes posterior departmental disparities, with summary statistics in Table 1.4 (recall that the estimates are calculated without a main coefficient on race and are centered at the overall racial disparity, not 0). The distributions have long tails that often dwarf the overall racial disparities estimated before. As the level of force increases, the distributions get tighter due to the events becoming rarer, and thus the "confidence" in each point estimate is lower and the estimates are shrunk more towards the grand mean. There are some clear outliers, which are a small sample problem. In the enhanced subplot, for example, the point around 0.9 corresponds to Point Pleasant, which used force against only two Black or Hispanic individuals in my sample. These observations were part of the same incident on subjects with nearly identical demographic variables and same type of force used against them, so $\text{Var}(\varepsilon_p^{BH})$ from Equation 1.5 is very small and the OLS estimate is barely shrunk. To prevent these estimates from receiving undue weight, I winsorize the distributions at the 1% and 99% levels.

The winsorized standard deviations of these estimates are sizable: about 7 percentage points for at least unenhanced, 4.5 percentage points for at least enhanced, and 0.6 percent-

age points for fired weapons. Comparing the dispersion of the estimates to the outcome means in Figure 1.8 reveals the large magnitude of the more extreme departmental disparities. Encouragingly, many departments do not seem to have racial disparities, at least for certain force outcomes. About 22% of departments have zero or negative racial disparities for at least unenhanced and enhanced, while the number is 39% for fired weapons.

I now use the multiple force outcomes to examine racial disparities within departments. Do departments that have greater racial disparities at one level of force also tend to have greater disparities at others? For example, do departments with larger disparities for at least unenhanced force typically to have larger disparities on the at least enhanced measure? Figure 1.9 plots the correlations of winsorized department-level disparities across force levels to address Perhaps surprisingly, I find essentially no correlations across force levels. Knowing that a department has a large, small, or no disparity at one force level tells us almost nothing about its racial differences at another level.

Both across and within departments, we see variation in racial disparities. This variation highlights the difficulties in identifying problematic departments and designing interventions targeting racial policing. It may be difficult to treat departments with such heterogeneous racial differences, and the variation in disparities even within a department further confounds efforts.

1.4.3 Predicting Departmental Racial Disparities

Having established the heterogeneity present in department-specific racial disparities, I turn to the question of what can predict these disparities. For example, plotting a heatmap of the disparities in Figure A4 rules out geographic factors such as proximity to New York City or Philadelphia or clusters of departments with similar race gaps. I proceed by running regressions of the estimated departmental disparities from the winsorized empirical Bayes estimates from Equation 1.5 on an array of possible correlates (results using unshrunk OLS estimates from Equation 1.3 are in Appendix A.3). Because the mapping from police

departments to their municipalities is almost one-to-one, I include regressors based on both the police departments themselves and the municipalities they serve. Doing so requires dropping the New Jersey State Police, as that department covers state highways and some municipalities that lack their own department.

For ease of reading and interpretation, I first make two modifications to the data. First, I multiply the outcome by 100, so that, for example, a coefficient of 0.2 corresponds with an increase in the outcome of 0.2 percentage points, not 20 percentage points. Second, following Gelman (2008), I divide non-binary variables by two times their standard deviations, so that all coefficients (including binary ones) are on approximately the same scale; coefficients on non-binary variables may be interpreted as the change in outcomes associated with moving from one standard deviation below the mean to one standard deviation above it.

I report the results of these regressions graphically in Figure 1.10 and numerically in Table 1.5. Despite the number of explanatory variables and the popular belief that at least some of them can affect racial policing, the models do a poor job of predicting the racial disparities (all R^2 values are at most 0.039), and only one can reject the null hypothesis of an F-test of joint significance for all covariates at the 5% level. Most variables' coefficients change greatly across models, both in magnitude and even sign. This may not be a problem, as we have already seen the weak correlations of within-department, across-outcome disparities. For example, departments with more Black or Hispanic officers could have smaller disparities on the at least unenhanced measure, as suggested by the model and consistent with works such as Wilkins and Williams (2008), Nicholson-Crotty, Nicholson-Crotty, and Fernandez (2017), and Ba et al. (2020), but diversity may not be a strong enough channel to impact officers' intensive margin decisions at the higher levels of force, hence the null coefficients there. Overall, however, this exercise underscores the difficulty in identifying potential causes of (and thus, solutions to) racial differences in policing.

With these observables unable to predict the disparities, the variation may be coming

from variables I do not observe. Natural candidates include items that are difficult to obtain or measure such as department policies and intangibles like culture. To attempt to get at these factors, I use the 2016 LEMAS survey to acquire more detailed information on the subset of municipal departments surveyed in that year (this is a non-random sample, as the LEMAS uses a stratified sampling design at the national level based on the number of sworn personnel and agency type, and response rates, while high, are not 100%). From the survey, I add to the regressions a number of variables that reflect departments' policies and/or culture, such as whether they have written policies on unbiased policing, whether they formally survey local residents to obtain feedback and improve policing, whether they report positive asset forfeiture, and whether the chief executive of the department is female or non-Hispanic White.

The results from these regressions are in Figure 1.11 and Table 1.6. Comparing coefficients on variables from LEMAS and not, the at least enhanced and fired weapon models in particular show that the LEMAS-sourced variables on policy and culture have the largest correlation magnitudes. Mechanically, R^2 increases from the addition of regressors (which doubles to 20) and the reduction in sample size (454 to 92), but ranges from 0.18 to 0.26. For the purposes of fit comparison, Table A10 presents regression results with the subset of departments present in the 2016 LEMAS but using only the non-LEMAS predictors and has R^2 values ranging from 0.05 to 0.13. Overall, it is clear that much of the variation is still coming from other sources. These may be other cultural or policy factors or department-specific idiosyncrasies, but the variation across and within departments makes clear the lack of a singular predictor of racial disparities in police use of force.

1.5 Conclusion

In this paper, I combine new analytical strategies with incident-level data on all recorded uses of force by municipal and state police in New Jersey between 2012 and 2016 to es-

timate racial differences in the severity of force used on a subject, conditional on force. This rich dataset allows me to adjust for factors such as the type of incident or the subject's actions to reduce the role of selection into the dataset and better examine the role of race in police violence. I present evidence of large disparities across races, in particular for Blacks, in police use of force that increase along the spectrum of force severity. The department seems to play a larger role in determining these disparities than the individuals officers. I then extend empirical Bayes methods and document substantial heterogeneity in these disparities across and within departments, finding some departments without disparities against Black and Hispanic subjects, but also a long tail that especially disfavors them, and no within-department correlations across force levels. Finally, I show that many commonly proposed factors do a poor job of explaining these departmental disparities. While the strongest evidence is for cultural and policy variables, the large amounts of unexplained variation suggest this is incomplete, at least for available measures. Although the intensive margin measure of racial differences at the core of this paper is not the only such measure, it is central to the discourse around race and policing and heretofore underexplored, broadening our understanding of the problems at the heart of one of the most important social issues today.

Much work remains to be done on race and police use of force, with policy and research closely intertwined. The presence of departments without estimated racial disparities is a promising sign that progress is possible. But this variation may also mean that a uniform treatment is sub-optimal, and the difficulties in predicting departmental disparities using observational data make targeting reforms difficult. Police departments and public officials must ensure that data availability and transparency continue to improve to aid progress. Initiatives such as New Jersey's new statewide data portal combined with its mandatory force reporting and the FBI's National Use-of-Force Data Collection will hopefully facilitate future research on how race and policing interact. These problems cannot be solved without close collaboration of researchers and policymakers, and better data give both the

tools needed to move forward.

Table 1.1: Summary Statistics for Force Reports

Statistic	N	Mean
Max force: compliance hold	39,322	0.50
Max force: takedown	39,322	0.06
Max force: hands/fists	39,322	0.25
Max force: leg strike	39,322	0.05
Max force: pepper spray	39,322	0.11
Max force: baton	39,322	0.03
Max force: canine	39,322	0.01
Max force: taser	39,322	0.002
Max force: fired weapon	39,322	0.004
Max subject action: resisted	39,322	0.63
Max subject action: physical threat/attack	39,322	0.34
Max subject action: blunt weapon threat/attack	39,322	0.01
Max subject action: knife threat/attack	39,322	0.01
Max subject action: vehicular threat/attack	39,322	0.01
Max subject action: firearm threat	39,322	0.01
Max subject action: fired weapon	39,322	0.001
Officer injured	39,322	0.10
Incident: crime in progress	39,322	0.27
Incident: domestic dispute	39,322	0.13
Incident: other dispute	39,322	0.11
Incident: suspicious person	39,322	0.11
Incident: traffic stop	39,322	0.09
Incident: other	39,322	0.33
Subject: White	39,322	0.48
Subject: Black	39,322	0.41
Subject: Hispanic	39,322	0.10
Subject: Asian/Pacific Islander	39,322	0.01
Subject: female	39,322	0.20
Subject: age	39,322	31.09

Notes: Data cover all police departments in New Jersey from 2012 through 2016. Data have been restructured so that each observation is a subject who had force used against them by police and cleaned as described in Appendix A.1. "Max force" refers to the highest level of force used in an incident; lower levels of force are not reported. "Max subject action" indicates the most severe action a subject took that could justify an office using force. Force used and subject actions are ordered from least severe to most severe. Incidents may have multiple types.

Table 1.2: Effect of Female Subject on Police Use of Force of At Least Specified Severity, Conditional on Force

	Unenhanced	Enhanced
Subject Black/Hispanic	0.014^{**}	0.028^{***}
	(0.007)	(0.005)
Subject female	-0.173^{***}	-0.064^{***}
	(0.009)	(0.005)
Subject Black/Hispanic $\times$ female	0.008	0.003
	(0.011)	(0.008)
Fixed effects	Dept.	Dept.
Clustering	Dept.	Dept.
Outcome mean	0.505	0.146
R^2	0.180	0.118
Num. obs.	39184	39184

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race interacted with an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department fixed effects.

Table 1.3: Effect of Being in Department with Above-Median Empirical Bayes Racial Disparity, Officers Switching Departments Only

	Unenhanced	Enhanced
Subject Black/Hispanic	-0.069	0.001
	(0.056)	(0.032)
Department above median	-0.076	0.066
	(0.063)	(0.042)
Subject Black/Hispanic $\times$ department above median	0.235^{***}	-0.005
	(0.070)	(0.052)
Fixed effects	Off.	Off.
Clustering	Off.	Off.
R^2	0.443	0.297
Num. obs.	1099	1099

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, department fixed effects, and an indicator for the department in which the officer served having above-median racial disparities from the empirical Bayes analysis. This regression only includes incidents from officers identified as switching departments.

Table 1.4: Summary Statistics for Empirical Bayes Department x Black/Hispanic Subject Interactions

	Unenhanced	Enhanced	Fired Weapons
SD	0.086	0.062	0.009
SD (Winsorized)	0.068	0.045	0.006
Min	-0.412	-0.237	-0.127
P01	-0.181	-0.178	-0.025
P05	-0.053	-0.086	-0.007
P25	0.004	0.006	-0.001
Median	0.015	0.028	0.001
P75	0.028	0.033	0.002
P95	0.094	0.056	0.009
P99	0.430	0.088	0.022
Max	0.778	0.895	0.038
Mean	0.021	0.014	0.000
$\% \leq 0$	0.223	0.225	0.386

Notes: Table reports empirical Bayes estimates of departmental racial disparities from Equation 1.5, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race interacted with a department indicator, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department fixed effects. Winsorization is done at the 1st and 99th percentiles.

Table 1.5: Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Variables Normalized by 2x Standard Deviation

	Unenhanced	Enhanced	Fired Weapons
Log median household income	0.563	−0.353	0.091
	(0.782)	(0.539)	(0.077)
Gini coefficient	−0.266	0.502	−0.068
	(0.746)	(0.494)	(0.054)
Log population	1.338	0.722	−0.010
	(1.440)	(1.105)	(0.134)
Log population per square mile	−0.005	0.580	0.129*
	(0.844)	(0.536)	(0.067)
Log number of police	−1.841	−0.398	−0.166
	(1.421)	(1.099)	(0.155)
% officers not White or Asian/PI	−1.134**	0.223	0.037
	(0.455)	(0.396)	(0.060)
% population 18-65 not White or Asian/PI	−0.521	−0.615	0.188
	(0.904)	(0.784)	(0.115)
Violent crime rate	0.679	0.913	−0.033
	(0.629)	(0.594)	(0.095)
2012 presidential election Romney vote %	−0.027	−0.629	0.115
	(1.173)	(0.857)	(0.122)
County mean Black-White IAT *D*-score	−0.962	0.503	0.053
	(0.672)	(0.577)	(0.072)
Prob > F	0.437	0.045	0.179
R^2	0.019	0.035	0.039
Num. obs.	454	454	454

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each variable on the right-hand side of the regression has been divided by twice its standard deviation.

Table 1.6: Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Numeric Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only

	Unenhanced	Enhanced	Fired Weapons
Log median household income	0.603	−0.730	−0.011
	(1.150)	(1.105)	(0.130)
Gini coefficient	1.442	−0.460	−0.103
	(1.060)	(1.139)	(0.186)
Log population	3.012	0.521	−0.041
	(2.314)	(1.946)	(0.456)
Log population per square mile	1.316	0.048	0.027
	(1.265)	(1.210)	(0.169)
Log number of police	−3.868	−0.537	−0.115
	(2.471)	(2.052)	(0.483)
% officers not White or Asian/PI	−1.207	0.158	0.022
	(0.817)	(0.881)	(0.156)
% population 18-65 not White or Asian/PI	2.254	0.661	0.133
	(1.869)	(1.584)	(0.226)
Violent crime rate	0.680	0.127	0.080
	(0.832)	(0.687)	(0.099)
2012 presidential election Romney vote %	0.704	−0.330	0.069
	(1.580)	(1.714)	(0.234)
County mean Black-White IAT *D*-score	−1.461	1.420	0.001
	(0.898)	(0.887)	(0.151)
Dept. has nonzero asset forfeiture	−1.332	−0.865	0.051
	(1.550)	(1.504)	(0.205)
Dept. uses written aptitude test in selecting recruits	0.466	1.893**	0.050
	(1.118)	(0.820)	(0.133)
Dept. maintains a written community policing plan	1.172	0.985	−0.074
	(0.838)	(0.946)	(0.131)
Dept. formally surveys local residents	−2.942	−0.604	−0.004
	(2.033)	(1.271)	(0.336)
Dept. has written policy on stop and frisk	0.470	−0.041	−0.230
	(1.393)	(1.630)	(0.193)
Dept. has written policy on foot pursuits	1.823	0.903	−0.123
	(1.195)	(1.190)	(0.142)
Dept. has written policy on unbiased policing	−1.546	−3.393***	0.229
	(1.788)	(1.279)	(0.242)
Dept. chief executive is female	0.228	−1.578	−0.893*
	(2.690)	(1.465)	(0.478)
Dept. chief executive is non-Hispanic White	−1.081	−2.922***	0.054
	(2.062)	(1.061)	(0.279)
Prob > F	0.227	0.023	0.182
R^2	0.246	0.197	0.203
Num. obs.	92	92	92

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each non-binary variable on the right-hand side of the regression has been divided by twice its standard deviation.

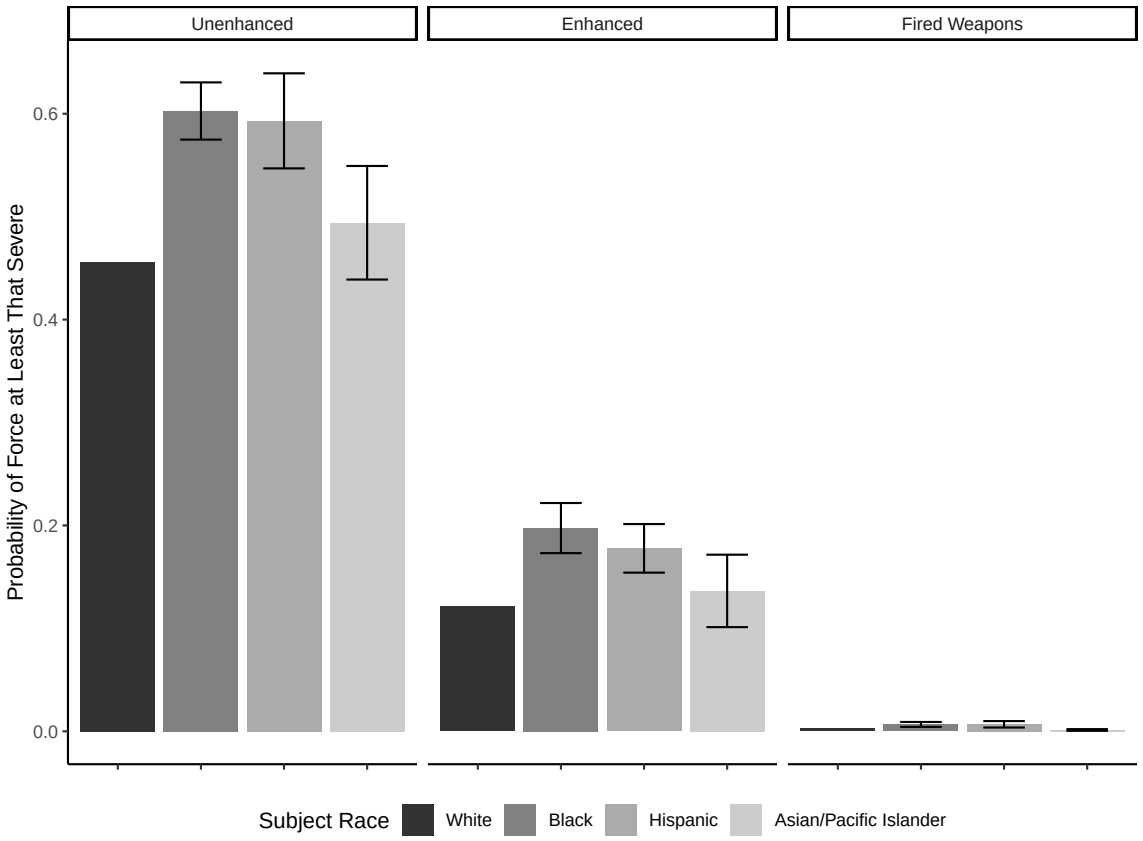


Figure 1.1: Overall Racial Disparities (No Controls)

Notes: Figure presents results from a series of OLS models regressing outcomes on a full set of racial dummies. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Non-White bars are obtained by taking the intercept (Whites) and adding the coefficient on that race. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the department level.

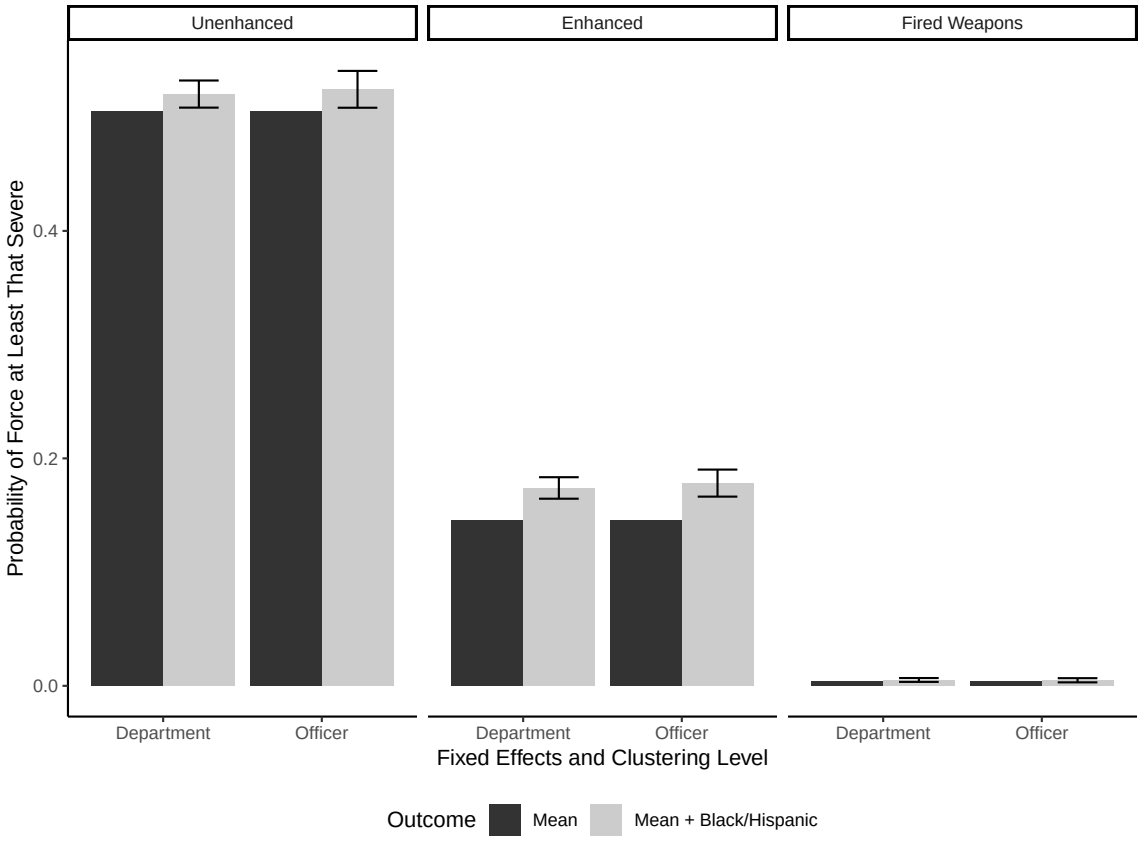


Figure 1.2: Overall Racial Disparities for Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

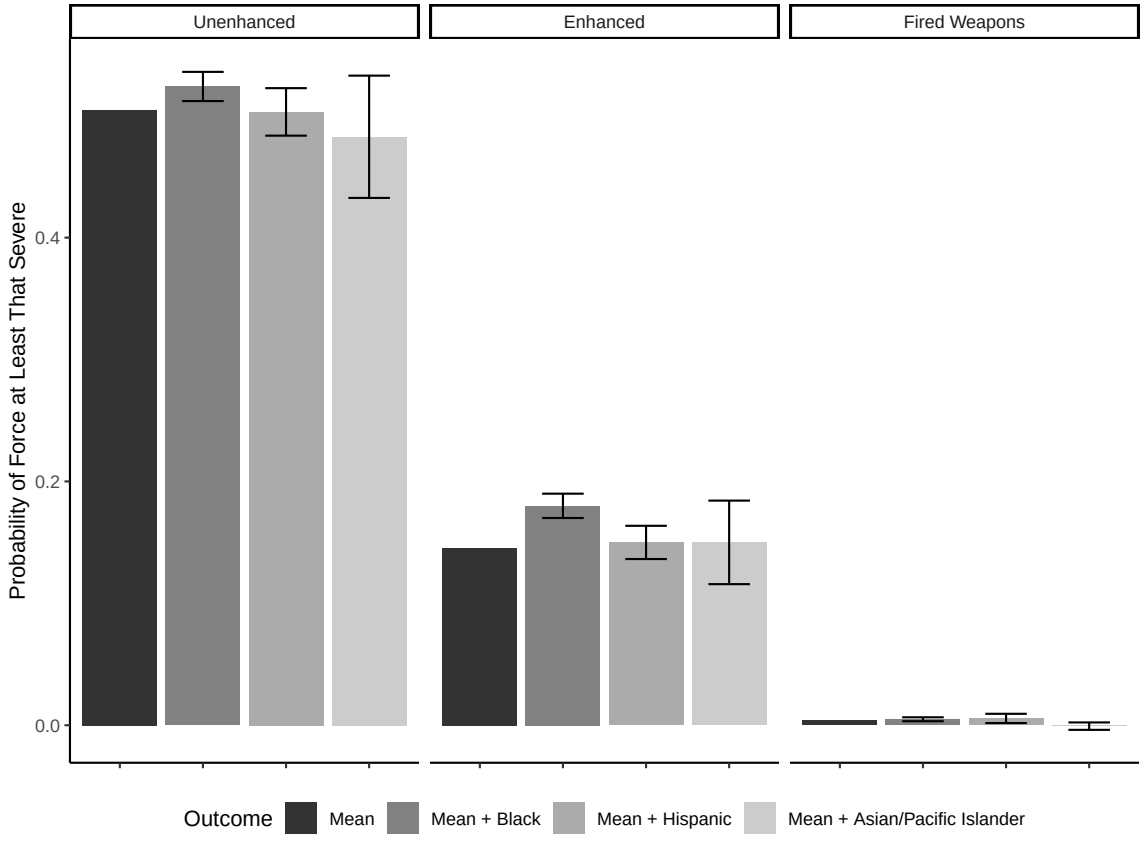


Figure 1.3: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Department Fixed Effects and Clustering)

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the department level.

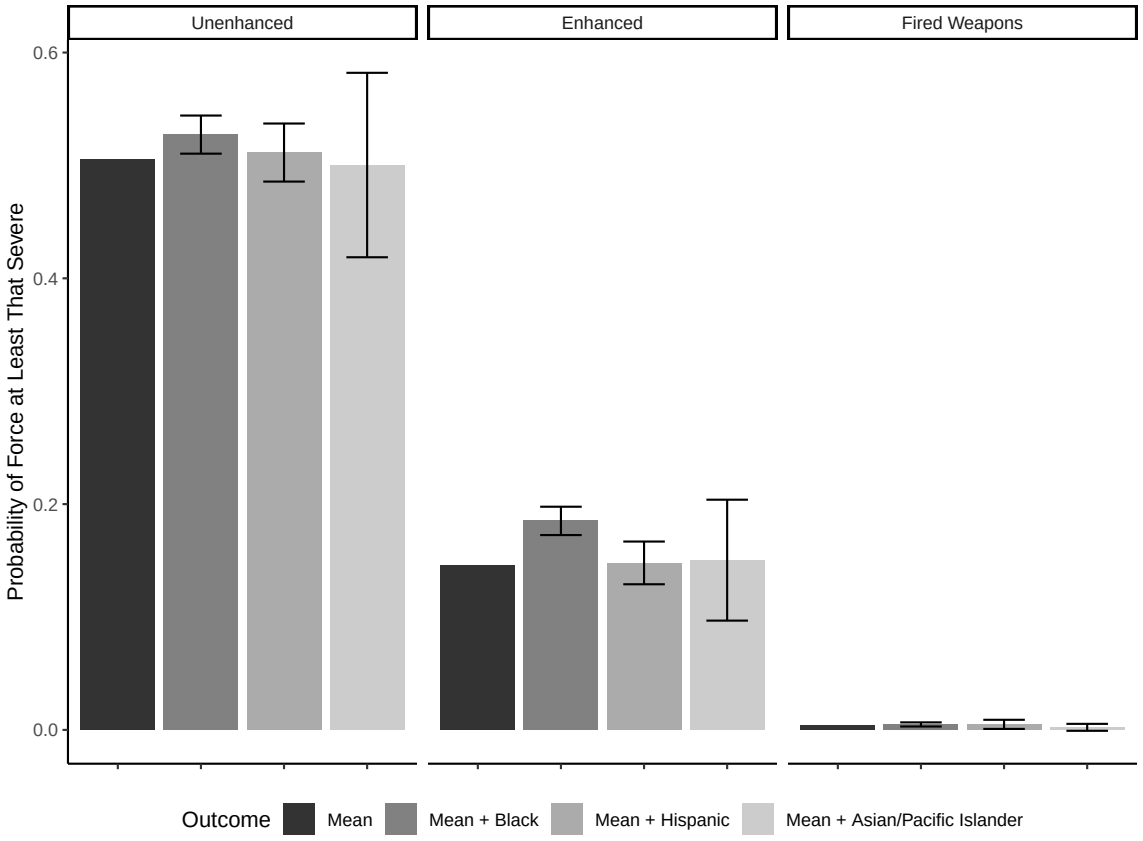


Figure 1.4: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Officer Fixed Effects and Clustering)

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the officer level.

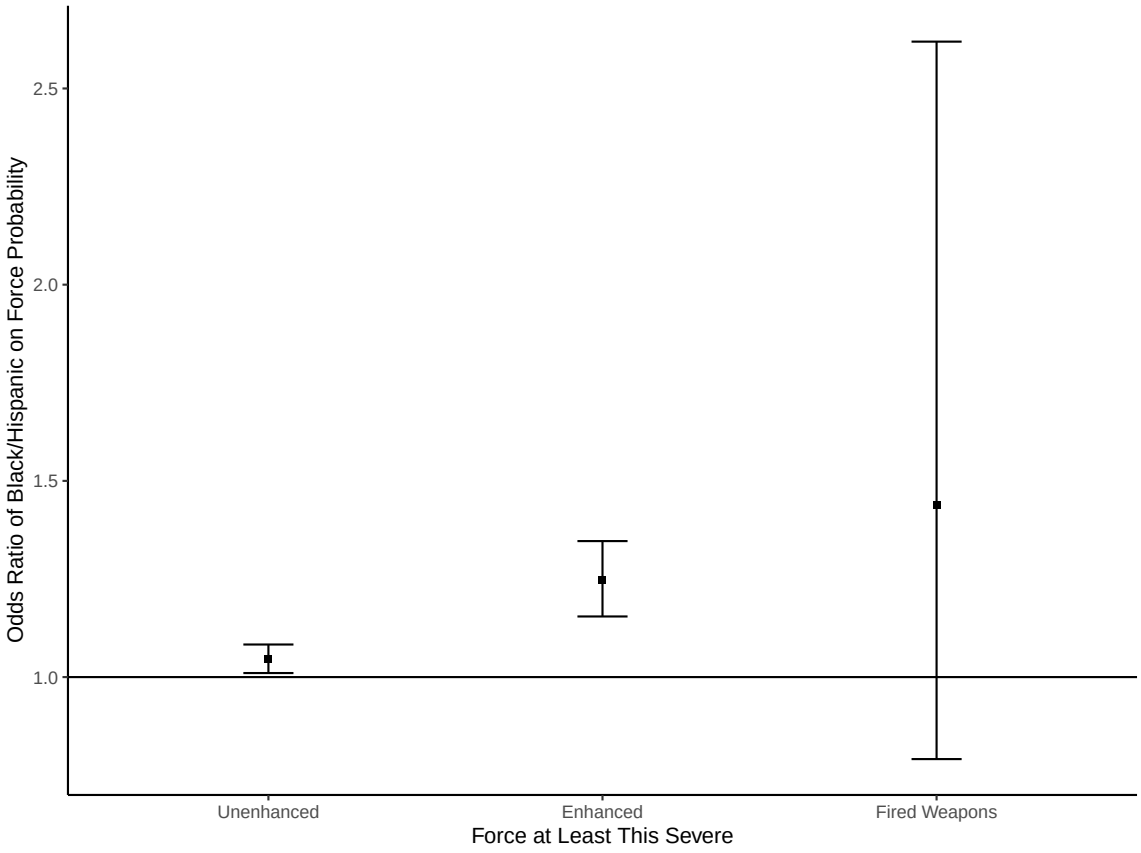


Figure 1.5: Conditional Logit Odds Ratios of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity

Notes: Figure presents results from a series of conditional logit models fit via Equation 1.2. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Points indicate the odds ratio of a Black/Hispanic subject. Confidence intervals clustered at the department level are obtained by exponentiating the logit coefficients from the regression and are asymmetric.

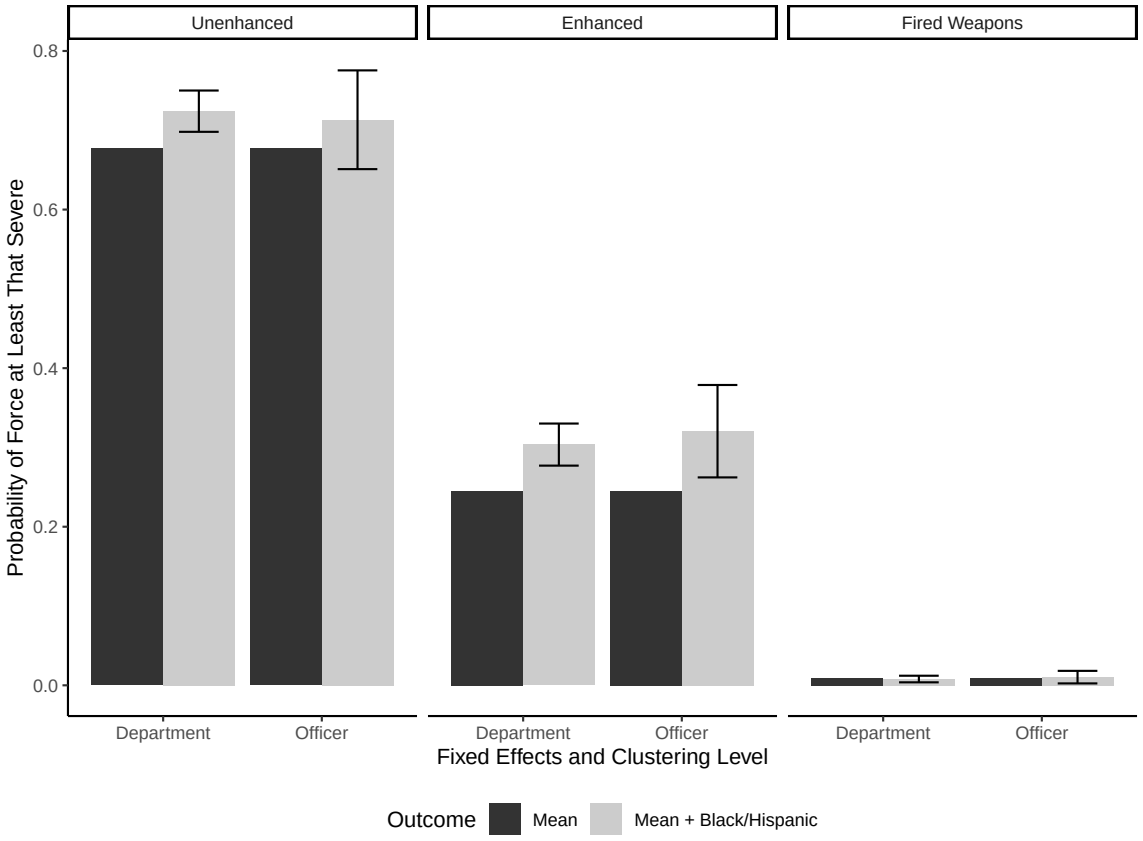


Figure 1.6: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject and Subject at Least Physically Threatened/Attacked Officer or Another)
Notes: Figure presents results from a series of OLS models estimated on the subset of the data where a subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night, and where the subject at least physically threatened or attacked an officer or another. The latter restriction is equivalent to dropping observations where the most severe actions by the subject was resisting. Regressions are fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

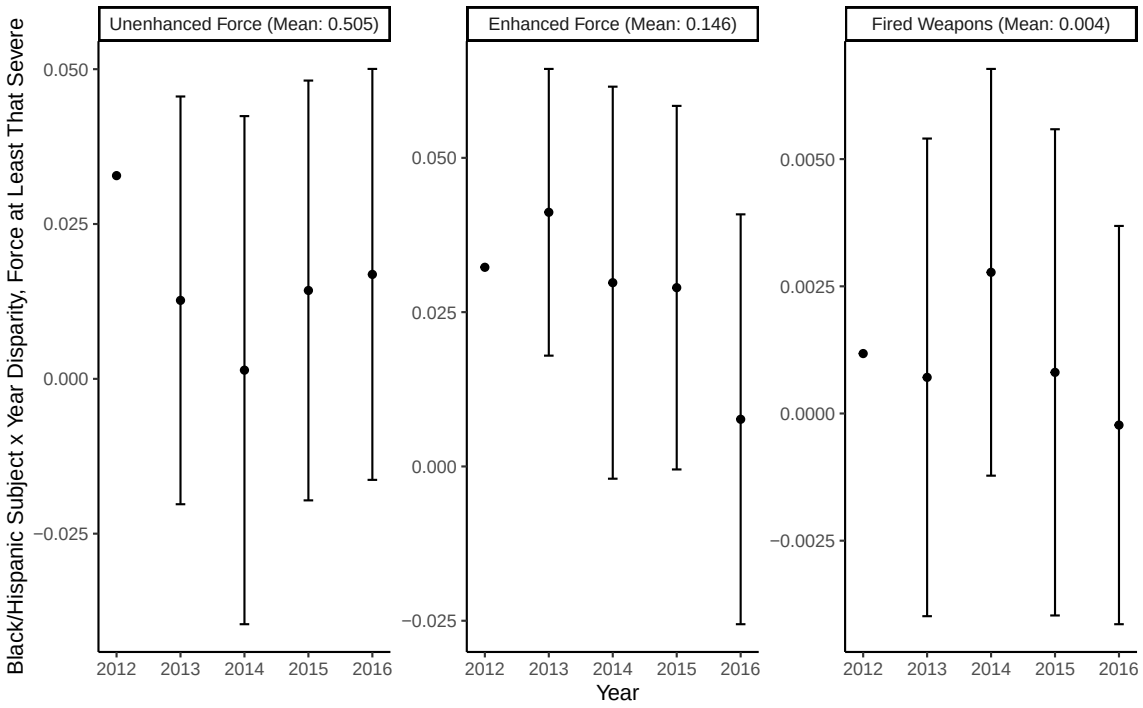


Figure 1.7: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity by Year

Notes: Figure presents results from a series of OLS models fit via Equation 1.1 with the Black $\times$ Hispanic indicator interacted with year (2012 is omitted). Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. 95% Confidence intervals are based on the corresponding year's coefficient. Standard errors are clustered at the department level.

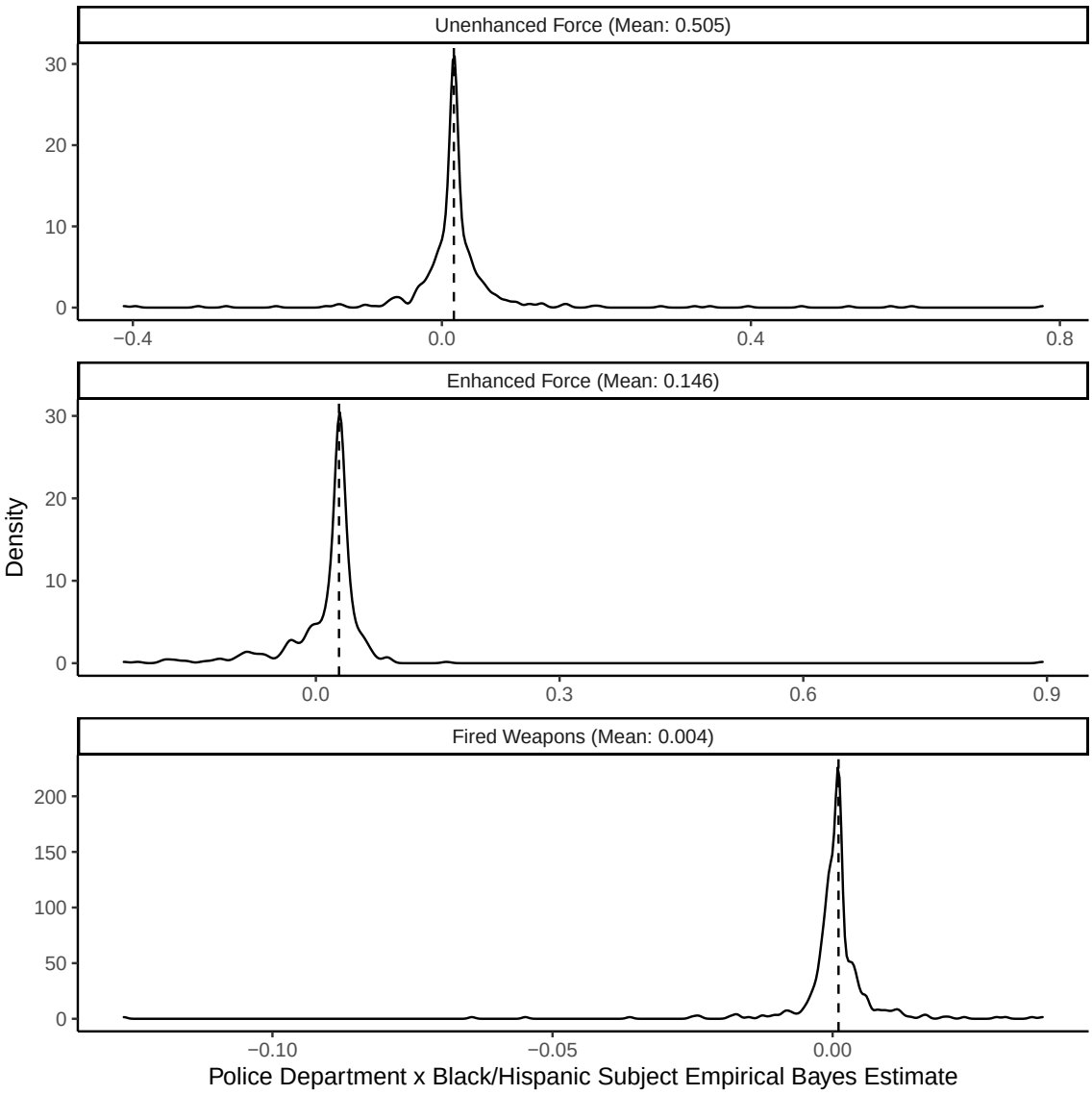


Figure 1.8: Distribution of Empirical Bayes Estimates of Department-Black/Hispanic Interactions

Notes: Figure presents kernel density estimates of department-specific racial disparities β between White/Asian/Pacific Islander subjects and Black/Hispanic subjects as estimated from the empirical Bayes estimator in Equation 1.5 with Gaussian kernels and the Silverman (1986) rule-of-thumb bandwidth. Each subgraph shows results from regressions with the specified outcome outcome: whether, conditional on any force being used, force of at least the specified severity was used. The dashed line indicates the mean of the prior.

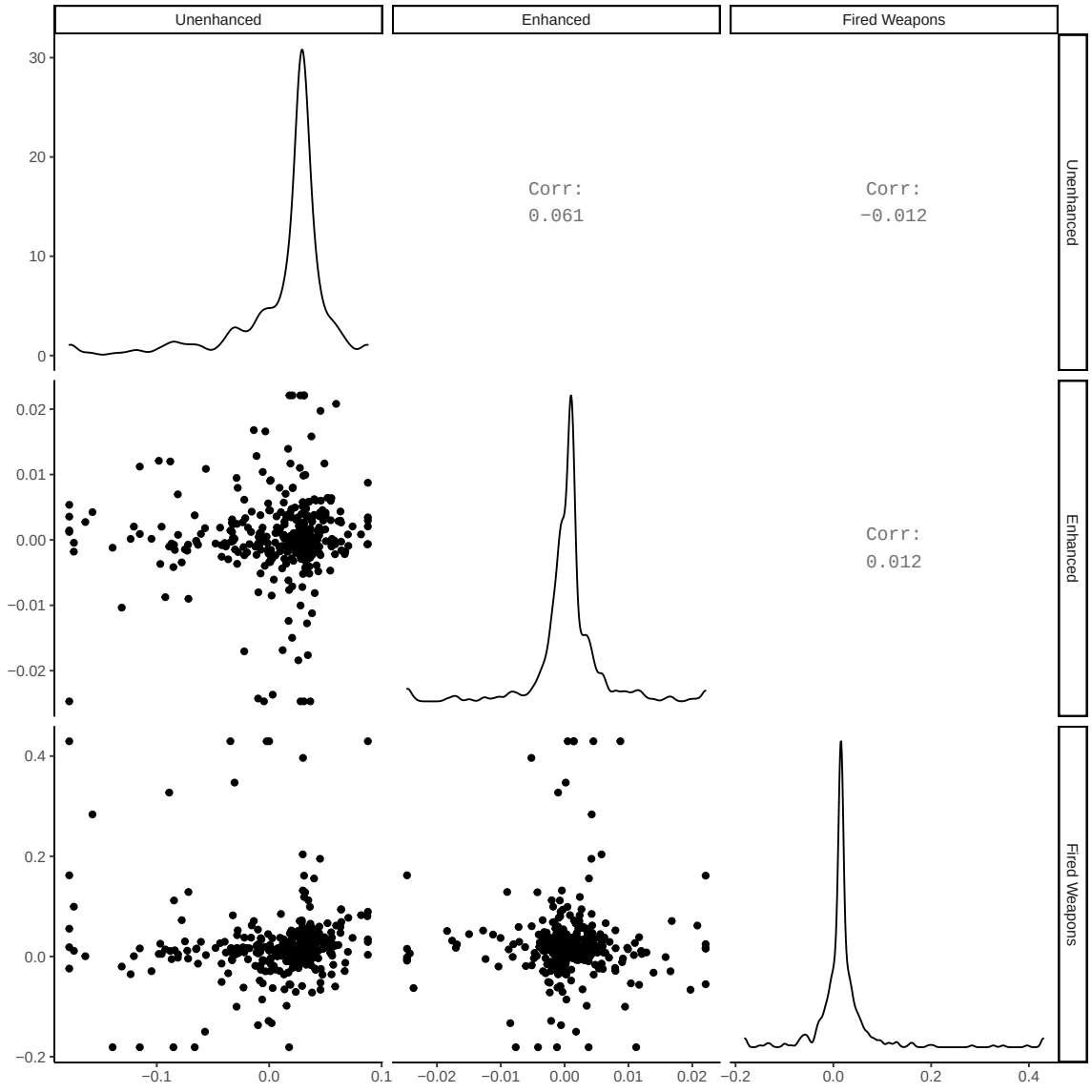


Figure 1.9: Within-Department Correlations of Winsorized Empirical Bayes Estimates of Department-Black/Hispanic Interactions Across Force Types

Notes: Figure presents the correlations across different force types of the winsorized department-specific racial disparities β between White/Asian/Pacific Islander subjects and Black/Hispanic subjects as estimated from the empirical Bayes estimator in Equation 1.5. Winsorization is done at the 1st and 99th percentiles. Every level of force includes all types of force at least that severe. The bottom left triangle depicts scatterplots of the estimates for each department for the force types in the corresponding column and row. The diagonal shows a histogram of the distribution of department-specific estimates. The upper triangle presents the correlation coefficient for the estimates of the corresponding force types.

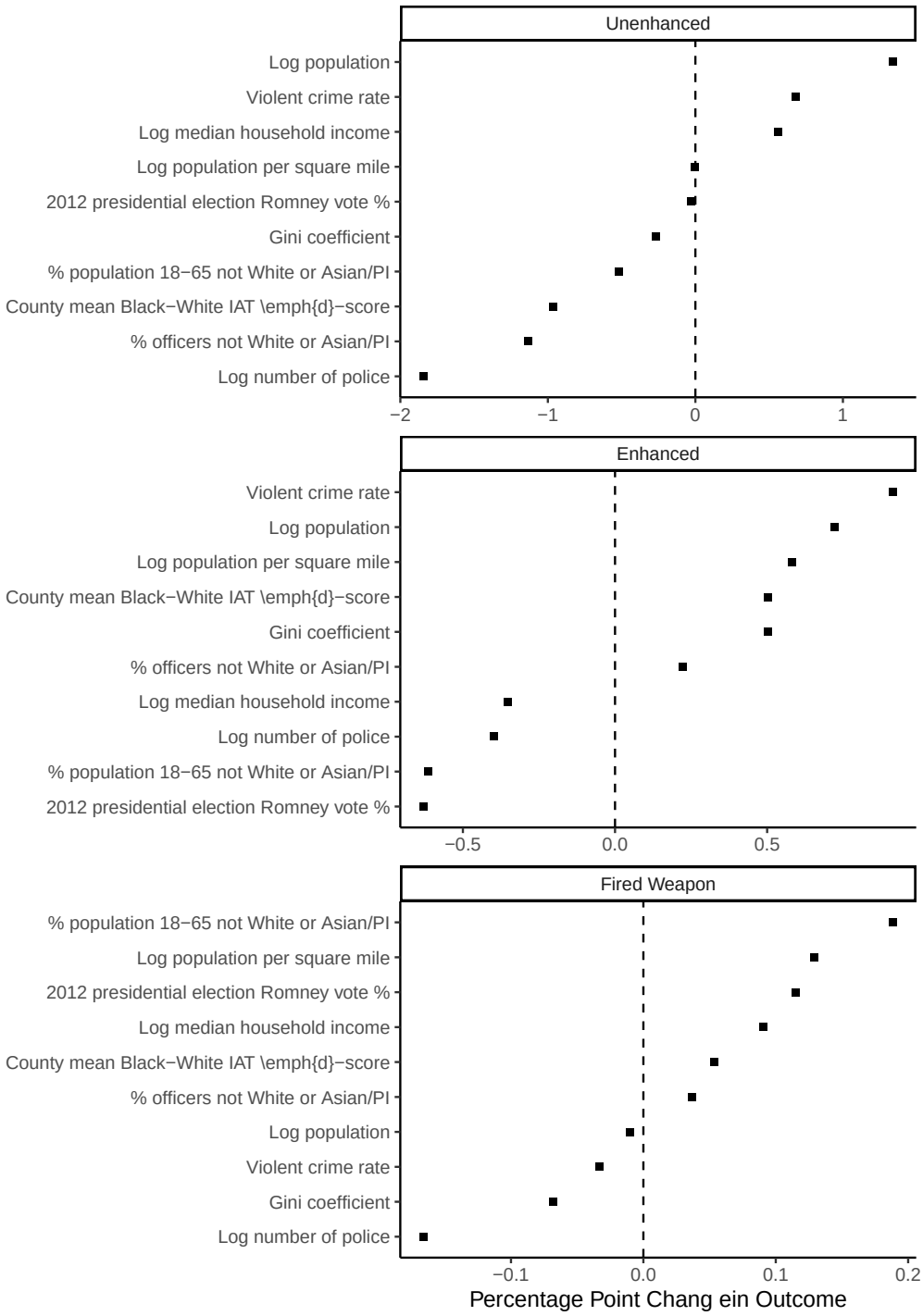


Figure 1.10: Coefficients from Regression of Empirical Bayes Estimates of Department-Black/Hispanic Interactions on Possible Correlates

Notes: Figure presents coefficients from a regression of winsorized empirical Bayes estimates of departmental disparities on the listed variables as described in Equation 1.6. Each variable on the right-hand side of the regression has been divided by twice its standard deviation to aid in interpretation.

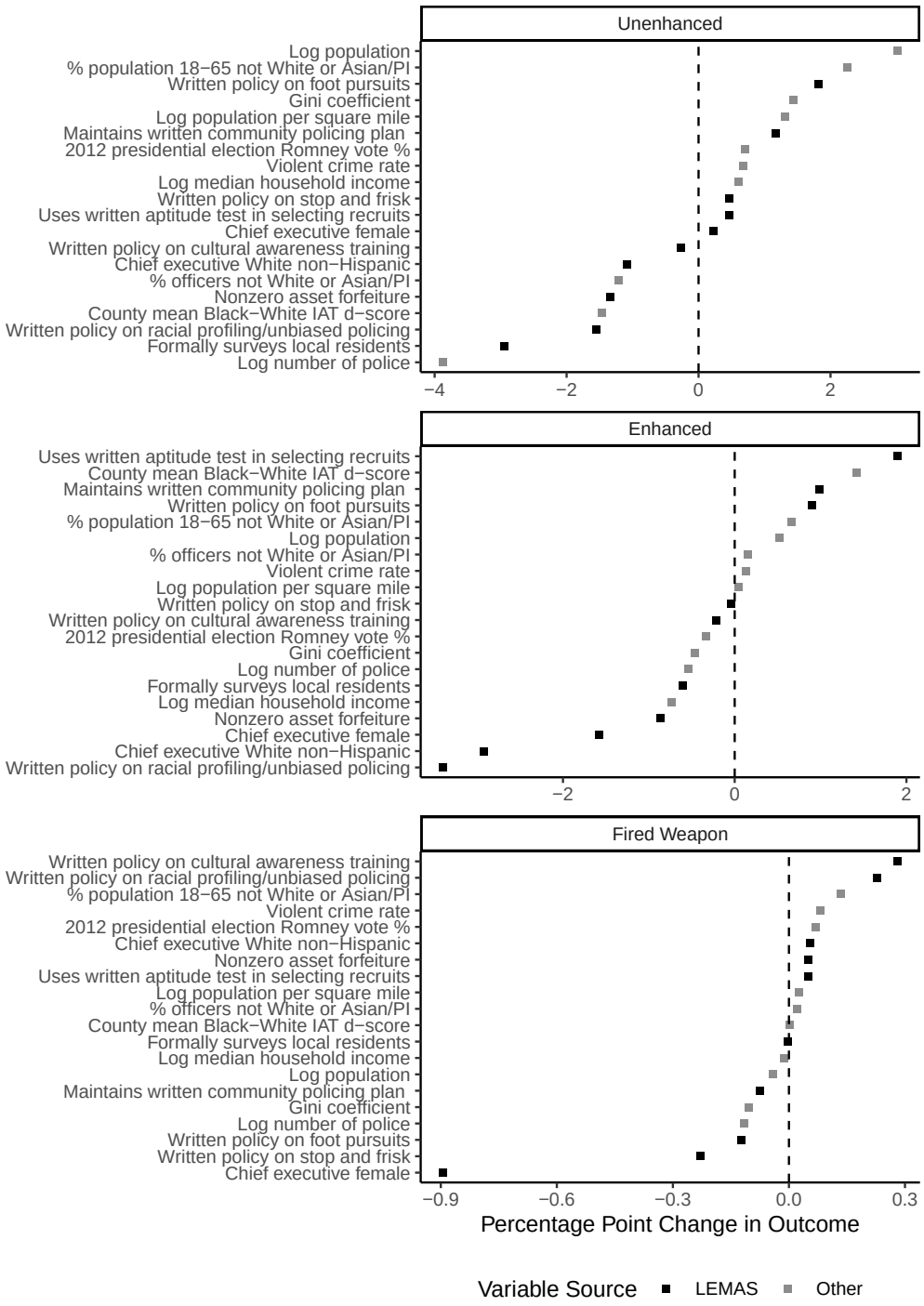


Figure 1.11: Coefficients from Regression of Empirical Bayes Estimates of Department-Black/Hispanic Interactions on Possible Correlates, Departments in 2016 LEMAS Only

Notes: Figure presents coefficients from a regression of winsorized empirical Bayes estimates of departmental disparities on the listed variables as described in Equation 1.6. Non-binary variables on the right-hand side of the regression has been divided by twice its standard deviation to aid in interpretation.

Chapter 2

Visual Inference And Graphical Representation in Regression Discontinuity Designs[1]

2.1 Introduction

"Few would deny that the most powerful statistical tool is graph paper."

— Geoffrey S. Watson (1964)

Despite the widespread use of graphical analysis in applied economic studies, there is little research on readers' ability to process the statistical information graphs are meant

[1]This chapter is coauthored with Christina Korting, Jordan Matsudaira, Zhuan Pei, and Yi Shen. We have benefited from comments by Colin Camerer, Matias Cattaneo, Damon Clark, Nathan Grawe, David Lee, Lars Lefgren, Thomas Lemieux, Pauline Leung, Jia Li, Adam Loy, Alex Mas, Doug Miller, Ted O'Donoghue, Steve Pischke, Jesse Rothstein, Rocio Titiunik, Stephanie Wang, and Andrea Weber, as well as seminar participants at Cornell, UC Irvine, and Princeton. Michael Daly, Rebecca Jackson, Motasem Kalaji, Xingyue Li, Fiona Qiu, and Tatiana Velasco provided research assistance, and we thank Brad Turner and Patti Tracey for providing logistical support. We are indebted to our friends for testing the experiment, to colleagues for participating in the study, and to Sahara Byrne for giving us access to her eyetracking lab. Financial support from the Cornell Institute of Social Sciences and the Princeton Industrial Relations Section is gratefully acknowledged. The authors have no relevant or material financial interests that relate to the research of this paper. This study is registered in the Open Science Framework and the AEA RCT Registry with ID AEARCTR-0004331. Preliminary versions of this chapter were presented at meetings of the Association for Public Policy Analysis and Management, the Society of Labor Economists, and the International Association for Applied Econometrics.

to convey (visual statistical inference or *visual inference* per Majumder, Hofmann, and Cook, 2013), especially in the empirical applications where graphs are most commonly used. In this paper, we evaluate the performance of visual inference in the regression discontinuity design (RDD or RD design). The popularity of RDD in the modern causal inference toolkit, which began in economics with Angrist and Lavy (1999), makes it an important setting in which to study visual inference. Standard practices in applying RDD today perhaps best embody the spirit of Watson's quote above, with graphs playing a central role in the presentation of findings. The key RD graph plots the bivariate relationship between outcome variable Y and running variable X and is meant to display a discontinuity (or lack thereof) in the underlying conditional expectation function (CEF) as X crosses a policy threshold. Influential practitioner guides by Imbens and Lemieux (2008) and Lee and Lemieux (2010) recommend creating this graph by dividing X into bins, computing the average of Y within each bin, and generating a scatter plot of these Y-averages against the midpoints of the bins.

We assess the performance of visual inference by studying whether people presented with this graph can correctly identify the existence or absence of a discontinuity at the policy threshold. This endeavor raises four conceptual and practical questions. First, because we can construct many different RD graphs based on the same data, how do graphical representation techniques affect visual inference? Second, as readers may have different abilities, whose visual inference should we evaluate? Third, what benchmark should we use to judge the adequacy of visual inference? And fourth, with the hope of improving it, can we understand the mechanics of visual inference?

To address the first question, we build on work pioneered by Eells (1926) and refined by Cleveland and McGill (1984) who experimentally evaluate the accuracy of visual perception across a variety of perceptual tasks (such as discerning the area, angle, or saturation of a graphical element). We conduct a series of randomized experiments to examine how different graphical parameters affect visual inference in RD. We present participants with RD

graphs produced from data generating processes (DGPs) based on microdata from 11 published papers that we randomly selected from a list of 110 empirical studies drawn from top economics journals (the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal of Business and Economic Statistics*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics and Statistics*). For each graph, we ask participants to identify the existence or absence of a discontinuity. We randomize respondents into different treatment arms and show participants within each arm graphs produced with particular graphical parameters such as small bin widths and evenly spaced bins.

There is limited research on how to choose these parameters in practice. For example, Calonico, Cattaneo, and Titiunik (2015) propose two bin width selectors: one that minimizes the integrated mean squared error (IMSE) of the bin averages, resulting in fewer, larger bins, and another that mimics the variability of the underlying data (mimicking variance or MV), which leads to more, smaller bins. While both proposals set a graphical parameter to satisfy an econometric criterion, practitioners are left with little basis to choose between them. Moreover, a host of other choices over graphical parameters remains with minimal guidance from the literature, such as including smoothed regression lines in the binned scatter plot, adding a vertical line to indicate the policy threshold, and choosing the axis scales.

By comparing the rates at which respondents correctly classify discontinuities across treatment arms, we can assess the pros and cons of different graphical parameters, which informs best practices in RD visual representation. We find that certain graphical parameters such as bin width and imposing smoothed regression lines create important tradeoffs between type I error (incorrectly identifying a discontinuity when there is none) and type II error (incorrectly identifying the absence of a discontinuity when there is one) rates. Relative to IMSE bins and imposing fit lines, using MV bins and not including the fit lines tends to decrease type I error rates but increase type II error rates.

To address the second question of evaluating visual inference and to assess how our estimates generalize across experience levels, we compare the performance of non-experts in our randomized experiments recruited through the Cornell University Johnson College's Business Simulation Lab with the performance of experts in fields that often employ RD designs. We find that the two groups perform comparably and are similarly affected by graphical methods. In the spirit of DellaVigna and Pope (2018), we also test whether experts are able to predict the graphical parameters that result in the highest rate of visual inference success by non-experts. We find that experts only partly anticipate the aforementioned tradeoffs from bin widths and fit lines.

Third, as a benchmark for the performance of visual inference, we compare the type I and type II error rates of our crowdsourced discontinuity/kink classifications with those from inference based on econometric estimators at the 5% (asymptotic) level. These comparisons confirm that the graphical representation of RD designs is a valuable tool for empirical analysis and complements econometric inference. We find that visual inference on graphs generated with methods leading to the lowest type I error rates performs very similarly to the Armstrong and Kolesár (2018) RD inference procedure (henceforth AK). RD visual inference achieves lower type I error rates than tests based on the Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) estimators (henceforth IK and CCT, respectively), but the two procedures exhibit considerable power advantages. The examination of the joint distribution of visual and econometric tests further illustrates their complementarity: while they commit similar type II errors, there does not appear to be a strong association in their type I errors.

Fourth, to understand the mechanics behind successful visual inference and investigate avenues for its improvement, we conduct an eyetracking study with both expert and non-expert participants. We track participants' visual attention to different parts of RD graphs while they form their discontinuity assessments. Although the six experts outperform the nine non-experts in this small lab study, their eyegaze patterns are similar along most di-

mensions, and various prediction exercises do not reveal eyegaze measures that robustly correlate with visual inference success. We tentatively conclude that the *processing* of visual signals, as opposed to their acquisition through ocular movements, is likely the driver of performance heterogeneity in visual inference.

We also evaluate visual inference in the closely related regression kink design (RK design or RKD). First, we use an analogous experimental approach to study graphical representation and additionally test the effects of visualization methods specific to RKD, namely plotting detrended outcomes as in Lee et al. (2019) and plotting first derivative estimates by adapting the generalized bin selector from Cattaneo, Farrell, and Feng (2020). Like in RD, we find similar tradeoffs in the choice of bin width, while plotting detrended outcomes decreases type II error rates and plotting derivative estimates leads to poor visual inference performance. Second, we document the performances of both experts and non-experts and observe that experts attain similar type I error rates but lower type II error rates compared with non-experts. Third, both experts and non-experts struggle to differentiate kinks from smooth peaks, which the default CCT inference procedure is better able to handle. However, the latter has a serious power disadvantage even at the largest kink magnitude we test. The lack of power can be alleviated by using conventional inference methods based on a local linear or quadratic estimator in conjunction with a larger bandwidth, consistent with the findings by Card et al. (2017).

The results from this paper lead to takeaways of RD/RK graphical representation and econometric inference for both practitioners and econometricians. For graphical representation in RD, if the primary concern is type I error rates, consistent with the notion of size control in econometric inference, one should use small bins with no fit lines as opposed to large bins with fit lines. Other parameters such as bin spacing (equally spaced versus quantile-spaced), axis scaling, and the presence of a vertical line indicating the policy threshold do not appear to matter, implying that researchers can adhere to reasonable personal preferences. For RK, one may consider plotting detrended outcomes, while plotting

derivative estimates is not recommended. The RD bin size result particularly underscores the value of using human perception as the criterion for evaluating graphs: while the larger IMSE bins are derived from a conventional statistical criterion, they have important drawbacks compared to the MV bins, which come from a non-standard statistical criterion and are constructed in a more ad-hoc way.

For econometric inference, the performance of visual inference provides an important benchmark: it informs the econometric procedure's value added above what readers glean from a carefully constructed graph. We find that the experts' performance is almost identical to that of the AK RD inference. By comparison, the IK and CCT RD procedures have elevated type I error rates (the CCT type I error rates are not significantly higher) but lower type II error rates. We can lower their type I error rates to that of the experts (just below 8%) by increasing the critical t-values to 2.46 for IK and 2.28 for CCT, and the resulting inference procedures retain their power advantages. Econometric methods add the most value in delivering estimates of the discontinuity magnitude: for example, the IK estimator yields lower mean squared errors than experts across all 11 DGPs.

This paper brings together a diverse set of literatures and makes the following contributions. i) We are the first to evaluate visual inference and graphical representation practices in two widely used quasi-experimental research designs in economics and other social sciences by drawing from four strands of the statistics literature that study the choice of graphical parameters (e.g. Calonico, Cattaneo, and Titiunik, 2015, Li et al., 2020), their effects on visual inference (e.g. Cleveland and McGill, 1984), the evaluation of visual inference through comparison with econometric inference (e.g. Majumder, Hofmann, and Cook, 2013), and the ocular process underlying visual inference via eyetracking technology (e.g. Hofmann et al., 2012). To guide our study design and help interpret our empirical results, we construct a general conceptual framework, which extends to future studies of visual inference in other contexts. ii) This is also the first paper within economics that uses lab experiments to study empirical methods, a new paradigm that can be applied to other

important areas (see Section 2.5 for more details). Using 11 DGPs calibrated to RD and RK microdata allows us to carry out a more comprehensive and empirically relevant evaluation of econometric inference procedures than econometric studies that typically rely on two or three DGPs. iii) This paper relates to the recent economics studies on methodological transparency and scientific communication (e.g. Angrist and Pischke, 2010; Andrews, Gentzkow, and Shapiro, 2020; Andrews and Shapiro, 2020). In particular, the bin width choice can be seen in the lens of Andrews, Gentzkow, and Shapiro (2020) as providing more or less data for the reader to draw inference on the parameter of interest while still being easier to process than the underlying microdata. And iv), we add to the literature on expert judgments (e.g. Camerer and Johnson, 1997) and expert forecasts of research results (e.g. Sanders, Mitchell, and Chonaire, 2015). Our finding that experts only partly anticipate our experimental results underscores the value of empirically evaluating visual inference and providing evidence-based guidance on graphical methods.

The remainder of this paper is organized as follows. In Section 2.2, we present our conceptual framework and formalize the measure we use to judge visual inference. In Section 2.3, we discuss the design of our experiments and studies, including the graphical parameters we test, the specification of the data generating processes, the design of the online survey, and the setup of our eyetracking study. Section 2.4 presents the empirical results. Section 2.5 concludes.

2.2 Conceptual Framework

In this section, we propose a conceptual framework for evaluating visual inference to guide study design and help with the interpretation of empirical results. The framework addresses the central questions arising from features of visual inference which set it apart from standard econometric inference. In particular, we outline how to aggregate visual inference performance across subjects given that they may reach different conclusions even when

viewing the same graph and delineate what meaningful parameter our aggregate measure corresponds to.

Although RD graphs may serve other purposes, we view their most important function as accurately conveying discontinuity existence/magnitude at the policy threshold. According to Lee and Lemieux (2010), other purposes of RD graphs include i) helping to assess regression specifications and ii) allowing for the inspection of discontinuities away from the policy cutoff. But ultimately these other functions are motivated by inference on the discontinuity at the policy threshold: i) can be viewed as reconciling visual and econometric inferences thereof and ii) informs the reader, under implicit global homogeneity assumptions, whether to believe the existence of a discontinuity at the policy threshold.

We therefore focus on binary classifications of a graph and treat type I and type II errors as the main performance measures for visual inference. However, the conceptual framework easily generalizes to assessing stated discontinuity magnitudes (which we elicit from experts).[2] A person commits a type I error in RD visual inference if she classifies a continuous graph as having a discontinuity, and a type II error if she classifies a discontinuous graph as continuous.

We introduce the following elements of the conceptual framework for RDD (the RKD framework is analogous). First, the vector γ denotes a combination of graphical parameters: bin width, bin spacing, the use of polynomial fit lines, inclusion of a vertical line, axis scaling, etc. Each entry of γ represents the value of a particular graphical parameter. In our experiments, we randomly assign each participant to view only graphs generated with a certain fixed value of γ. Ideally, one could run a large experiment with a full factorial design to test all combinations of the graphical values outlined above, but resource constraints force us to test a subset of the graphical parameter space via a sequence of studies as described in Section 2.3.3.

[2]Theoretically, we could attempt to obtain the confidence set of visual inference by asking the respondents whether the discontinuity is equal to d for a range of possible values for d. Given the practical challenges in implementation, however, we leave this approach to future work.

The combination (g,d) denotes the probability model underlying an RD dataset. g encompasses four elements: i) the distribution of the running variable X; ii) the conditional expectation function $E[\tilde{Y}|X=x]$ which is continuous at $x=0$; iii) the distribution of the error term u where $\tilde{Y} \equiv E[\tilde{Y}|X=x]+u$; iv) the sample size N. We note that the variable Y can represent the outcome, baseline covariates, or treatment take-up, so this framework applies to all graphs typically included in RD studies, including those from a fuzzy design. Intuitively, g specifies everything in the probability model except for the discontinuity, including the shape of the conditional expectation function. The discontinuity then results from shifting the right arm of the smooth function $E[\tilde{Y}|X=x]$ by some discontinuity level d, that is, $Y=\tilde{Y}+d\cdot 1_{[X\geqslant 0]}$ (which implies that $E[Y|X=x]=E[\tilde{Y}|X=x]+d\cdot 1_{[X\geqslant 0]}$, and we provide a graphical illustration in Section 2.3.2). Typically, the (g,d) combination is jointly referred to as the "data generating process," but we separate out the discontinuity level d and call g the DGP for ease of exposition.

We think of each (bivariate) RD dataset as a realization from the probability model (g,d). Realizations are indexed by ε, which can be thought of as the seed in statistical software. Taken together, the triplet (g,d,ε) uniquely determines an RD dataset and, in fixing the graphical parameters for said dataset, the quadruplet (γ,g,d,ε) determines an RD graph. We will refer to (γ,g,d), that is the set of all possible graphs of type γ resulting from a given probability model, as a "meta-graph."

When presented with the same RD graph, there are several reasons why participants may draw different visual inferences. Some participants may be more skilled than others at classifying a discontinuity because they have received more training in statistics, have more experience with RD graphs, or otherwise have superior ability. Participants may also apply different intrinsic "p-value" cutoffs, as some may be more likely than others to classify any given graph as having a discontinuity. We use ϕ to capture these human characteristics that affect graph perception, whether or not they are observed by us.

The probability that a participant with characteristics ϕ reports that a discontinuity ex-

ists in RD graph $(\gamma, g, d, \varepsilon)$ is denoted by $\tilde{p}(\gamma, g, d, \varepsilon, \phi)$. From casual observation, we know that the same participant may be influenced by idiosyncratic elements not encapsulated in ϕ and classify the same graph differently on different days. The probability formulation $\tilde{p}$ allows these factors to affect visual inference.

We now define the type I and type II error probabilities which we use to gauge participant performance. First, averaging $\tilde{p}$ over realizations ε and participant characteristics ϕ leads to the quantity

$$p(\gamma, g, d) \equiv E_{\varepsilon,\phi}[\tilde{p}(\gamma, g, d, \varepsilon, \phi)].$$

This is the probability that a randomly chosen participant reports a discontinuity in the meta-graph (γ, g, d). A high value of p indicates a high classification error probability for DGP g when the true discontinuity d is zero (type I error), but a low classification error probability when d is nonzero (type II error). Formally, the DGP-specific or g-specific type I and type II error probabilities for graphical parameter γ are defined as:

$$g\text{-specific type I error probability: } p(\gamma, g, 0)$$
$$g\text{-specific type II error probability: } 1 - p(\gamma, g, d) \text{ for } d \neq 0.$$

Conceptually, we can further average $p(\gamma, g, d)$ over the space of DGPs $\mathscr{G}$ (discussion of $\mathscr{G}$ after Assumption 1 below) to arrive at the *overall* discontinuity classification probability for γ:

$$\bar{p}(\gamma, d) \equiv E_{g \in \mathscr{G}}[p(\gamma, g, d)].$$

Correspondingly, the overall type I and type II error probabilities are defined as

$$\text{overall type I error probability: } \bar{p}(\gamma, 0)$$
$$\text{overall type II error probability: } 1 - \bar{p}(\gamma, d) \text{ for } d \neq 0.$$

In this paper, we design experiments to estimate the type I and type II error probabilities

as defined above. For each meta-graph (γ, g, d), we randomly sample M participants and realizations. Participant i is shown RD graph $(\gamma, g, d, \varepsilon_i)$, where i and ε_i both take on values in the set $\{1, ..., M\}$, and is asked to assess the presence of a discontinuity. Let the binary variable $R_i(\gamma, g, d, \varepsilon_i)$ denote participant i's discontinuity classification, which equals one if the participant reports a discontinuity at the policy threshold. If the M participants and realizations come from simple random sampling from the distributions of ϕ and ε, then the following assumption holds:

Assumption 1. *For a given DGP g, the $R_i(\gamma, g, d, \varepsilon_i)$'s are i.i.d. with $E[R_i(\gamma, g, d, \varepsilon_i)] = p(\gamma, g, d)$.*

A natural estimator for $p(\gamma, g, d)$ is the sample average of discontinuity classifications:

$$\hat{p}(\gamma, g, d) = \frac{1}{M} \sum_i R_i(\gamma, g, d, \varepsilon_i).$$

In Proposition 1 in Appendix B.1.1, we state the distribution of $\hat{p}(\gamma, g, d)$ and show the estimator to be unbiased and consistent as $M \to \infty$ for $p(\gamma, g, d)$ under Assumption 1.

To estimate the overall probability $\bar{p}(\gamma, d)$, we need to sample from the DGP space $\mathscr{G}$. While it is difficult to formally define $\mathscr{G}$, we think of the data used in empirical RD research as realizations when sampling from $\mathscr{G}$. To that end, we can specify J DGPs that approximate data from existing research and present graphs generated with discontinuity d for each DGP g_j $(j = 1, ..., J)$ to a distinct group of M participants for a total of $M \cdot J$ participants and visual discontinuity classifications.

Assumption 2. *The DGP g_j's are randomly sampled from $\mathscr{G}$.*

A natural estimator for $\bar{p}(\gamma, d)$ is

$$\hat{\bar{p}}(\gamma, d) \equiv \frac{1}{J} \sum_j \hat{p}(\gamma, g_j, d) = \frac{1}{M \cdot J} \sum_{i,j} R_i(\gamma, g_j, d, \varepsilon_i),$$

the average of discontinuity classifications across the $M \cdot J$ classifications. In Proposition

2 in Appendix B.1.1, we state the distribution of $\hat{\bar{p}}(\gamma,d)$ and show the estimator to be unbiased and consistent (as $J \to \infty$) for $\bar{p}(\gamma,d)$ under Assumptions 1 and 2 (given that $J = 11$ in our experiments, consistency here is a conceptual statement implying that were we to incorporate DGPs from more RD studies, our type I and type II error rates would be closer in probability to the population parameters of interest).

For brevity, we refer to the type I error probabilities $p(\gamma,g,0)$ and $\bar{p}(\gamma,0)$ as the DGP-specific and overall sizes of visual inference respectively, and refer to one minus the type II error probabilities—$p(\gamma,g,d)$ and $\bar{p}(\gamma,d)$ for $d \neq 0$—as the respective powers. This nomenclature is consistent with conventions in the statistics literature except for calling the overall type I error probabilities $\bar{p}(\gamma,0)$ size. Conventionally, size is defined as the *supremum* of the type I error probability over the set of DGPs, whereas $\bar{p}(\gamma,0)$ is an *average*. We choose not to adopt the conventional size definition for two reasons. First, finding the DGP that results in the highest type I error probability for visual inference is difficult. Second, that DGP may be very different from typical DGPs that underlie actual RDD datasets, which suggests that adopting the conventional size definition may be overly conservative.

We can also define the type I and type II error probabilities of the inference procedure based on a given econometric estimator, $\hat{\theta}$, but with two adjustments. First, γ is no longer an argument in these expressions because we directly apply $\hat{\theta}$ to the dataset (g,d,ε). Second, we need to specify the level of the testing procedure based on $\hat{\theta}$, which we set to 5%, the prevailing standard in empirical studies. Because the definitions of these probabilities and their estimators are similar to the quantities defined above, we omit them here.

In subsequent sections, we empirically trace out $\hat{\bar{p}}(\gamma,d)$ as functions of d, which are *power functions* that concisely summarize the type I and type II error probabilities. We evaluate visual inference by comparing its power functions $\hat{\bar{p}}(\gamma,d)$ across γ, between expert and non-expert populations, and against the corresponding power functions for various econometric inference procedures. We discuss the implementation of the inference procedure on the differences between the visual and econometric inference functions when we

present our empirical results in Section 2.4, as its details depend on the design of our experiments.

In summary, we use the type I and type II error rates to measure the performance of an inference method. The type I error rate for RD visual inference is the fraction of continuous graphs participants incorrectly classify as having a discontinuity, and the type II error rate is the fraction of discontinuous graphs incorrectly classified as being continuous. Our framework allows us to interpret these rates as unbiased and consistent estimates of the type I and type II error probabilities, which represent the type I and type II probabilities a randomly chosen person commits when classifying a graph generated from a representative DGP.

2.3 Description of Experiments and Studies

2.3.1 Graphical Parameters Tested

We test the effects of bin width, bin spacing, parametric fit lines, vertical lines at the policy threshold, and y-axis scaling. Additionally, we test two RK-specific parameters: detrending the data with a global linear regression prior to creating the graph and plotting estimated first derivatives instead of the levels of the data so that a kink appears as a discontinuity. We discuss each of these treatments in detail below.

The most studied graphical parameter in RD is the width of each bin in the binned scatter plot. The first class of bin width selection algorithms comes from Lee and Lemieux (2010): start with some number of bins, double that number, test whether the additional bins fit the data significantly better, and repeat until the test fails to reject the null. Calonico, Cattaneo, and Titiunik (2015) propose two bin width selection algorithms based on different econometric criteria. The first and more in line with the convention of the nonparametric regression literature is the bin selector that minimizes the IMSE of the bin-average estimator of the CEF, where the resulting number of bins increases with the sample size n at the

rate $n^{1/3}$. For their second bin selector, the MV selector, Calonico, Cattaneo, and Titiunik (2015) state that they "choose the number of bins so that the binned sample means have an asymptotic (integrated) variability approximately equal to the amount of variability of the raw data." The resulting number of bins increases with the sample size more quickly, at the rate $n/\log(n)^2$. The IMSE-optimal bin width therefore selects fewer bins, and hence has larger bin widths, than the MV algorithm. In addition to these two algorithms, Calonico, Cattaneo, and Titiunik (2015) provide an interpretation for any given number of bins as the output of a weighted IMSE-optimal algorithm. We use this result later to extend the MV bin selection algorithm to plotting estimated derivatives in RKD. Applying the Lee and Lemieux (2010) algorithms to our datasets typically leads to bin numbers in between the IMSE and MV selectors that tend to be closer to those of the IMSE. We therefore restrict our analysis to the IMSE and MV bin selectors which we describe in more detail in Appendix B.1.2.

Although the prevailing approach is to adopt evenly spaced bins, with 98% of the more than 100 RDD and RKD studies we compile and review for current practices taking this approach, this method has drawbacks in that the resulting bins may contain vastly different numbers of observations, or even none at all. This can happen when the distribution of the running variable is far from uniform. As a remedy, Calonico, Cattaneo, and Titiunik (2015) also propose quantile-spaced bins where each bin contains the same number of observations. Both spacings support IMSE and MV bin selectors, and we test each of these combinations.

Following suggestions by Imbens and Lemieux (2008) and Lee and Lemieux (2010), RD and RK graphs frequently feature parametric fit lines and a vertical line at the treatment threshold. Both papers suggest fit lines improve "visual clarity" by approximating the conditional expectation functions, and the default in the popular `rdplot` command by Calonico, Cattaneo, and Titiunik (2015), for example, uses piecewise global quartic regressions on each side of the policy threshold. 10 of the 11 RDD papers on which we calibrate

our DGPs include fit lines. For the six of these papers that generate fit lines using polynomial regressions, we use the same polynomial order as in the source graph, and in the remaining cases, our team unanimously decides on the fit that best matches the original fit line or the data (we could also use a formal data-driven approach to select the polynomial order, but different criteria from Lee and Lemieux (2010) lead to widely different recommendations as shown in Table B1). The presence of a vertical line at the cutoff is designed to visually separate observations above and below that cutoff. We test all combinations of these two treatments except for including the fit line and not having a vertical line, which is used infrequently in practice (our literature review shows that fewer than 10% of papers use this combination).

The motivation for rescaling graph axes comes from Cleveland, Diaconis, and McGill (1982), who note that correlations on scatter plots seem stronger when scales are increased (see Figure B1 for an example). We use two axis scaling options in our experiments. First is the default output returned by Stata 14. Second, we double that range by recording the range of the *y*-variable from the default graph, increasing the bounds by 50% of the original range in each direction, resulting in a graph where the data are condensed along the vertical axis. We do not manipulate the scale of the *x*-axis because our survey of the literature suggests that this is not a common adjustment. A related decision researchers encounter is the range of the running variable to use in producing graphs: should they use the entire dataset or only a subsample close to the policy threshold? We do not test this margin of adjustment in our experiments due to the difficulty in generalizing the findings from such an exercise. Suppose we find that selecting 50% of the observations closest to the threshold improves visual inference, should researchers "chop" the sample they are already planning to use? And after doing so, will it be beneficial to chop again? One could argue for testing the effect of using the full sample versus the subsample falling within the IK or CCT bandwidth (see Section 2.4.4 below for review), but these bandwidths themselves depend on the full sample—the first step in bandwidth calculation is a (semi-)global regression—

and it is not even clear how we should define the "full" sample. In fact, some of the replication data used for our DGP calibration are already subsets of a larger sample.

There are other RD graphical parameters we do not test in our experiments. One is plotting confidence bands around the binned averages or fit lines. However, confidence intervals are too complex to explain to the non-experts in our short tutorial without potentially affecting the way participants think about the classification task itself, and therefore we do not experimentally test their effects on visual inference. Because the moderate size of our expert pool makes it unsuited to randomized experiments, we refrain from testing other graphical parameters.

The first of our RK-specific treatments is detrending the data. This is simply running a global linear regression and plotting the residuals instead of the original data, as shown in Figure B2. Plotting these residuals may make kinks more visually apparent and is adopted by Lee et al. (2019).[3]

We test one new treatment that has not previously been used in practice: estimating and plotting local first derivatives rather than the levels of the data. We follow Cattaneo, Farrell, and Feng (2020), who examine partitioning-based nonparametric regression and derive an IMSE-optimal number of partitioning knots to estimate the level or derivatives of a CEF. As in conventional RD and RK plots, the knots are determined separately on each side of the policy threshold with no overlap. Within each bin defined by these knots, we run a linear regression, and for each bin, we plot the corresponding linear coefficient. A kink in levels then appears as a discontinuity in first derivatives, as in Figure B3. Because the MV bin selector does not apply to this setting, we instead extend the results from Cattaneo, Farrell, and Feng (2020) to approximate a small-width bin selector. For each of our RD datasets, we compute the implied weights that justify the MV bin selector as weighted-IMSE-optimal, take the median value, and use that weight for a weighted IMSE calculation of how many bins to use for the small bins treatment when plotting estimated derivatives for RK.

[3]In principle, one could detrend by residualizing against any smooth function; we use linear detrending for simplicity.

Because the RKD DGPs we use are based on large administrative datasets, as is typical for RKD, both bin selectors will generally select many more bins than in the RD setting. Using equally spaced MV bins tends to result in many near-empty bins in the sparse tails of our datasets, as in Figure B4, which likely hinders kink perception at the policy threshold. Because such graphs never appear in papers implementing RKD, we use quantile spacing for all kink graphs.

2.3.2 Creation of Simulated Datasets and Graphs

For both RDD and RKD, we specify data generating processes based on the actual data used in published research. For the RDD components of our study, we randomly sample 11 from a set of empirical RD papers published in the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal of Business and Economic Statistics, Journal of Political Economy, Quarterly Journal of Economics, Review of Economic Studies,* and *Review of Economics and Statistics* that have replication data available to create our DGPs. We refer to these DGPs as RD DGP1-DGP11.

The calibration of each DGP g entails the specification of its four components: the distribution of the running variable X, the continuous conditional expectation function $E[\tilde{Y}|X = x]$, the distribution of the error term u, and the sample size N. Out of the 11 studies, two plot residualized outcomes to account for covariates, conceptually consistent with the covariate-adjusted RD estimation by Calonico et al. (2019) and the ideas of Angrist and Rokkanen (2015). We use the empirical distribution of the running variable from each of the 11 papers but normalize it to lie in $[-1,1]$ with 0 representing the treatment cutoff. We also remove the most extreme observations where $|X| > 0.99$, following Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014). For two papers which feature semi-discrete running variables, we add small amounts of normal noise to the running variable to match the regularity conditions from Calonico, Cattaneo, and Titiunik (2015). To create a continuous CEF, we fit global piecewise quintics (still following

Imbens and Kalyanaraman, 2012 and Calonico, Cattaneo, and Titiunik, 2014) and vertically shift the right arm. We specify the distribution of the error term u as i.i.d. normal with mean zero and standard deviation σ, which we set as the root mean squared error of the piecewise quintic regression. We use the same number of observations as the original paper minus any observations removed while trimming the data. Plots of the resulting CEFs before we vertically shift their right arm to make them continuous are in Figure B5, and Figure B6 illustrates the construction process. We leave the full details of DGP specification to Appendix B.2.

Because the outcomes from the 11 papers are measured in different units, we specify discontinuity levels d as multiples of σ. d takes on 11 values: $0, \pm 0.1944\sigma, \pm 0.324\sigma$, $\pm 0.54\sigma, \pm 0.9\sigma, \pm 1.5\sigma$. We include $d = 0$ in order to measure type I error rates. We choose the upper bound $|d| = 1.5\sigma$ based on our own visual judgment: it represents the point at which we expect every reasonable person to say a graph from any of our 11 DGPs features a discontinuity. The nonzero magnitudes of d are equally spaced on the log scale. We use a log rather than linear scale to generate more graphs with smaller discontinuities, which are harder to detect, in order to better capture the shape of the power functions.

Our discontinuity magnitudes are similarly distributed to those observed in the main outcome graphs in our literature review. The average absolute value of the t-statistics in our datasets from piecewise quintic regressions is 5.0 with a standard deviation of 5.3, compared to the observed mean of 3.9 with a standard deviation of 6.4. If we instead compare the distributions of the absolute value of the discontinuity divided by the control magnitude (the left intercept of the CEF), the means are similar, 1.3 in our datasets and 1.9 in the field, while our standard deviations are somewhat smaller at 2.2 compared to 7.7.

In the RKD case, most studies use restricted administrative datasets, resulting in a lack of publicly available replication microdata. Therefore, we fit 11 DGPs based on three studies: Card, Lee, and Pei (2009), Card et al. (2015c), and Card et al. (2015a). The DGPs are fitted on seven subsamples: the reduced-form sample of Card, Lee, and Pei (2009),

the pre- and post-recession reduced-form samples of Card et al. (2015c), and the top and bottom kink first-stage and reduced-form samples of Card et al. (2015a). Although the latter two papers use restricted data from Austria and Missouri respectively, a member of our team has specified the DGPs from past work on these projects.

The process for calibrating RK DGPs is similar to that for RD. Data confidentiality creates a complication in that the actual data cannot be shared among team members, so instead of using the X variable from the actual data, the running variable used to create each graph is sampled from an estimated kernel density. For each of the seven samples, we obtain a CEF from piecewise quintic regressions like for their RD counterparts, except here we impose continuity as implied by the continuous treatment rule and the economic model in Card et al. (2015a). In addition, for each of the four subsamples of Card et al. (2015a), we follow the Monte Carlo simulations of Card et al. (2015b) and Card et al. (2017) and specify an additional CEF from a *global* quintic regression. We create these four DGPs that are smooth at the policy threshold out of concern that using only the other seven may be too favorable to visual inference based on the results of an RKD pilot (the pilot revealed both very low type I error rates and high power in a setting that did not require participants to distinguish between curvature and kinks at the policy threshold). We plot all resulting CEFs in Figure B7. The CEFs for RK DGPs 5, 6, 9, and 10 are the global quintics, while the rest are piecewise quintics. We create CEFs with different kinks by rotating the right arm of the unkinked CEF, and we specify the kink magnitude as multiples of ξ, which denotes in radians the asymptotic standard deviation of the angle estimator between the left and right arms of the CEF from a piecewise quintic regression. We use nine different kink levels: $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$. Similar to our RD DGPs, we also use a normal distribution for the error term u, but the specification of its standard deviation is more complex. We discuss in detail the calibration of our RK DGPs in Appendix B.2.

As argued in Section 2.2, we want our DGPs to be representative. Although we select the papers randomly, we also need to evaluate how well our DGPs approximate the actual

data. To do this, we adapt the lineup protocol from Buja et al. (2009) and Majumder, Hofmann, and Cook (2013), which uses visual inference to conduct hypothesis testing.[4] In our case, we test the null hypothesis that the original datasets come from the calibrated DGPs. Specifically, we present one graph of the original data randomly placed among 19 graphs from datasets drawn from the corresponding DGP. The goal is to identify the true dataset by choosing the graph that least resembles the others. If the viewer does not select the original graph, then we cannot reject the null hypothesis. The probability of identifying the graph produced from the original data (or the type I error probability) among the 19 simulated datasets is 5% $(1/20)$ for a single reader. For our lineup protocol, each graph is a binned scatter plot using the MV bin selector. An example of the combined plots is in Figure B8, with the coordinates of the true graph in this footnote.[5]

Based on visual testing among the authors, we cannot identify the graph from the true data for eight out of our 11 RD DGPs, which supports the idea that our DGPs approximate the original datasets well. For three DGPs, however, there is an obvious difference, such as in Figure B9. All three "fail" seemingly because of the misspecification of the variance structure of the error term u. Recall that we specify u as being i.i.d. across observations and homoskedastic. But in Figure B9, for example, the running variable is time, and there is positive serial correlation in the outcome. As a consequence, the outcome variability in the binned scatter plot is understated when u is assumed to be i.i.d. Nevertheless, we adhere to the i.i.d. specification because it is standard in Monte Carlo exercises to evaluate RD estimators and makes it straightforward to implement the uniformly most powerful test in Section 2.4.

Another caveat of our DGP specification is that using quintic regressions can lead to overfitting. Some of our graphs feature high variation across bins in the tails due to the sensitivity of the fitted piecewise quintics. We follow the practice of Imbens and Kalyanara-

[4]Because of confidentiality concerns with the RK microdata from Card et al. (2015c) and Card et al. (2015a), we do not present lineups for the RK DGPs.

[5]The graph produced from data used in the original paper is in row $-3 \cdot 2 + 7$ and column $\sqrt{4+5} - 2$. The remaining graphs are generated from our specified DGP.

man (2012) and Calonico, Cattaneo, and Titiunik (2014) in our DGP specification while acknowledging this potential drawback of using quintics. These caveats notwithstanding, the lineup protocol offers a novel and transparent method to evaluate DGP specifications, and it is because of its usefulness that we can identify the aforementioned limitations at all.

2.3.3 Non-Expert Experiments

In our randomized experiments, we present non-expert participants with binned scatter plots made from our DGPs and ask them to classify the graphs as having a discontinuity/kink or not. We conduct five phases of computer-based experiments online through Cornell University Johnson College's Business Simulation Lab. Our subject pool consists of current and former Cornell students, Cornell staff, and non-student local residents with an expressed interest in focus groups or surveys. Although these educated laypeople are not the primary audience for academic research, RD graphs are sufficiently transparent that they are featured in popular media articles in publications such as *The New York Times*, *The Washington Post*, and *The Atlantic* (Dynarski, 2014; Sides, 2015; Rosen, 2015), suggesting the participants in our sample should be capable of interpreting the graphs.

Before the experiment, participants watch a video tutorial explaining how the graphs are constructed.[6] We do not instruct participants on how to make their decisions, e.g. whether only to look at points near the cutoff or mentally to trace out the CEF. The video contains an attention check with a corresponding question later in the experiment to ensure that subjects pay attention to the instructions. After the video, participants complete a series of interactive example tasks and receive feedback on their answers. As part of the instructions, we tell the participants explicitly that all, some, or none of the 11 graphs they classify may feature a discontinuity/kink.

In each phase of the experiment, we present participants with a series of RD or RK graphs using data generated as described in Section 2.3.2. In the RD experiments, partici-

[6]The video tutorial is available at `https://www.dropbox.com/s/3q5bj0adxis4vf2/RD_video_tutorial.mp4?dl=0`.

pants see two graphs with zero discontinuities, one each of $\pm 0.1944\sigma, \pm 0.324\sigma, \pm 0.54\sigma, \pm 0.9\sigma$, and one of either 1.5σ or -1.5σ. Participants see one graph from all 11 DGPs in a randomized order. We have up to 88 participants per treatment arm, and every graph we generate is seen by only one participant. The setup is similar for RK, except we present each participant with 11 graphs plotting levels of the data and another 11 graphs plotting estimated first derivatives, for a total of 22 classification tasks. For each graph, we ask participants whether they believe there is a discontinuity (or kink for RK) at $x = 0$. To stimulate participant engagement and elicit participants' confidence in their response, participants can choose a bonus that is either based on a monetary wager which pays 40 cents if their judgment is correct but nothing otherwise, or a fixed payment of 20 cents irrespective of their performance. Under this bonus scheme, risk- or loss-averse participants will choose the fixed payment whenever their subjective probability of being correct is close to 50% (see Appendix B.3.1 for details).

Because running an experiment with $2^5 = 32$ RDD treatment arms is infeasible with our resources, we conduct our experiment in phases, testing only a few treatments in each phase. To inform best practices in graphical representation, we are mainly interested in the direction of the effects of graphical parameters, e.g. whether using small bins increases or decreases type I error rates relative to using large bins. Therefore, we assume that our graphical parameters do not have "qualitative interactions" (see Gail and Simon, 1985), i.e., the direction of the effect of one parameter does not change with other parameters. We can test this assumption on the two treatments that we interact in each phase, and it appears to hold.

Tables B2 and B3 list the graphical parameters we test and hold fixed for various experimental phases. In Phase RDD1, we test both bin width and axis scaling options. In Phase RDD2, we test bin widths and bin spacings. In Phase RDD3, we test imposing fit lines and a vertical line at the treatment threshold. In Phase RKD, we test combinations of bin widths and detrending the data and separately test bin widths for graphs of local

derivatives. Because RKD participants see graphs both with levels of the data and estimated first derivatives, we have a between-subjects design for the RK graphical treatments of bin width and detrending, but a within-subject design for plotting levels and derivatives.

Participants receive a base pay of $3 for being in the experiment and a bonus of 0, 20, or 40 cents per classification depending on the participant's wager selection and whether they are correct. We do not give participants real-time feedback but do report earnings and the total tally of correct classifications after a short exit survey soliciting demographic information and comments at the end of the experiment. The study takes participants approximately 15 minutes to complete.[7]

2.3.4 Expert Study

In addition to our non-expert experiment, we conduct a study with researchers in economics and related fields who work on topics that often employ RDDs and RKDs. We collect data at three technical social science seminars and online by contacting randomly selected members of the National Bureau of Economic Research in applied microeconomic fields (aging, children, development, education, health, health care, industrial organization, labor, and public) and Institute of Labor Economics (IZA) fellows and affiliates. After removing a total of six responses from participants who completed the survey more than once, did not provide a valid email address for payment, or were not part of our recruited sample, we are left with 143 expert responses for RD and 48 for RK.

This expert study allows us to answer two questions. First, how do classification accuracy and the impacts of graphical techniques differ between experts and non-experts? And second, can experts correctly predict which graphing options perform best for our non-expert sample? This second question speaks to experts' ability to predict which visualization choices are most suited for interpretation by a lay audience. Because the success

[7]The experiments are programmed in oTree (Chen, Schonger, and Wickens, 2016) and pre-registered at the AEA RCT registry (Korting et al., 2019a) and the Center for Open Science's OSF platform (Korting et al., 2019b).

of a graphical technique ultimately lies in the reader's correct perception of graphs using it, it is important to understand whether experts' intuition regarding the relative advantages and drawbacks of alternative representation choices aligns with the evidence we find in practice. In related work on experts' ability to predict non-expert performance, DellaVigna and Pope (2018) find that economic experts are better than non-experts at estimating the effect of alternative incentive schemes on performance in a real effort task, but perform similarly to non-experts in terms of a simple ranking of incentive schemes.

Our expert study consists of two parts. The first is similar in structure to the non-expert experiment. Participants see a series of RD or RK graphs and are asked to classify them as (not) having a discontinuity/kink. To assess the accuracy of point estimates in addition to binary classifications of discontinuities, we also ask participants in the RD module for an estimate of the discontinuity magnitude whenever they report a discontinuity. Due to sample size limitations, we do not randomize graphical treatments in the expert study, and all participants see graphs with no fit lines, default axis scaling, and a vertical line at the treatment threshold. All expert RD graphs use equally spaced small bins, except for one seminar where participants see large bins, and all expert RK graphs use large bins with quantile spacing. Four randomly selected participants receive a base payment of $450 plus a bonus payment of $50 per correct discontinuity/kink classification. The bonus payment in the RD case does not depend on the accuracy of the magnitude estimate.

The second part of the expert study asks about experts' preferences and their beliefs regarding non-expert performance across alternative graphical parameters. We present experts with three discontinuity or kink magnitudes: $0, 0.54\sigma, 1.5\sigma$ for RD and $0, 0.324\xi, 0.9\xi$ for RK. For each RD magnitude, we present four graphs using the same underlying data, one for each combination of bin width and fit lines. We show graphs from the DGP where visual inference performs most closely to the average over all DGPs. For RK, we show only two graphs with both bin width options and without fit lines. We randomize graph treatment order between participants. At each magnitude, we ask the

experts to indicate which of the four (two) treatment options they prefer and which of the four (two) treatment options they believe perform best and worst in our non-expert sample. To evaluate the experts' predictions about non-expert performance, we run a supplemental phase RDD4 in which we test all four treatments simultaneously across subjects. This is not necessary for RK, where we test all treatments in one phase.

2.3.5 Eyetracking Study

A followup to our observed performance differences across individuals is to ask why some people are better at classifying discontinuities than others and whether visual patterns can predict performance. To answer these questions, we conduct an eyetracking study through Cornell University's Mobile Communication Lab. Eyetrackers provide insights into cognitive processes by tracing participants' eye patterns as they navigate visual information. These eye patterns are typically differentiated into *saccades* (rapid eye movements) and *fixations* (maintaining visual gaze on a single location). Fixations are believed to be a valid indicator for attention during complex information processing tasks such as reading (Just and Carpenter, 1980; Rayner, 1998) and have previously been used to study how physicians analyze electrocardiograms (Bond et al., 2014), how children "read" bar graphs (Garcia Moreno-Esteva et al., 2017), and how alternative presentations of health messages online improve recall (Bol et al., 2016).

The structure of the eyetracking study is similar to that of our non-expert experiment, and both experts and non-experts participate in the study. We recruit non-expert participants from the same Business Simulation Lab subject pool. All non-expert subjects receive the same instructions and training as in the online experiment before completing the classification task at a physical eyetracking terminal (see Appendix B.3.3 for details). We recruit experts from a pool of experienced graduate students and faculty at Cornell University. Expert and non-expert participants receive participation fees of $10 and $5, respectively, plus 40 cents for each correctly classified graph. In total, we recruit 11 non-expert and six

expert subjects. We drop one non-expert participant from the sample because of the poor quality of their eyetracking data resulting from a low calibration score.

The eyetracking approach allows us to investigate the extent to which visual patterns predict participants' correct classification of a graph. Similar to the RD phases of our primary experiment, we present subjects with a series of 11 RD graphs, one for each DGP, but this time showing the same graphs in the same order to all participants. We keep the set of graphs fixed because the visual patterns that lead to correct discontinuity classification may vary from graph to graph, and due to the risk of visual fatigue affecting eye movements over time, we also keep the same graph order between subjects.

2.4 Results

2.4.1 RD Non-Expert Experiment Results

For each combination of graphical parameters in all phases of the regression discontinuity experiment, we compute power functions based on participants' classifications of graphs as having or not having a discontinuity. Using notation from Section 2.2, a DGP-specific power function represents the estimates $\hat{p}(\gamma, g, d)$ for graphical parameters γ and DGP g across different levels of discontinuity d. An overall power function represents the estimates $\hat{\bar{p}}(\gamma, d)$.

The intercept of a power function indicates the type I error rate or size as defined in Section 2.2. At all other discontinuity magnitudes, the power function represents the proportion of graphs with discontinuities that participants classify correctly, which can be interpreted as the statistical power in detecting the discontinuity when the DGP is chosen uniformly randomly from the 11 possibilities. The ideal inference method has a small intercept before quickly rising to high power. Plots of the overall and DGP-specific power functions for each phase are in Figures 2.1 and B10 respectively. The x-axis in these graphs is the magnitude of the discontinuity divided by the DGP-specific σ. This normalization

facilitates aggregation and comparisons across DGPs, which then have six identical discontinuity magnitudes: 0, 0.1944, 0.324, 0.54, 0.9, and 1.5. Estimated treatment effects are in Tables B4-B7. Because we effectively adopt stratified randomization in the design of our experiments as described in Section 2.3.3, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude, we obtain these estimates by regressing the participants' responses on treatment indicators and stratum fixed effects.

Phase RDD1 tests the four combinations of the bin width treatments (large IMSE-optimal bins and small MV bins) with the *y*-axis scaling treatments (the default in Stata 14 and double that range). Comparing power functions, the large bins have significantly higher type I error rates relative to the small bins. While both small bin treatments lead to a size of approximately 5%, the large bins have a size of around 20% to 25%. The large bins have a power advantage over the small bins, but the treatments converge in power as the discontinuity magnitude increases. In contrast to bin widths, axis scaling has little effect on participant perception. Based on this result, we use Stata's default scaling for all subsequent phases.

In Phase RDD2, we again test the two bin width treatments, this time interacted with the two bin-spacing treatments: even spacing and quantile spacing. Note that large bins and small bins with even spacing appear in both Phases RDD1 and RDD2. With this design, we can gauge the stability of visual inference across different samples from the non-expert population, and it is encouraging to see the results for these two treatments being virtually identical across phases. Comparing the power functions for these repeated treatments with their new quantile-spaced versions, we see that the evenly spaced and quantile-spaced power functions are very similar. We conclude that bin spacing has a small or null effect on visual inference for the DGPs we test.

Phase RDD3 tests three treatments: the inclusion of a vertical line at the treatment threshold with and without polynomial fit lines and the omission of both the vertical line and fit lines. We find that the vertical line at the cutoff makes little difference in perception.

Fit lines, on the other hand, appear to increase both size and power in this phase, in line with a common concern that they may be overly suggestive of discontinuities.

Jointly, these three phases of experiments suggest that the presence of fit lines and the bin width choice have the largest impact on visual perceptions of discontinuities. We therefore base our analysis of expert preferences and expert predictions about non-expert performance on the interaction of these two treatments and run a final phase of experiments, Phase RDD4, directly comparing the four possible treatment combinations. Interestingly, while the effects of bin width choice are once again robust across phases, the effects of fit lines are more muted in this phase. In particular, the treatment with small bin and fit lines has a size of only 0.052 in Phase RDD4 but 0.175 in Phase RDD3. This finding suggests that we cannot conclude that fit lines unequivocally result in an increase in type I errors.

We provide additional results in Appendices B.4 and B.5. Appendix B.4 includes evidence for the balance of covariates across treatment arms and the measurement of predictive power of demographic and DGP characteristic variables on visual inference performance. Appendix B.5 provides an answer to the question of how large the t-statistics need to be for readers to visually detect a discontinuity.

2.4.2 RK Non-Expert Experiment Results

We conduct one phase of the non-expert experiment focusing on the regression kink design. Here we test two separate sets of treatments. In the first set, we compare all combinations of IMSE and MV bin widths with detrending and not detrending the data. Power functions based on the raw or detrended data are plotted in Figure 2.2 with the treatment effect estimates in Table B8. Type I error rates are universally high for kinks, around 15% to 20%. Breaking the DGPs into the seven piecewise quintics and the four smooth global quintics with peaks at $x = 0$ in Figure B11 shows that the high type I error rate is driven by the smooth DGPs. Small bins again decrease the type I error rate, but the effect is much smaller than for discontinuities at about one percentage point. Detrending the data prior to

plotting has positive effects on power while maintaining a similar type I error rate.

The power functions from plotting the estimated first derivatives in the data and identifying a discontinuity in them are also shown in Figure 2.2 with corresponding treatment effect estimates in Table B9. Results for this treatment are not compelling, with high rates of both type I and II errors. For the small bin treatment, the size is above 20% and the power peaks at around 30%, while the large bin treatment's size is almost 40% and the power does not reach 70%. These power functions are dominated by the levels treatments, and we advise against plotting estimated derivatives in RKD applications.

2.4.3 Expert Study Results

RD Expert Study Results

For the experts in our RD sample, we show most participants (95 out of 143) graphs with small bins, no fit lines, even spacing, default y-axis scaling, and a vertical line at the policy threshold, which have the lowest type I error rate in our non-expert experiment. We present the expert power functions and the non-expert power functions based on graphs generated with the same parameters in the left panel of Figure 2.3, and we plot the differences between the power functions in Figure B12. This figure also includes pointwise 95% confidence intervals generated by applying the large sample approximation described at the end of Appendix B.1.1 and by assuming independence between the experts and non-experts. The two groups perform similarly, with experts having slightly worse size (approximately 8% to the non-expert 5%) and slightly greater power. The only statistically significant differences are for the experts' marginally superior powers at the 0.1944 and 0.324 discontinuities.

In addition to the aforementioned treatment, we show experts in one seminar pool (48 out of 143) graphs using the large bins and no fit lines treatment. We compare them with non-experts under the same treatment in the left panel of Figure 2.3. The two groups perform similarly. Experts again have a slightly worse type I error rate than non-experts

here, and both are well above the rates for experts and non-experts with small bins. The expert and non-expert power functions are not statistically significantly different anywhere.

RK Study Expert Results

Experts in our RK sample see graphs with large bins, no fit lines, quantile spacing, and a vertical line at the policy threshold. In addition to using quantile-spaced bins as mentioned in Section 2.3.1, we use only large bins because on our kink datasets based on large administrative data, the small bin algorithm selects almost 1,000 times more bins than the large bin algorithm, resulting in graphs that do not look like those that commonly appear in RKD papers (for RD, the small bin algorithm only selects about 16 times more bins). We compare their performance with that of non-experts in the same treatment in the right panel of Figure 2.3 and Figure B13. Although there is virtually no size difference, experts have substantially better power for moderately sized kinks.

Expert Preferences and Predicting Non-Expert Performance

For RD, we present experts' preferences and their beliefs about non-expert performance across the four considered treatments in Figure B14. When asked about graphing options for the main graph of a paper that conveys the treatment effect, most experts report preferring small bins, usually with fit lines. These results hold at all three discontinuity magnitudes considered, including zero. Experts' predictions about the most effective treatments for non-experts tend to mirror their preferences. By a large margin, experts believe small bins with fit lines to be the most efficacious treatment for non-experts at all discontinuity magnitudes. Conversely, most experts view large bins without fit lines least favorably in the context of non-expert performance.

Comparing the expert predictions for RD to our experimental data from Phase RDD4, we find substantial discordance for the effects of bin width choice on non-expert classification accuracy. The best- and worst-performing treatments at each discontinuity magnitude

have + and - signs, respectively, in Figure B14. The actual power functions are shown in Figure 2.1, and the ones based only on the DGP used in the example graphs shown to experts are in Figure B15. While a majority of experts correctly identifies the bin width treatment with the best size control (i.e. most experts prefer small bins at the zero discontinuity level, either with or without fit lines), there is also significant expert support for the large bin with fit lines treatment, even when there is no discontinuity, which exhibits the greatest type I error rate in our sample. In addition, experts fail to predict the type I vs type II error tradeoff presented by the bin width choice: most experts expect large bins to perform worst even at large discontinuities, while we find this treatment arm has the greatest power in those cases. Although in the actual power functions, the effects of bin width are much more pronounced than the effect of fit lines, we find more expert disagreement regarding non-expert performance along this dimension. We also find expert predictions to be similar whether their own visual inference performance is above or below the median.

Figure B16 shows experts' personal preferences and beliefs about non-expert performance for RK graphs with small bins and large bins, both without fit lines. For both themselves and the non-experts, experts clearly prefer large bins at every kink magnitude. Experts pick only one of the three best performers in the non-expert results in Figure 2.2, though differences between these treatments are generally smaller than in the RD setting, making prediction more difficult.

2.4.4 Visual vs Econometric Inference

RD Visual vs Econometric Inference

In this section, we compare the performances of visual inference from our small bin expert sample with econometric inference based on various RD estimators. For each estimator, we present graphs of both the overall power functions and the difference between them. For a fair comparison, we base the estimators' power calculations on their rejection deci-

sions over the same set of datasets underlying the graphs seen by the experts. That is, the estimators "see" the same data as the experts, preventing differences driven by variation in sampling from the same DGP.[8]

As a benchmark, our first estimator comes from a correctly specified model: a global piecewise quintic regression with homoskedastic standard errors. The power function for a 5% test based on this estimator compared with human performance is presented in the left panel of Figure 2.4 with the differences between them in Figure B17. The midpoint of each vertical bar in Figure B17 represents the difference between point estimates from regressions of the inference decision on indicators for each discontinuity magnitude. We compute confidence intervals for these differences by assuming independence between visual and econometric inferences.[9] We cluster the visual inferences at the participant level and the econometric inferences at the DGP-seed level (there are 88 such combinations). The type I error rate is just under 8% for the experts and approximately 6.8% for the piecewise quintic, which is near 5% by design. The piecewise quintic estimator has higher power, though the difference is only significant at the 0.324 discontinuity.

Next, we implement the IK, CCT, and AK inference procedures, again plotting the corresponding power functions in the left panel of Figure 2.4 and their differences in Figure B17. All three procedures build upon local linear regressions but take different approaches to conduct inference.

Strictly speaking, Imbens and Kalyanaraman (2012) do not study inference but propose an MSE-optimal bandwidth selector, the IK bandwidth:

$$\hat{h} = C \cdot \left(\frac{\hat{\sigma}^2(0^+) + \hat{\sigma}^2(0^-)}{\hat{f}(0) \cdot (\hat{\mu}^{(2)}(0^+) - \hat{\mu}^{(2)}(0^-))^2 + \hat{r}} \right)^{1/5} \cdot n^{-1/5}$$

[8] Out of the 11 RD studies, eight report main treatment effects and plot the data without controlling for covariates, and two report covariate-adjusted estimates and plot residualized outcomes. For these ten studies, visual and econometric inferences are based on the same data input, and our DGP specifications allow for an apples-to-apples comparison.

[9] As we show in Appendix B.6, we cannot reject the independence assumption at the zero discontinuity level. Visual inference tends to be positively associated with econometric inference at nonzero discontinuity levels, and thus the confidence intervals for the differences are conservative there.

where C is a constant determined by the kernel, $f(\cdot)$ is the density of the running variable, $\mu^{(2)}(0^{\pm})$ and $\sigma^2(0^{\pm})$ represent the second derivatives of the CEF and conditional variance on both sides of the threshold, r is a regularization term to prevent very large bandwidths, and the hats on the various quantities indicate that they are estimated. We refer to the "conventional" (terminology from Calonico, Cattaneo, and Titiunik, 2014) inference procedure practitioners typically implement in conjunction with the IK bandwidth as the "IK" inference procedure. Specifically, letting τ denote the true RD parameter and $\hat{\tau}$ the local linear estimator, we have

$$\sqrt{nh}(\hat{\tau} - \tau - h^2 B) \Rightarrow N(0, \Omega) \tag{2.1}$$

where B denotes the bias constant, which is a function of the second derivatives of the CEF on both sides of the threshold. Even though the normal distribution is centered around the constant $\tau + \sqrt{nh^5}B$ when h shrinks at the optimal rate $h = O(n^{-1/5})$ (as is the case for IK), "conventional" inference ignores the bias term $\sqrt{nh^5}B$. As seen in Figure 2.4, the 5%-level IK procedure is powerful, even more so than the piecewise quintic estimator, and is significantly better than visual inference at detecting discontinuities up to 1.5σ. But with this power comes a significant size distortion. When there is truly no discontinuity, the estimator still rejects the null hypothesis in 22.6% of datasets.

Third, we assess the performance of the inference procedure proposed by Calonico, Cattaneo, and Titiunik (2014) (CCT) as implemented in the `rdrobust` Stata package. This procedure differs from IK and creates robust confidence intervals by estimating the bias B in expression (2.1) using another ("pilot") bandwidth b, creating a bias-corrected estimator $\hat{\tau}^{bc} = \hat{\tau} - h^2\hat{B}$, and accounting for the sampling variation in $\hat{B}$ under unconventional asymptotics where h/b converges to a positive constant. Calonico, Cattaneo, and Titiunik (2014) also generalize IK and propose a new class of MSE-optimal bandwidth selectors, which we refer to as the CCT bandwidth(s). Note that although CCT's size is approximately 12.5%, as seen in the left panel of Figure 2.4, this could be a small sample problem.[10] Because

[10]While the CCT bandwidth remains the default and modal choice of `rdrobust`, new work by Calonico,

there are only eight draws from each DGP, i.e. ε takes only eight values for each g, we effectively have only 88 datasets, and all RD datasets within a DGP-seed pair are identical modulo the discontinuity level. In a separate Monte Carlo simulation with 1,000 draws, the size is in line with that of the experts at 7.0%. However, it is possible that the type I error rate of visual inference also decreases over graphs based on these alternative datasets—the realization of the disturbance term can impact both econometric and visual inference—and therefore we keep the comparison based on the 88 datasets. The CCT inference procedure is powerful, enjoying a significant 15 to 20 percentage point power advantage over expert visual inference at intermediate discontinuity levels.[11]

Finally, we apply the AK inference procedure from Armstrong and Kolesár (2018) and implemented in the R package `RDHonest`, which adapts Donoho (1994) and produces asymptotically valid and minimax optimal confidence intervals over the Taylor class of conditional expectation functions

$$\mathscr{F}_T(C;\chi_\pm) = \left\{ f_\pm : |f_\pm(x) - f'_\pm(0)x| \leqslant C_T|x|^2 \text{ for all } x \in \chi_\pm \right\}.$$

C_T is a researcher-supplied bound on the second derivatives on both sides of 0, which is an important tuning parameter and dictates the bias magnitude of a local linear estimator (hence, unlike CCT, AK's bias correction is nonrandom). When using the true second-derivative bounds, which requires omniscience about the DGPs, the AK inference procedure has a size of approximately 7%, and the power function is very close to that of the experts.[12] The package `RDHonest` also implements a rule-of-thumb curvature estimator as

Cattaneo, and Farrell (2020a) proposes inference-optimal RD bandwidth selectors. As seen in Figure B18, using these bandwidths reduces the size distortion relative to visual inference while retaining the power advantage.

[11] One may be concerned that the better size control of visual inference is a consequence of our experimental design where the majority (nine out of 11) of RD graphs feature a discontinuity. If subjects speculate that only about half the graphs feature a discontinuity, our size result is biased in favor of visual inference. As mentioned in Section 2.3.3 and Appendix B.3, we explicitly tell participants that "all, some or none of the 11 graphs you see in this survey may feature a discontinuity," and we present evidence in Appendix B.4.3 to further alleviate this concern.

[12] Armstrong and Kolesár (2020) propose analogous confidence intervals that maintain coverage and enjoy minimax optimality over a Hölder class of functions, which is determined by a global, as opposed to local,

a default bound. This rule of thumb is obtained via a piecewise quartic regression, which approximates our piecewise quintic CEFs very well. Unsurprisingly, inference with these estimated bounds performs very similarly to inference using the true bounds, with a very small drop in size and a modest decrease in power. To further gauge how AK performs as a function of the tuning parameter, we use the maximum of the true bounds across all 11 DGPs. These larger bounds result in a decrease in size of about four percentage points but a large drop in power, up to around 25 percentage points at intermediate magnitudes and 10 percentage points at the smallest and largest ones.

Inference procedures based on the IK and CCT estimators at the 5% level exhibit higher type I and lower type II error rates than visual inference. It is difficult to say whether these econometric inference procedures perform better or worse because there is no guidance on how to trade off type I and type II errors. We circumvent this problem by size-adjusting the inference procedures: we search for alternative critical t-values (or equivalently, critical p-values) such that the resulting type I error rate of the econometric inference procedure is equal to that of visual inference and then use those critical values to conduct inference. For IK, this critical t-value is 2.46, and it is 2.28 for CCT. We present the results for these size-adjusted inference procedures in the right panel of Figure 2.4. Despite the extent of their size differences relative to visual inference, both econometric procedures' powers only suffer by around 5-10 percentage points from this adjustment. When using these conservative critical values, the IK and CCT procedures retain their statistically significant power advantages over visual inference at moderate discontinuities while matching visual inference in size, as shown in Figure B19.

In addition to asking experts to classify graphs as having or not having a discontinu-

bound on the second derivative of the CEF. The corresponding power functions (not plotted) are similar when using the true global second-derivative bounds. Like Armstrong and Kolesár (2018) and Armstrong and Kolesár (2020), Imbens and Wager (2019) also adapt the idea of Donoho (1994). They propose an RD estimator through numerical optimization that is minimax mean-squared-error optimal over CEFs with a global second derivative bound. Because it performs similarly to Armstrong and Kolesár (2018) in simulations by Pei et al. (2018) and can be computationally demanding, we do not implement the Imbens and Wager (2019) procedure here.

ity, we ask them to estimate the magnitude of the discontinuity. We take several steps to make human and econometric point estimates comparable. Because we ask participants to round their estimates to the nearest hundredth, we round estimators with the same precision. We similarly replace econometric estimates with 0 when the test fails to reject the null hypothesis.

Figure B20 presents the root mean squared error (RMSE) of each estimator's point estimates by DGP, averaged over all discontinuity magnitudes. We do this by DGP to prevent scaling issues, as units for each DGP are different and results from one DGP are not directly comparable to those from another. Although there are a handful of DGPs where experts perform similarly to the econometric estimators, their point estimates generally have a greater RMSE than the estimators, and experts overall are worse at estimating magnitudes. We also decompose the human MSE gap relative to the piecewise quintic in Figure B20 into its bias and variance components (see Appendix B.1.3 for details about this calculation). The DGPs are divided approximately evenly into those where bias or variance accounts for most of the MSE. In summary, although humans perform quite well at identifying the existence of discontinuities, their ability to estimate their magnitudes is not as strong.

RK Visual vs Econometric Inference

Paralleling the exercise for RD, we compare the performance of our expert pool with that of econometric inference procedures. We first try two 5% tests on the kink parameter using a correctly specified OLS regression model (piecewise quintic with homoskedastic standard errors) and the default CCT `rdrobust` procedure, which is based on a bias-corrected local quadratic estimator for kinks (Imbens and Kalyanaraman, 2012 and Armstrong and Kolesár, 2018 do not provide RK extensions of their RD methods). We compare the power functions of visual inference with those for the piecewise quintic estimator in Figures 2.5 and B21. Despite being the correctly specified model, the piecewise quintic estimator has low power even at the largest kink level in our graphs. This is not surprising, as a kink of

0.9ξ corresponds to an asymptotic t-statistic of 0.9 for the piecewise quintic kink estimator, well below the 1.96 critical value used in the test. Results are similar for the default CCT `rdrobust` procedure, which has excellent size control but is highly conservative as seen in the same two figures.

The large divergence in the performances of visual and econometric RK inference raises the question of what the humans see that the two econometric procedures fail to capture. While this is a difficult problem, we can provide a partial answer by looking for econometric inference methods that better mimic visual inference. Beginning from the default CCT `rdrobust` procedure (bias-correction, local quadratic, triangular kernel, bandwidth regularization, and robust standard errors), we test alternative inference methods to find power functions that more closely resemble the experts'. We also plot the power functions of these additional inference procedures in Figure 2.5. Using a linear estimator instead of quadratic dramatically increases the power at larger kink magnitudes. Using conventional as opposed to robust standard errors and removing bandwidth regularization further increases power and brings the size closer to the experts'. The expert power function is sandwiched between those of conventional inferences based on local linear and local quadratic estimators with unregularized CCT bandwidths and uniform kernels.[13] We conclude that experts may be approximating a local linear or quadratic regression (or something in between) as a heuristic to classify graphs.

Although the econometric inference based on the correctly specified quintic regression lacks power on our simulated datasets, it is not the most powerful test we can conduct. Because of the normality of the error term in our DGP specification, we can implement a simple two-sided test, the uniformly most powerful unbiased test, for whether the kink is zero by imposing the true values of all parameters in the model except for the parameter of interest. The size of the test is approximately 8% on our sample and it has a power of one

[13]For the local linear and quadratic estimators, the average regularized CCT bandwidths are 0.23 and 0.33, respectively. The corresponding unregularized bandwidths are 0.41 and 0.56. A bandwidth of 1 indicates a global regression.

at all nonzero kink levels.[14] If we do not impose the true values of the other parameters (or nuisance parameters), the problem of computing the power function for the uniformly most powerful test becomes difficult (see recent progress by Elliott, Mäller, and Watson, 2015), and we leave it to future research.

To this point, we have presented the "marginal" power functions for visual and econometric inference procedures. In Appendix B.6, we extend this analysis to discuss the joint distribution of visual and econometric discontinuity/kink tests. Knowing the joint distribution i) informs us of their complementarity, i.e. whether visual and econometric inferences tend to be wrong on different data, and ii) is useful for combining inferences. We find that type II errors by experts are predictive of type II errors by various econometric inference methods, but the same is not true for type I errors. We conceptually illustrate how combining the binary visual inference decision with its econometric counterpart can refine overall inference.

2.4.5 Eyetracking Study Results

In our eyetracking study, we investigate whether certain eye patterns are predictive of correct discontinuity inference. Common practice in eyetracking is to divide the displayed image into "areas of interest" (AOIs) ex-ante and to measure the visual attention these AOIs receive. In a study by Byrne, Kalaji, and Niederdeppe (2018), for example, the AOIs include warning labels on images of cigarette packs, and the authors test whether the degree of visual attention therein predicts stated health risk beliefs and other outcomes ex-post. Our case is different: we are agnostic about the AOIs ex-ante, and instead search for eye patterns correlated with correct discontinuity inference.[15] If we identify robustly predic-

[14] We implement the same test for our RD DGPs as well, and its power function is practically the same.

[15] Because we show 11 different graphs, it is difficult to define consistent AOIs across the graphs. One might argue for using an optimal bandwidth to define AOIs. Because we are interested in the bandwidth for visual inference, we apply the `rdrobust` procedure to each of the binned datasets in the 11 graphs presented in the eyetracking study (the sample size of each dataset is the number of bins in the graph). The discontinuity classification based on the resulting 95% robust confidence intervals is only correct for six of the 11 graphs (55%), performing worse than our participants who classify 59% of graphs correctly on average.

tive eye patterns, we could then test in a follow-up experiment whether training participants based on these findings improves performance.

As explained in Appendix B.3.3, we choose the 11 graphs for this study by identifying those with the maximum difference in visual inference performance between non-experts and experts from the first three non-expert phases and two expert seminar samples. We begin by testing for differences in visual inference and attention between experts and non-experts in our eyetracking sample and present the results in Table B10. Consistent with results from our early expert sample, the experts in the eyetracking study perform better—they correctly classify the existence of a discontinuity in 68% of the graphs, higher than the 53% rate for the non-experts. On average, the six experts spend more than twice as long on each graph and fixate on more than twice as many locations.

We then transform the eyegaze pixel coordinate data into the units of the (x, y) variables of each graph and compute a set of summary measures from the eyetracking micro-data. These summary measures include the number of fixations, quantities that proxy for "visual bandwidth" (the horizontal range and horizontal standard deviation of the eyegaze coordinates), and the square root of the average squared deviation of the eyegaze coordinates from the true CEF scaled by σ (see Appendix B.7 for details). We refer to the latter measure as the "eyegaze RIMSE" (root integrated mean squared error). We also fit RDD models to the eyegaze data using the CCT inference procedure and record the conventional and bias-corrected local linear estimates. We then compute the absolute deviation of these estimates from the true discontinuity in the DGP, which we also scale by σ.

The only measures that are significantly different between the experts and non-experts at the 5% level are time spent on each graph and whether the local RD eyegaze estimates are computable (which depends on the number of distinct eyegazes)—fewer RD estimates are computable from non-experts' eyegaze data. Significant at the 10% level, experts cover 30% more horizontal range (1.40 out of a maximum of 2).

Next, we assess the extent to which eyegaze patterns predict visual inference success.

As mentioned above, the goal is to identify patterns that are causally linked to successful inference. To that end, our prediction exercise requires interpretability, and therefore we use the simple eyegaze summary measures in Table B10 as inputs to various prediction models. In addition, we include the 11 graph identifiers as explanatory variables—eyegaze patterns differ across graphs, and graphs vary in their difficulty, so omitting the graph identifiers may lead us to erroneously conclude that certain eyegaze patterns are important. Because of the large number of covariates relative to the number of graphs, we fit penalized logit, pruned tree, bagging, and random forest models to prevent overfitting and we choose tuning parameters via five-fold cross validation where applicable.

The most important predictive variables appear to be the graph identifiers. Some of the graphs we test are easy for participants to classify while others are difficult, and knowing which graph a participant looks at is the most potent predictor. We do not detect any ocular patterns that robustly predict successful visual inference based on our small eyetracking study (the full details of the prediction exercise are in Appendix B.7). We tentatively conclude that although experts outperform non-experts in this study, their eyegaze patterns are similar along most dimensions we consider, and the better performance is likely driven by the processing, as opposed to the acquisition, of visual signals.

2.5 Conclusion

This paper evaluates the performance of visual inference in RD and RK designs. Through a series of experiments and studies, we provide answers to questions in four related areas. First, how do graphical representation techniques affect visual inference? Second, do experts and non-experts differ in their visual inference performances, and are experts able to identify the most effective treatments for non-experts? Third, how does visual inference perform compared to common econometric inference procedures in RDD and RKD? And fourth, can we identify ocular patterns predictive of successful visual inference?

To answer the first question, we experimentally assess how five RD graphical parameters impact visual inference accuracy. We find that generating graphs with the Calonico, Cattaneo, and Titiunik (2015) IMSE (large) bin selector leads to higher type I error rates but lower type II error rates relative to their MV (small) bin selector. Similarly, imposing fit lines leads to increased power but may come at the expense of size distortion. Bin spacing, a vertical line at the policy threshold, and y-axis scaling have little effect. For RKD, larger bin widths again lead to higher type I error rates, though to a smaller extent than in RDD, but generally lower type II error rates. Among the two additional RK parameters we test, detrending may improve power without distorting size, while plotting derivatives leads to worse size and worse power.

For the second set of questions, we recruit experts to participate in a similar study. Experts perform similarly to non-experts in reading RD graphs, with marginally lower type II error rates at moderate discontinuity levels. Experts' RK visual inference is considerably more powerful. When predicting different graphical representation techniques' effects on non-expert visual inference, experts only partly anticipate the tradeoff between type I and type II errors that arises.

Third, visual inference from RD graphs generated using the method with the smallest size achieves a lower type I error rate than econometric inference at the 5% level based on the Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) estimators, though the two econometric inference procedures offer considerable power advantages. The performance of visual inference is very similar to that based on the procedure suggested in Armstrong and Kolesár (2018). For RKD, type I error rates are higher for visual inference, but only for DGPs with curvature that peaks near the threshold. Analyzing the joint distribution of visual and econometric tests reveals their complementarity: while they commit similar type II errors in RDD, there does not appear to be a strong association in their type I errors or in RKD. Systematically studying the settings in which visual inference outperforms econometric inference could be an important next step.

Finally, our eyetracking study does not reveal eyegaze patterns that robustly predict success in visual inference. Although the experts in this sample perform better than the non-experts, we cannot attribute this to their ocular movements, at least not in terms of summary measures such as their "visual bandwidth" or how well they trace out the underlying conditional expectation function.

A limitation of our study is the set of parameters we are able to experimentally test. As mentioned above, we do not impose fit lines with confidence intervals in our graphs, which researchers sometimes do, due to the difficulty in explaining it to non-expert participants. We also do not vary the size and color of the dots. However, these choices may impact inference due to their effect on visual attention and visual complexity as suggested by the literature on the psychological and neurological mechanisms underlying the processing of (visual) information (Hegarty, Canham, and Fabrikant, 2010; Kriz and Hegarty, 2007; Rosenholtz, Li, and Nakano, 2007; Wolfe and Horowitz, 2004). We leave these investigations to future work.

Another caveat for our results is that they are based on a specific set of DGPs. For example, while the bin spacing choice—equally spaced or quantile spaced—appears immaterial in our RD experiments, it could be important when the distribution of the running variable is farther from uniform than in our DGPs. On the other hand, the number of DGPs used in Monte Carlo simulations that lead to methodological recommendations is typically far lower than our 11, and those DGPs sometimes bear no semblance to real-world data. In addition, we test the validity of our simulated datasets by adapting the lineup protocol from Majumder, Hofmann, and Cook (2013) to assess the degree to which our DGPs approximate the original data.

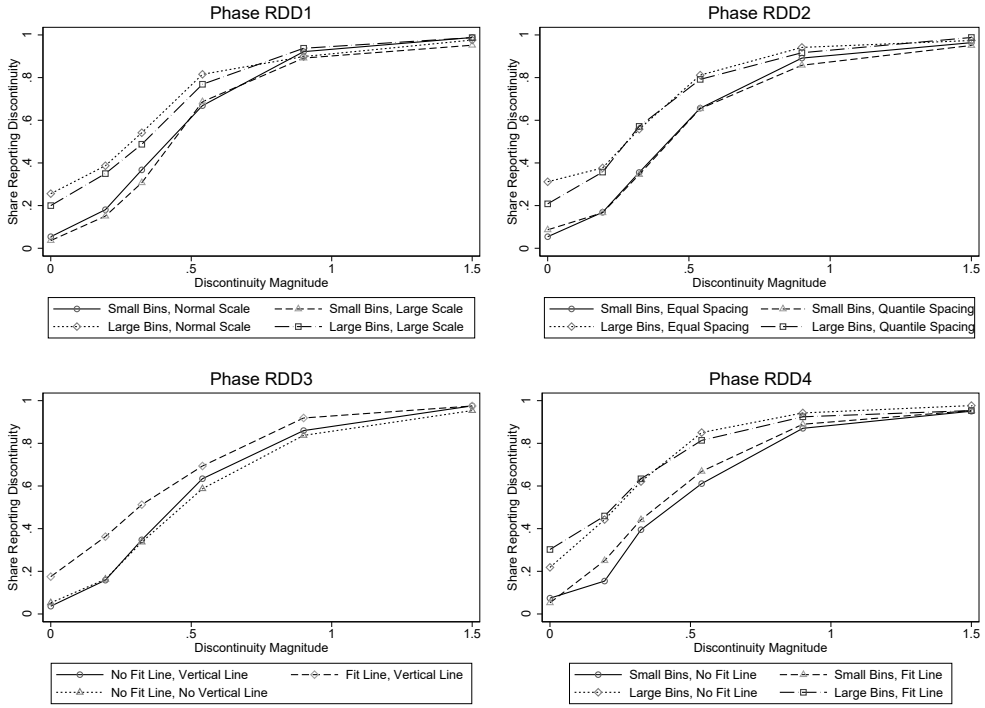
Our study answers the call by Leek and Peng (2015) to provide empirical evidence on best practices in data analysis, and we hope it will spur similar investigations. One related topic to study follows the recent work by Cattaneo et al. (2019), who, among other contributions, propose econometric tests of linearity and monotonicity based on binned

scatter plots, which are motivated by studies such as Chetty et al. (2011) and Chetty et al. (2014). One could assess the impact of the graphical parameters on reader perception and compare visual and econometric linearity/monotonicity tests. Another closely related topic is structural breaks in time series econometrics, in which graphs serve practically the same purpose as those in RDD/RKD. Within time series econometrics, studying visual inference for unit root/stationarity analysis may also be promising.[16] In an influential textbook, Stock and Watson (2011) conduct an augmented Dickey-Fuller (ADF) test for the presence of a unit root in U.S. inflation. Upon finding that the test rejects a unit root at the 10% level but not at the 5% level, Stock and Watson (2011) write "The ADF statistics paint a rather ambiguous picture...Clearly, inflation in [the figure] exhibits long-run swings, consistent with the stochastic trend model." In this case, Stock and Watson (2011) apply visual unit root inference when the test statistic is marginal, which raises the question: can we leverage our eyes to begin with?

[16]In recent work, Shen and Wirjanto (2019) propose a new framework for stationarity tests, which formalizes the intuition that a visual characteristic of stationary time series is the infinite recurrence of "simple events" asymptotically.

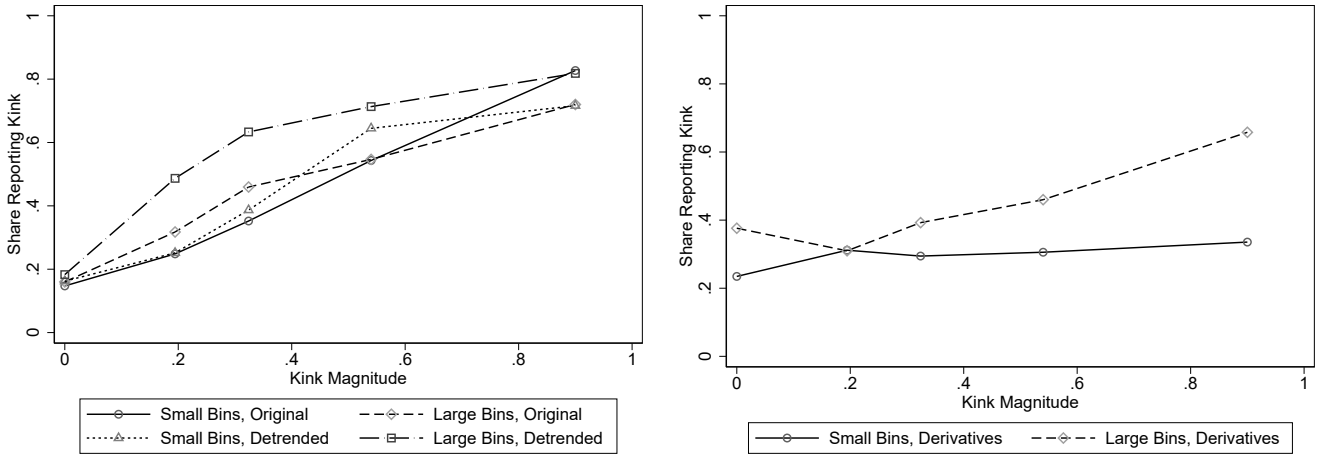
Figures

Figure 2.1: RDD Phases: Power Functions



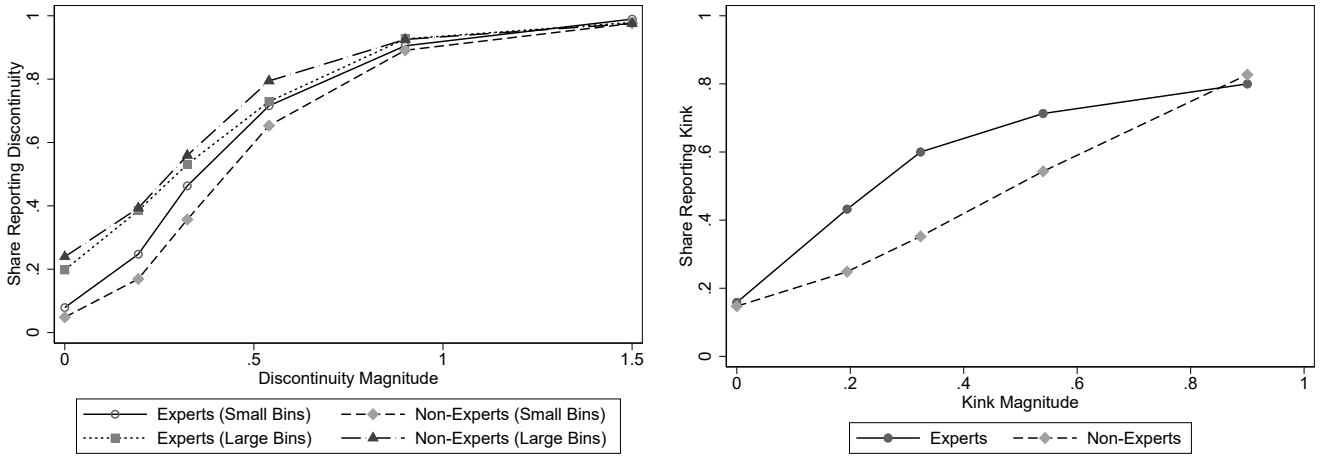
Notes: Plotted are empirical power functions from the four non-expert RDD experiments. The power functions are defined in Section 2.2. The discontinuity magnitude on the *x*-axis is specified as a multiple of the error standard deviation. The *y*-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure 2.2: RKD Phase: Power Functions



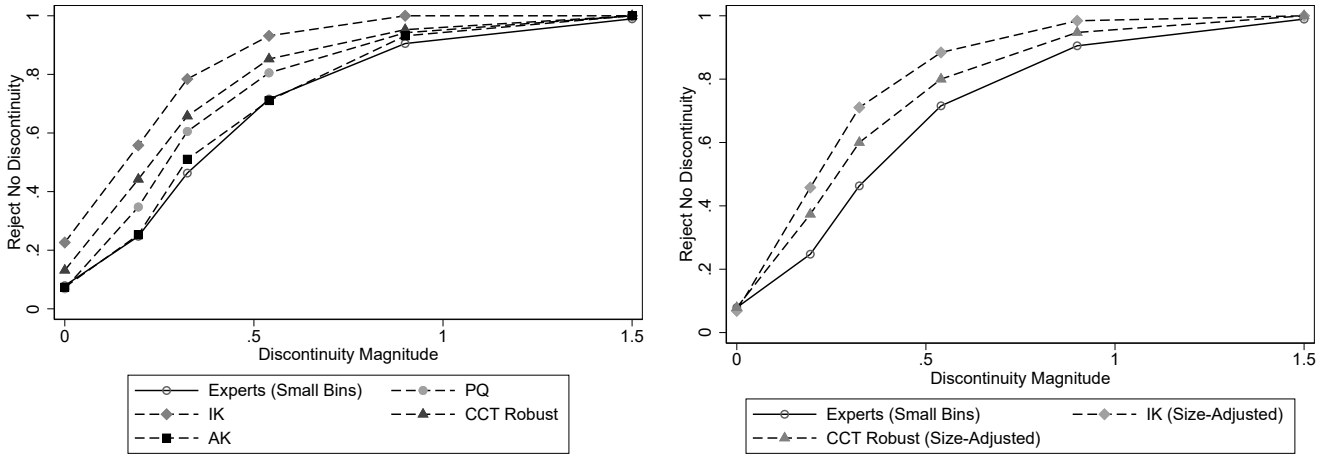
Notes: Plotted are empirical power functions from the RKD experiment. The power functions are defined in Section 2.2. The discontinuity magnitude on the x-axis is specified as the asymptotic t-statistics of the angle estimator between the left and right arms of the CEFs (see Section 2.3.2 and Appendix B.2 for details). The y-axis represents the share of respondents classifying a graph as having a kink at the policy threshold. The left figure compares the power functions for the four RK treatments that plot the outcome in levels. The right figure compares the power functions for the two RK treatments that plot the first derivatives of the outcome with respect to the running variable.

Figure 2.3: Expert vs Non-Expert Performance



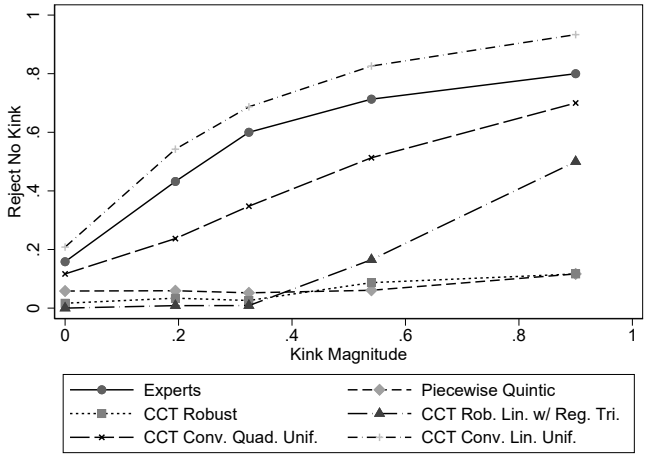
Notes: Plotted are the power functions for the experts and non-experts. The left and right figures compare their performances from reading RD and RK graphs, respectively.

Figure 2.4: RDD: Expert Visual vs Econometric Inference



Notes: PQ uses a correctly-specified regression model using piecewise quintics above and below the treatment threshold and assuming homoskedasticity. IK is based on a local linear estimator using the IK bandwidth. CCT Robust is the default RDD inference procedure from CCT's `rdrobust`, which uses the robust confidence interval based on a bias-corrected local quadratic estimator with triangular kernel and regularized (smaller) bandwidth. AK uses the `RDHonest` procedure with the true bound on each DGP's second derivative.

Figure 2.5: RKD: Expert Visual vs Econometric Inference



Notes: Piecewise Quintic is a correctly specified regression model using piecewise quintics above and below the treatment threshold and assuming homoskedasticity. CCT Robust is the default RKD inference procedure from CCT's `rdrobust`, which uses the robust confidence interval based on a bias-corrected local quadratic estimator with triangular kernel and regularized (smaller) bandwidth. CCT Rob. Lin. w/ Reg. Tri. is the same as CCT Robust, except it is based on a linear estimator. CCT Conv. Quad. Unif. is the inference procedure using the conventional confidence interval based on a uncorrected local quadratic estimator with uniform kernel and unregularized (larger) bandwidth. CCT Conv. Lin. Unif. is the same as CCT Conv. Quad. Unif., except it is based on a linear estimator.

Chapter 3

Migration and Local Minimum Wages[1]

3.1 Introduction

In recent years, local minimum wages, those for jurisdictions such as counties or municipalities below the state level, have expanded rapidly in the United States. Prior to 2013, there were only three such active policies in the country, while there are now approximately 60. They now cover tens of millions of Americans and typically set wage levels significantly higher than state or federal requirements, which have been comparatively static over the same period, with the latter being fixed at $7.25 per hour since 2009. While economists have explored the impacts of national and state minimum wages in a long, contentious literature, the effects of local minimum wages are much less understood beyond high-level outcomes such as earnings and employment. In a recent summary, Dube and Lindner (2021) write that "surprisingly little research has been devoted to some important aspects of [local] minimum wages," noting in particular the possibility of spatial changes or distortions.

This paper explores the broader impacts of local minimum wages by focusing on migration responses. When a local minimum wage is enacted in an area, are there changes in its migration rates? If so, are people more likely to move in or out, where are they coming

[1]I am grateful to Leah Boustan, David Lee, and Alex Mas for helpful comments.

from or going to, and are there impacts on untreated neighboring areas? Further, is there selection in these moves, where higher- or lower-earning individuals are more likely to enter or exit the area? Depending on the magnitudes of these responses, the distortions created by local minimum wages could be large and counteract some of the benefits of such a policy. The ability or efficiency of local minimum wages to address, for example, regional variation in costs of living, local monopsony power, or an undesirable income distribution may be threatened if geographic mobility allows people to arbitrage away gains, and the frictions for these moves are likely to be smaller than those for minimum wages set at the state level.

I address these questions using detailed migration data derived from US tax filings. These returns approximate the universe of US households and allow for analyses of (non)-migration, including in- and out-migration separately, as well as the adjusted gross income (AGI) of movers-in, movers-out, and non-movers. Leveraging variation in the timing of the passages of these local minimum wages through a matched difference-in-differences framework, I find several key results. First, in counties enacting local minimum wages, there is a significant, slightly delayed increase in out-migration rates. These movers-out appear negatively selected, as mean AGIs in this group decline after the minimum wage's passage. For foreign migration, however, there is a precipitous *decline* in out-migration after the passage of a local minimum wage, with no apparent changes in selection based on earnings. Finally, local minimum wages do not appear to result in significant spillovers into adjacent untreated counties, suggesting that they may not create large distortions that would limit their utility as a policy. The exact mechanism prompting the overall out-migration effects (e.g., disemployment or higher home prices) is unclear, but an analysis of typical home values indicates that some of the increase in wages of these policies may be passed through to the real estate market. These results collectively suggest major impacts of local minimum wages beyond their previously studied first-order effects on the earnings and employment of certain workers.

This paper contributes to several strands of the literature. First and foremost is the effort to understand the spatial effects of minimum wages. Monras (2019) is the most closely related work, developing a generalized Rosen-Roback spatial equilibrium model to find the key result that the direction of migratory responses of minimum wages depends on the elasticity of labor demand, with additional predictions based on how unemployment benefits are funded.[2] The paper presents corroborating empirical evidence based on state-level minimum wages. Pérez (2020), in a paper similar in question but different in execution, uses data from the Longitudinal Employer-Household Dynamics Program's Local Origin and Destination Employment Statistics on residences and commutes. Comparing nearby areas on opposite sides of borders of states with different minimum wages, he finds that minimum wage increases cause a drop in commuters to that area and a shift of jobs from the high-minimum wage area to the lower minimum wage area, while the migration results are less conclusive but "do not rule out emigration from higher minimum wage areas."

There is also a related sub-strand of this line of research devoted to identifying the effects of minimum wages specifically for immigrants, with mixed results. Using state-level variation from federal minimum wage increases, Giulietti (2014) argues that minimum wage increases were a "pull factor" for low-skilled (but not high-skilled) immigrants, as they increased wages without decreasing employment. Boffy-Ramirez (2013) also focuses on low-skilled immigrants, finding that the increase in wages without disemployment raised expected earnings and attracted low-skilled workers, but that this effect is concentrated among immigrants who have been in the US between two and four years. But Orrenius and Zavodny (2008) and Cadena (2014), both using data from the Current Population Survey (CPS) and variation in state-level minimum wages, find that low-skilled immigrants are less likely to live in areas with higher or less stagnant minimum wages, which they argue could be evidence of adverse labor market effects of the policies. And focusing instead

[2]The specific relevant predictions for local minimum wages are in Proposition 2, which states that "in general [. . .] [the region enacting the minimum wage] gains or loses population depending on the local labor demand elasticity. That is, there is a threshold value of the labor demand elasticity above which [the region] loses population when minimum wages increase."

on the transmission of minimum wage effects across regions, earlier studies by Castillo-Freeman and Freeman (1992) and Santiago (1993) show that migration from Puerto Rico to the United States increased after Puerto Rico's minimum wage increased in 1974.

Distinct from state and federal minimum wages due to their typical magnitude, hyper-locality, and political background, local minimum wages have also received specific attention from economists. Like the broader minimum wage literature, however, there is disagreement on the existence and magnitude of disemployment effects. Early analyses of the handful of local minimum wage early movers (pre-2013) such as Potter (2006), Dube, Naidu, and Reich (2007), and Schmitt and Rosnick (2011) support the idea that these policies increase earnings without material disemployment effects. Studies of the more recent wave of local minimum wages (the focus of this paper) are less congruous about the presence of disemployment, though negative estimates are still not large. Studies by Jardim et al. (2017) and Jardim et al. (2018) focusing on Seattle's 2015 minimum wage increase find significant drops in hours worked. In contrast, Allegretto et al. (2018) use coarser data from the Quarterly Census of Employment and Wages for six cities' minimum wage increases, finding that earnings increased without significant negative employment effects, a result also found in a separate analysis of the Seattle policy by some of the same authors (Reich, Allegretto, and Godoey, 2017). Dube and Lindner (2021) provide a history of local minimum wage studies as part of a broader summary of the policies.

Finally, this paper adds to our understanding of the empirical tools commonly used to evaluate the effects of minimum wages. Comparisons of geographically contiguous treated and untreated areas have formed the basis of numerous key results on the minimum wage in labor economics, but the possibility of spatial distortions from these policies may pose problems for such strategies. Confounders like population changes could then cause us to misinterpret measures such as employment rates or average earnings. Although a few studies, such as Allegretto et al. (2013), Neumark, Salas, and Wascher (2014), and Allegretto et al. (2018), have employed techniques like matching or synthetic controls that may

partially address this, the notion of spillovers to untreated areas has received insufficient attention in labor economics given this technique's ubiquity and the first-order nature of the bias. To my knowledge, a robustness check using interior state counties that finds no spillovers in Dube, Lester, and Reich (2010)'s analysis of minimum wages with a state border discontinuity design is the sole example in the literature, though Kuehn (2016) and McKinnish (2017) consider commuting spillovers and do find rises in out-of-state commuting after minimum wage increases. However, my results for untreated neighbors are largely consistent with null responses, with some caveats, supporting prior estimates based on geographic neighbors.

The sum of my results paints a complex picture of the effects of local minimum wages. On the positive front, local spatial distortions appear minimal, which would support the use of local minimum wages as part of the place-based policies toolkit to address matters such as regional variation in costs of living. Further, there are not large increases in migration into treated areas, and so the benefits of the policy are likely not being arbitraged away by people who did not live in these areas prior to the minimum wage's passage. However, there are downsides, with the primary concern being the overall out-migration from local minimum wages. While I cannot pin down a mechanism, two possible culprits are disemployment effects of minimum wages and pass-through to housing costs, the latter of which is directly supported by my empirical results, though these channels are neither mutually exclusive nor exhaustive. And as a practitioner's note, when interpreting estimates of the effects of minimum wages on employment, the extent to which a measure or empirical strategy may be affected by migration should be considered.

3.2 Data and Institutional Settings

3.2.1 Local Minimum Wages

The interested reader should refer to Dube and Lindner (2021) for a succinct overview of

the past and present of local minimum wages in the United States. Broadly speaking, these policies are set at the city or county level and affect all work conducted within the area's borders. By design, they set wages above the prevailing state or federal wage floor, and indeed often go significantly above them (Dube and Lindner, 2021 document that 22 of 42 city minimum wages in the US are at least $15). They appear in cities and counties both large and small, but are predominantly found in more left-leaning areas, having been banned or preempted in many states with Republican leadership.

Prior to 2013, there were three local minimum wages in the United States: Washington, DC, Santa Fe, and San Francisco (New Orleans had attempted to raise its minimum wage, but the state of Louisiana preempted the law). Since then, the policies have exploded in popularity, and as of this writing, there are approximately 60 sub-state minimum wages in the US. I focus my attention on this current wave of the policies. Due to limitations in the migration data's availability described below, I will focus on county minimum wages passed by the end of 2014 (there were no county minimum wages passed in 2015, making this a natural stopping point). The 2013 passages were in Bernalillo County, NM; Montgomery County, MD; and Prince George's County, MD. The 2014 cohort was Santa Fe County, NM; Washington, D.C.; and San Francisco County, CA (Washington, D.C. straddles the line between city and state but has a county FIPS code and complete data availability, so I include it in my analysis).

Table 3.1 presents some basic information for each of my in-sample minimum wages. The prevailing minimum wages prior to the new policies for the Maryland ($7.25) and New Mexico ($7.50) counties were lower than for Washington, D.C. ($8.25) and San Francisco ($11.05), with the latter two having previously passed local minimum wage ordinances. Compared to the pre-treatment minimum wages, the post-treatment target wages are on average 40.0% higher. The time between enactment and when the policies started ranges from several months to a almost year, allowing for possible anticipation effects that will be important to account for in empirical analyses.

3.2.2 IRS Statistics of Income

My primary data source is the Internal Revenue Service's (IRS) Statistics of Income (SOI) program. The primary feature of these data are county-to-county flows for tax filers whose reported county of address changes from one year to the next. In particular, the data contain both the origin and destination counties, as well as the number of filers who did not move outside of the county, allowing the computation of gross and net migration changes benchmarked to the county tax filer population. The data also include the total adjusted gross income for each of these three groups (movers-in, movers-out, and non-movers). These migration numbers are provided for both filers (approximating the number of households) and exemptions claimed (approximating the number of people); I focus on the filers totals, as the household is likely the level at which most migration decisions are made.

The IRS SOI data have many advantages, but also has some limitations. Per Molloy, Smith, and Wozniak (2011), the population of tax filers is not necessarily representative of nonfilers, though they note that the CPS states that 87% of household heads filed tax returns between 1992 and 2009. Two major groups that may not file taxes are the elderly (age 65 or older) and the poor earning under certain cutoffs that depend on age and marital status. While the elderly are not a key population of interest in this paper, as youth or working-age adult are generally the focus of economists analyzing minimum wages, poor individuals and families are more relevant. Although the cutoffs for needing to file taxes are low enough that a full-time worker covered under one of the local minimum wages considered would be unlikely not to be required to file, disemployment effects could result in more nonfilers. However, even when not required to file, there are still large incentives to do so to be eligible for certain tax credits. Molloy, Smith, and Wozniak (2011) also observe that in the CPS data, tax filers tend to migrate more frequently than nonfilers. This gap has stayed constant over time, so while it may result in higher estimates of migration changes than if using the ACS or CPS, the lack of differences over time limits more problematic biases

from trend changes. Finally, there were fundamental changes to the IRS data beginning in 2011 that render data from before and after that period incomparable, and so I limit my sample to 2011 through 2017, the final year currently available.

3.2.3 Additional Data

To obtain a panel of county labor market conditions, I combine data from two sources. First is the US Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW), which provides quarterly labor market statistics for every county in the United States. I use two data points from the QCEW, each taken from the first quarter of the year. First is total employment for each county, which combined with population data, offers an employment-to-population measure. Second, the QCEW breaks down employment totals within different North American Industry Classification System (NAICS) sectors, which are helpful in understanding the types of industries in which workers in different counties are likely to be employed. Both of these are key covariates likely to affect the impact of minimum wages.

Additionally, I use data from the Longitudinal Employer-Household Dynamics (LEHD) program's Quarterly Workforce Indicators (QWI). Like the QCEW, The QWI provides quarterly county-level employment statistics such as the number of workers and their earnings based on the LEHD's linked employer-employee data, which cover 95% of US private sector jobs. For every county-year, I collect the first quarter employment totals for all provided age bins (14-18 years, 19-21, 22-24, 25-34, 35-44, 45-54, 55-64, and 65-99). Age is strongly correlated with whether a minimum wage is binding for a given worker, making this another important factor in an area's impact from a minimum wage.

Finally, I supplement this panel with auxiliary data from the US Census Bureau. To obtain the denominator for the employment-to-population ratios, I use the Census' intercensal county population estimates. And to identify the geographic neighbors of counties with local minimum wages, I use the Bureau's county adjacency files. These data allow

for counties with borders over water to be considered adjacent, such as San Francisco and Alameda Counties in California.

3.2.4 Summary Statistics

Table 3.2 presents some key summary statistics for a weighted average of all US counties (column 1), treated counties that passed minimum wages by the end of 2014 (column 2), and their untreated neighbors (column 3). Summary statistics for treated and adjacent counties are based on the year prior to the passage of the minimum wage, while the US statistics use data from 2012.

Counties enacting minimum wages in the sample period are exceptional in certain ways. Their average weekly earnings, for example, are substantial higher than those of the average US county's, $1,200 to $700. And their mean population is also seven times that of the national average. Local labor markets appear stronger in treated counties, with significantly higher prime-age (25-54) employment-to-population rates, and the workforces also have higher prime-age worker shares.[3] Based on NAICS one-digit sectors, treated counties have much lower rates of unskilled blue-collar jobs such as agriculture and mining (sectors 1-2-3) and much more skilled white-collar work (sectors 5-6-9). Migration rates across treatment status, however, are quite similar, with the exception of annual across-state migration rates being higher for treated counties, five percent to two percent.

The untreated adjacent counties also differ from the national average, but only along certain dimensions and in a less extreme fashion. Mean populations are now 3.4 times of the US average, and average weekly earnings are about $300 higher. Although prime-age employment-to-population ratios and workforce shares are about the same for the US average and adjacent counties, the NAICS sector breakdown is much closer to that of the treated counties. As with the treated counties, migration rates are very similar to the na-

[3] When using these figures to compute employment-to-population ratios with the QWI data, values greater than one are possible because the QWI numerator is based on the place of work, while the Census denominator is based on residence. To avoid outlier values, I cap the ratios at one.

tional averages.

3.3 Empirical Strategy

3.3.1 Coarsened Exact Matching

This paper intends to estimate the causal effects of local minimum wages, not only on the areas that enacted such policies, but on their neighbors. To that end, conventional approaches such as difference-in-differences comparing treated areas and their neighbors are inappropriate, especially if there are spillovers. Additionally, the areas with local minimum wages tend to be exceptional in certain ways, for example, having higher pre-policy wages, different industry composition, or larger populations, and their neighbors may be poor counterfactuals.

To solve these problems, my preferred empirical strategy is based around coarsened exact matching (CEM) methods, described, for example, in Iacus, King, and Porro (2012). With CEM, a particular strength is that the control units may come from anywhere geographically, and I can ensure that treated and adjacent observations have separate, appropriately similar counterfactuals. Further, the large number of counties in the United States (more than 3000) and the availability of high quality annual county data makes matching a natural fit. Similar approaches have been taken in the minimum wage literature with synthetic controls methods, such as in Allegretto et al. (2013) and Allegretto et al. (2018), but I am not aware of any efforts to use these techniques to estimate policy spillovers to neighboring areas.

Using the county-year as my level of observation, I implement CEM based on the share of workers employed in three bins of North American Industry Classification System (NAICS) one-digit sectors, the employment-to-population ratio for prime-age adults, and the percentage of the workforce consisting of prime age workers. For the NAICS codes, I group sectors with leading digits 1-2-3, (agriculture/forestry/fishing/hunting, mining/quar-

rying/oil/gas, utilities, construction, and manufacturing), 4-7-8 (wholesale trade, retail trade, transportation/warehousing, arts/entertainment/recreation, accommodation/food services, and other services) and 5-6-9 (information, finance/insurance, real estate/rental/leasing, professional/scientific/technical services, management, administrative/support/waste management/remediation services, education, health care/social assistance, and public administration). Because this partition covers every industry, I omit the 4-7-8 category from matching. I exclude counties that are eventually treated or that contain municipal minimum wages from consideration for matching for the untreated adjacent counties.

These measures provide a sense of the industry composition, labor market strength, and ages of workers in the area, which may all affect the impacts and passage of local minimum wages. I conduct CEM using 10 percentage point bins for each of these variables and match using data from the year prior to the enactment of the minimum wage. For example, to find matched comparisons for Bernalillo County, New Mexico, whose minimum wage passed in 2013, I take all counties with NAICS 1-2-3 and 5-6-9 sectoral shares, prime age employment-population ratio, and prime age workforce share in the same 10% bins in 2012. I explicitly do not match on outcomes to provide a more transparent approach to the construction of a counterfactual.

3.3.2 Two-Way Fixed Effects and Callaway and Sant'Anna (2020) Difference-in-Differences

After identifying appropriate control units and their weights from the CEM algorithm, the natural empirical approach is difference-in-differences (DiD). In recent years, economists have used "two-way fixed effects" (TWFE) models when dealing with staggered event times such as this, where not all counties experience the event in the same period (year).

The standard TWFE model for my setting is as follows.

$$Y_{it} = \sum_{\tau=-2}^{3} \beta_\tau D_{i\tau} + \alpha_i + \gamma_t + \varepsilon_{it} \tag{3.1}$$

The unit of observation is the county-year, and the relevant event in $\tau = 0$ is the passage of the minimum wage, rather than its start date, to limit the possibility of anticipatory responses (in practice, the passage is either in the same year as the minimum wage's start or the year before). To avoid multicolinearity problems between event-time and year indicators, it is necessary to drop one event-time indicator, conventionally $\tau = -1$. With staggered event times, researchers "bin" together and do not report event years outside of the common support.

In common practice, under the assumption of parallel trends for an outcome between the units of interest and their matched controls, the above regression was used to estimate the causal effects of the local minimum wages. There are two main benefits to this approach. First, the inclusion of policy leads allows for a visual test of the pre-treatment parallel trends assumption. Second, the non-parametric dynamic estimates are more appropriate for a setting in which we are relatively agnostic about treatment effects and temporary or one-time effects are each perfectly plausible. Local minimum wages may, for example, result in a one-time migration change as marginal agents move in or out. Or they might result in more permanent changes, such as an increase in migration in as people age into the workforce and relocate to these areas with higher wages for inexperienced workers.

However, while still commonly used by practitioners, difference-in-differences estimates with staggered treatment times have received considerable attention by econometricians in recent years, highlighting issues surrounding bias and interpretability. A wave of papers[4] explores the inner workings of these types of estimators (and their event study analogs), with the key results being that even in fully dynamic models with pure control

[4]See, in particular, Borusyak and Jaravel (2017), Sun and Abraham (2020), De Chaisemartin and d'Haultfoeuille (2020), Callaway and Sant'Anna (2020), Goodman-Bacon (2020), Imai, Kim, and Wang (2020), and Athey and Imbens (2021).

units, the parameters identified are an unusual weighted average of pairwise comparisons without obvious interpretation and with problems of negative weights.

To address these issues, I employ the estimator described by Callaway and Sant'Anna (2020) (hereon referred to as "CS"), referring the reader to their work for full details. The key innovation of the paper builds around what the authors call the "group-time average treatment effects," or the treatment effects for each cohort at a given point in (event) time. In the potential outcomes framework, this average treatment for the treated (ATT) for cohort g (i.e., counties whose minimum wage passed in a certain year) t years after the event can be written as

$$ATT(g,t) = \mathbb{E}[Y_t(g) - Y_t(0)|G_g = 1] \tag{3.2}$$

for outcome Y and indicator variable G taking value 1 once an observation receives its absorbing treatment. Under straightforward assumptions (limited anticipation, conditional parallel trends, and an overlap condition), $ATT(g,t)$ can be identified in a manner that avoids the undesirable comparisons that make up the standard TWFE coefficients.

In addition to deriving unbiased estimation strategies for these quantities, the authors identify aggregation procedures to turn these group-time average treatment effects into more conventional quantities, such as the event study estimates I will present, as opposed to separately considering each cohort's treatment estimate in every event year. Specifically, I employ the scheme described in Equation 3.6 of Callaway and Sant'Anna (2020) for aggregated estimates using a "balanced" panel (i.e., the composition of the treated units does not change over event time), reproduced here with a simplified notation, as I compute aggregated ATTs only for identified, cohort-balanced years.

$$\theta(e) = \sum_{g=2013}^{2014} ATT(g, g+e)P(G=g) \tag{3.3}$$

This is simply a weighted average of the group-specific ATTs. The CS method thus provides an elegant, interpretable, drop-in replacement for TWFE difference-in-differences

practices that avoids many of the issues around estimating and aggregating treatment effects. Further, the authors' software implementation allows for simultaneous confidence bands as opposed to pointwise, which "asymptotically cover the *entire path* of the group-time average treatment effects with fixed probability and take into account the dependency across different group-time average treatment effect estimators."

3.4 Results

3.4.1 Match Quality

The benefit of CEM comes from the process' ability to identify an appropriate artificial control sample for the treated units. To evaluate the quality of these matches, I compare the different samples of the data before and after matching in Table 3.2. Columns 4 through 7 describe the treated counties matched by CEM, their matched controls, and the analogs for untreated adjacent counties. The matched counterfactual figures are weighted using the CEM weights used in the actual regressions. For experimental counties, the data used are from the year prior to the relevant event.

Note that the number of observations within the difference-in-differences samples changes from the full samples in columns 2 and 3 to their post-CEM analogues in columns 4 and 5. Because CEM is an exact matching algorithm, experimental counties with no partners in the same cells across all matching variables must be dropped from the analysis. This results in one treated county (out of six) and five adjacent counties (out of 21) being dropped for the difference-in-differences regressions. This process does not materially change the group means for the variables in Table 3.2.

Comparing column 4 with 5 for treated counties and 6 with 7 for the untreated adjacent counties allows us to see how well the matched counterfactuals align with the experimental samples. Overall, the quality of the matches appears good, especially considering that these counties enacting minimum wages are so exceptional. Average weekly earnings are

about 30% greater in treated counties than their matched comparisons, or approximately one standard deviation. For adjacent counties compared to their controls, they are slightly more than 20% greater. Unsurprisingly, treated counties are significantly larger in terms of population than their matched counterfactuals. There is no way around this, as these counties include some of the largest metro areas in the entire country. Populations are also larger in the adjacent experimental counties, though here the difference represents about two-thirds of a standard deviation. Moving beyond these level variables to the proportion and rate variables shows more positive results. Encouragingly, despite not matching on outcomes, migration rates are balanced very well by the process, including their second moments. And by design, the variables used for matching (NAICS sectoral shares, prime-age employment-to-population ratios, and the share of the workforce that is prime-age) are all well matched.

3.4.2 Migration Results

My primary results, those using the CS difference-in-differences estimators, are reported in Tables 3.3 through 3.8 and Figures 3.2 through 3.7. Results using conventional TWFE estimators are in Appendix C.2. The CS point estimates in pre-period t are relative to period $t-1$, and the post-period estimates in period t are relative to the final pre-treatment period $t=-1$.[5] By necessity, the first period results are omitted, but because I present the aggregated results for the balanced years only, all periods in the results have point estimates and standard errors. Note also that outcomes are reported as changes in rates for tax filers (analogous to households). A negative result for in-migration, for example, does not imply that the considered population is shrinking in the county, only that the rate of gross in-migration has dropped. Most of the following results will be constrained by the coarseness of the small number of effective observations, but I hope to revisit these results with finer administrative microdata in the future.

[5] I thank Olivia Healy for her slides detailing this and other fine details of implementing the CS estimator.

Figure 3.2 plots the difference-in-differences results for total in- and out-migration rates for treated and adjacent counties, as well as the mean log AGI of movers-in/out for each county type. For treated counties, the pre-period coefficients for migration out are flat and fairly close to zero, which is encouraging for the validity of the empirical strategy. There is no significant increase in migration out in the year of the minimum wage's passage, but starting in the first year after, there is a sustained increase of several tenths of a percentage point (the average county has an annual out-migration rate of seven percent; calling the average increase in the final three post-periods 0.5 percentage points, this would be a 7.1% increase in annual migration out). For in-migration, there is a positive pre-trend before an estimated increase in the in-migration rate of 0.4 percentage points (5.7%) in the year that the minimum wage is enacted. This one-year increased rate then erodes back to zero within the data availability horizon. Looking at the AGI of these movers, the average value is unchanged across the pre- and post-periods for movers-in. However, movers-out appear to have AGIs about 5% lower immediately after the minimum wage is enacted. In other words, the movers-out are more negatively selected: people with lower earnings are more likely to move out, though we cannot deduce the cause, such as disemployment from minimum wages, nor whether they also had low earnings before the minimum wage, from these results.

This increase in out-migration is appreciably large. A 0.5 percentage point increase in gross migration out sustained for (at least) three years would imply that almost one percent of the county's (tax-filing) households leave within that span. Given that the average household in the US has 2.5 members and the average treated county has a population of 700,000, if the average mover-out filer also represented a 2.5-person household, this would imply a population loss of around $700,000 \times 0.005 = 3,500$ individuals per year, or 10,500 in the three years following the minimum wage's passage, significant when considering the local employment and housing markets, as well as the tax base.

Turning our attention to the results for adjacent[6] counties in Figure 3.2, we see a different pattern of results. For adjacent out-migration, there may be a positive pre-trend, and post-treatment point estimates are all fairly near zero (in interpreting these and other adjacent pre-trends, recall that although there may certainly be pre-trends, these counties' status is based on a decision, the local minimum wage, made by their neighbor, not themselves). However, as with the treated counties, there does appear to be a drop in mean AGIs among movers-out after the policy of around 5% to 10%. For adjacent in-migration, point estimates do not show any apparent pattern of change, and there is no change in selection of movers-in on income for adjacent counties. Overall, these overall results for adjacent counties suggest a more null result than for treated counties on migration. The negative selection of movers-out may indicate some degree of spillovers from treated counties, but moreso on who moves rather than the number of moves.

Next, I look at subsets of migration to better understand the different flows resulting from local minimum wages. Although the results for domestic migration (left to the appendix) are mechanically similar to the total results due to their making up the vast majority of migration, the estimates for foreign migration in Figure 3.3 show stark differences. For foreign migration out, there is a large *drop* in out-migration (about 0.03 percentage points compared to an annual average of 0.01 percentage points) in the second year after the minimum wage's passage, while there no change in migration rates inward after the minimum wage's passage. These results are visually compelling but less precise than the main estimates due to the small number number of foreign moves, and the out-migration results, for example, likely reflect hundreds of moves per year rather than thousands. If there is any change in selection, based on the AGI results, both movers-in and movers-out appear to be negatively selected, but the magnitudes of any such changes are not enough to put significant weight on those signs. Other than a possible increase in migration rates right after passage for adjacent counties, there are no apparent spillovers or changes in selection.

[6]I use the term "adjacent" as a shorthand for "untreated and adjacent;" neighboring treated counties are excluded.

Next, I examine within-state and across-state migration in Figures 3.4 and 3.5. The within-state rates should be of particular concern to the states with treated counties, as they are the next level of government responsible for addressing externalities and spillovers that may be created by local minimum wages, such as unemployment benefits from job loss or the fiscal ramifications of a shrinking tax base due to out-migration. However, any within-state migration impacts appear modest at most. For treated counties, the results for migration out do not show a sustained increase, while any mild increase in migration in immediately after the policy enactment is suspect due to the positive pre-trends and a decline two and three years after. For adjacent county migration, treatment effects are null. Looking at migration involving different states, there is one striking result for treated counties: an increase in migration out of the state, rising throughout the post period until a point estimate of 0.5 percentage points in the final estimated period. Interestingly, however, we do not see a decline in the AGI of movers-out of this group.

Finally, to look at the most local distortions, Figure 3.6 presents results for migration inflows from and outflows to neighboring counties. Despite the previous results of increases in gross migration, these estimates suggest that the observed migration responses are not local, with no detectable changes in migration rates in or out for treated or adjacent counties, nor differences in selection based on the AGIs of those who do move (the point estimates in the post period for out-migration for treated counties are all marginally higher than in the pre-period, but the magnitudes of these coefficients are so small that it is premature to attach significance to them).

3.5 Discussion

Having established that local minimum wages impact migration, the next steps are to understand how and why they do. Note that despite the overall migration changes, there is minimal movement between neighboring counties. While good in the sense that the poli-

cies avoid distortionary "bunching" from small moves across local borders, this means that people are leaving the area entirely (and indeed, state, based on the previous results), which *does* have implications for state budgets, programs, etc. that matter for these counties.

Broadly, the overall results are consistent with Monras (2019)'s theoretical predictions and empirical results, the latter being based on state-level minimum wage changes. I observe an increase in out-migration from treated areas and that these movers-out report lower earnings in their tax filings, possibly because they are lower-skilled workers (I will discuss the mechanisms behind this result in more detail momentarily, as there are multiple possible causes). In the setting described by Monras (2019), this implies that in treated counties, employment decreases by more than wages increase, making the area less attractive in the Rosen-Roback sense. This does not necessarily imply that the local minimum wages are bad or inefficient, but only there are costs and benefits that policymakers must weigh. However, it *is* consistent with minimum wages having material disemployment effects, and it also implies that the labor demand elasticity is sufficiently high for disemployment effects to make the area less desirable despite the higher wages.

The lack of an immediate response to local minimum wages in some of these migration outcomes may help elucidate mechanisms. Meer and West (2016) examine the effects of minimum wages on the labor market and find drops in employment that grow until and plateau at approximately the three-year mark. This timeline matches the increase in out-migration that I observe and is evidence consistent with a delayed propagation of disemployment effects through the labor market. The null/noisy inward migration results do not provide much support either way, however; we might expect an immediate increase in migration inward if wages are rising before disemployment were felt, but if workers wait for in-hand jobs before moving to an area, that may not happen.

The results for immigrants add to another area of the minimum wage literature without total consensus: whether minimum wages attract immigrants. As mentioned in the earlier literature review, papers such as Orrenius and Zavodny (2008), Cadena (2014), and Giuli-

etti (2014) disagree on whether the policies serve as push or pull factors for immigrants. I find that while local minimum wages may not be a pull factor for international migrants in the sense of an increase in migration inflows, they do make these areas less likely to lose population to foreign countries. My results neither affirm nor refute previous findings directly, as I am examining the totality of foreign migration rather than the migration of foreign-born individuals. Responses may also vary with, for example, whether an individual is a documented immigrant.[7] But these international questions are fairly novel, and further investigation, such as with Census Bureau microdata, is warranted.

The finding of increased and negatively selected out-migration may hold important implications for the study of minimum wage impacts more broadly. Economists often focus on metrics, such as earnings or employment rates, that are conditional on employment.[8] Out-migration may mask disemployment effects of minimum wages on these measures. However, disemployment is not the only possible channel that may cause adverse outcomes. Pass-through to the housing market, for example, could also push marginal residents out of these areas. Indeed, the results from Figure 3.7 point to an increase in typical home values as measured by the Zillow Home Value Index of about 5%, which could not only force some individuals out of the treated counties, but also diminish the benefits of the policies for those who stay. Importantly, real estate prices affect everyone who lives in the area, not just workers at the margin for minimum wages. This may induce out-migration from people earning well above minimum wage. It also would make local minimum wages less effective as a redistributive tool, by allowing real estate holders to capture some of the increased labor earnings, and as a method to combat high local costs of living. For the problem of costs of living, as a back-of-the-envelope calculation, the mean percentage change in minimum wages for in-sample counties is 40.0%, which compared to the increase in

[7]Orrenius and Zavodny (2008) and Cadena (2014), for example, use data that include undocumented immigrants. Although it is estimated that 50 to 75 percent of undocumented immigrants pay local, state, and federal taxes (Merrell, 2007), undocumented filers are a self-selected group that may categorically different from, for example, a general population estimate from the Current Population Survey.

[8]There are, of course, exceptions, such as Cengiz et al. (2019), which focuses on total employment.

home values of about five percent in the years after the minimum wage's passage, implies an elasticity of 0.125. In interpreting this quantity, consider that Cengiz et al. (2019) show that minimum wages have "ripple" effects higher on the wage distribution that would cause minimum wages to increase earnings for a larger share of the population than just those directly impacted by the policies; they estimate that 40% of the total increase in wages is due to ripple effects. Although these numbers may be plausible, note that the confidence intervals on the post-period home value coefficients cannot rule out null passthrough effects. And the passage of local minimum wages may coincide with other policies that affect home values or even serve as a local amenity through warm glow effects. In any case, it is clear that minimum wages may have implications beyond first-order labor market effects, and understanding these impacts and their mechanisms is crucial for optimal policy design.

For practitioners, the data and policy variation in local minimum wages result in a natural setting in which to examine effects not just on treated counties, but also their untreated neighbors. Economists have for decades estimated the effects of minimum wages using geographic borders between treated and untreated areas, with the key identifying assumption being a lack of spillovers on the untreated regions. The mostly null impacts on untreated adjacent counties support prior estimates based on this strategy. However, researchers should think critically about such strategies. There may be different impacts for larger or smaller economic hubs or labor market centers compared to their suburbs, and these relationships could affect results. The literature on spatial minimum wages is younger than the broader literature on minimum wages, and future work should continue to explore how spatial impacts may bias our understanding of the labor market effects of the policies.

3.6 Conclusion

Using evidence from county minimum wages with coarsened-exact matching and new difference-in-differences estimators, I have shown that local minimum wages result in

changes in migration patterns. There is an increase in out-migration from counties after they enact minimum wages, and the people who move out after the policies are likely to have report lower earnings. There is also a *decline* in migration to foreign countries after the policy. For in-migration and migration to and from untreated adjacent neighboring counties, results are largely null. These results add to the complexity of our understanding of the labor market and spatial implications of minimum wages. While the mechanisms behind these results, such as disemployment or housing passthrough, are not entirely clear, the mixed signage of the above estimates may suggest that (local) minimum wages serve as "push" and "pull" factors for different groups. Overall, these findings provide more specific evidence of the tradeoffs of local minimum wages.

The policy implications of this research are not so much to discourage local minimum wages but to better understand their costs to weigh them against their benefits, which have been previously estimated. That local minimum wages induce migration out of treated areas and are accompanied by increases in home values needs to be weighed against the benefits of higher wages and prior findings of increased total labor earnings. These effects have implications beyond just the labor market effects of minimum wages, and states should be cognizant of changes in the labor force and tax base as a result of these policies. Encouragingly, however, there does not seem to be significant local bunching at treated/untreated borders, at least not based on migration results, which may have created distortions and arbitraged away many of the benefits of the minimum wages.

Future research should explore these migratory responses in more detail to better understand the mechanisms driving them. In future work, I plan to use administrative microdata to revisit these topics. With finer data, such as detailed information on moves, demographics, and earnings histories, one could estimate more precise impacts, including for city minimum wages, and make clearer the roles of disemployment effects or real estate prices or other channels. Moreover, the ability to follow such individuals before and after a move would more complete a more complete understanding of earnings or welfare effects. The

results of such an analysis would be invaluable for labor, public, and urban economists seeking to understand the full impacts of minimum wages as well as for policymakers determining whether they should consider a local minimum wage to improve outcomes for their constituents.

Table 3.1: Descriptive Statistics for In-Sample County Minimum Wages

County	MW Passed	MW Started	MW Before	MW After
Bernalillo, NM	April 2013	July 2013	$7.50	$8.50
Montgomery, MD	November 2013	October 2014	$7.25	$11.50
Prince George's, MD	December 2013	October 2014	$7.25	$11.50
Santa Fe, NM	January 2014	July 2014	$7.50	$10.66
Washington, D.C.	January 2014	July 2014	$8.25	$11.50
San Francisco, CA	November 2014	May 2015	$11.05	$14.00

Notes: Table reports descriptive statistics for county minimum wages in the analysis sample. Post-treatment minimum wages are the maximum wage targeted in the initial passage of the relevant policy; some areas later amended theirs to target higher wages.

Table 3.2: Summary Statistics for Treated and Adjacent Counties and Matched Control Counties

	(1) US (2012)	(2) Treat	(3) Adj	(4) CEM Treat	(5) CEM Treat Cont	(6) CEM Adj	(7) CEM Adj Cont
Avg Wkly Earnings ($1000s)	0.7 (0.2)	1.2 (0.4)	1.0 (0.4)	1.1 (0.4)	0.8 (0.3)	0.9 (0.3)	0.7 (0.2)
Population (1000s)	100.7 (322.3)	699.8 (275.2)	342.1 (425.9)	709.6 (300.5)	492.7 (618.1)	308.0 (422.9)	181.2 (288.9)
Prime-Age Emp/Pop Ratio	0.55 (0.19)	0.76 (0.19)	0.56 (0.21)	0.71 (0.17)	0.66 (0.16)	0.49 (0.17)	0.47 (0.17)
Age 25-54 Workforce Share	0.64 (0.03)	0.68 (0.03)	0.67 (0.03)	0.68 (0.03)	0.67 (0.03)	0.66 (0.02)	0.65 (0.03)
NAICS 1-2-3 Share	0.25 (0.15)	0.08 (0.03)	0.12 (0.05)	0.09 (0.03)	0.10 (0.05)	0.14 (0.04)	0.13 (0.05)
NAICS 4-7-8 Share	0.38 (0.12)	0.33 (0.06)	0.36 (0.09)	0.35 (0.05)	0.32 (0.05)	0.39 (0.07)	0.40 (0.09)
NAICS 5-6-9 Share	0.37 (0.15)	0.60 (0.08)	0.52 (0.13)	0.57 (0.07)	0.57 (0.07)	0.47 (0.08)	0.46 (0.10)
Total Out-Migr. Rate	0.07 (0.02)	0.08 (0.01)	0.08 (0.03)	0.07 (0.01)	0.08 (0.02)	0.07 (0.01)	0.07 (0.02)
Total In-Migr. Rate	0.06 (0.02)	0.07 (0.01)	0.08 (0.03)	0.07 (0.01)	0.07 (0.02)	0.07 (0.02)	0.07 (0.03)
Domestic Out-Migr. Rate	0.07 (0.02)	0.07 (0.01)	0.08 (0.03)	0.07 (0.01)	0.08 (0.02)	0.07 (0.01)	0.07 (0.02)
Domestic In-Migr. Rate	0.06 (0.02)	0.07 (0.01)	0.08 (0.03)	0.07 (0.01)	0.07 (0.02)	0.07 (0.02)	0.07 (0.03)
Foreign Out-Migr. Rate	0.001 (0.002)	0.001 (0.001)	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)	0.002 (0.001)	0.001 (0.002)
Foreign In-Migr. Rate	0.002 (0.002)	0.001 (0.000)	0.002 (0.001)	0.001 (0.000)	0.001 (0.002)	0.001 (0.001)	0.002 (0.003)
Same-State Out-Migr. Rate	0.04 (0.02)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.04 (0.02)	0.04 (0.01)	0.04 (0.02)
Same-State In-Migr. Rate	0.04 (0.02)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)
Diff.-State Out-Migr. Rate	0.02 (0.02)	0.05 (0.02)	0.04 (0.02)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.02)
Diff.-State In-Migr. Rate	0.02 (0.02)	0.05 (0.02)	0.03 (0.02)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.02)
Neighbors Out-Migr. Rate	0.03 (0.01)	0.03 (0.01)	0.04 (0.02)	0.03 (0.01)	0.03 (0.02)	0.03 (0.01)	0.03 (0.01)
Neighbors In-Migr. Rate	0.03 (0.01)	0.03 (0.01)	0.04 (0.02)	0.03 (0.01)	0.02 (0.02)	0.03 (0.01)	0.03 (0.01)
n	3141	6	21	5	48	16	322

Notes: Table reports the mean and (standard deviation) of key summary statistics for seven groups: (1) all counties in the United States in 2012; (2) treated counties that enacted a local minimum wage by the end of 2014; (3) counties adjacent to those that enacted a local minimum wage by the end of 2014; (4) treated counties that enacted a local minimum wage by the end of 2014 with matched control counties via coarsened exact matching; (5) counties adjacent to those that enacted a local minimum wage by the end of 2014 with matched control counties via coarsened exact matching; (6) control counties matched to those that enacted a local minimum wage by the end of 2014 via coarsened exact matching; (7) control counties matched to those adjacent to counties that enacted a local minimum wage by the end of 2014 via coarsened exact matching. Members of the control columns are weightened according to the weights assigned by the coarsened exact matching process. All statistics are taken from the year prior to the enactment of a minimum wage (in the case of matched controls, this is the year prior to the enactment of the matched unit's local minimum wage).

Table 3.3: Calloway and Sant'Anna Difference-in-Differences Estimates, Total Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	0.002	-0.004	0.070	0.090	0.000	0.003	0.035	-0.021
	(0.003)	(0.002)	(0.109)	(0.125)	(0.003)	(0.002)	(0.075)	(0.067)
t=-1	0.002	0.000	0.011	0.040	0.004	0.001	0.017	-0.014
	(0.001)	(0.001)	(0.034)	(0.042)	(0.001)	(0.002)	(0.046)	(0.024)
t=0	0.002	0.004	-0.050	0.002	0.001	0.004	-0.030	-0.023
	(0.002)	(0.002)	(0.042)	(0.040)	(0.001)	(0.001)	(0.062)	(0.049)
t=1	0.006	0.003	-0.049	0.056	-0.002	0.001	-0.105	0.039
	(0.003)	(0.003)	(0.037)	(0.098)	(0.002)	(0.002)	(0.046)	(0.036)
t=2	0.004	0.001	-0.032	-0.052	-0.003	-0.001	-0.051	0.008
	(0.003)	(0.002)	(0.034)	(0.032)	(0.004)	(0.002)	(0.047)	(0.022)
t=3	0.005	-0.001	-0.002	-0.025	0.000	0.005	-0.039	0.003
	(0.003)	(0.002)	(0.046)	(0.028)	(0.004)	(0.003)	(0.043)	(0.031)
Clusters	53	53	53	52	338	338	337	337

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

Table 3.4: Calloway and Sant'Anna Difference-in-Differences Estimates, Foreign Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	0.0001	0.0000	-0.105	0.082	0.0001	0.000	0.376	0.131
	(0.0002)	(0.0001)	(0.187)	(0.083)	(0.0001)	(0.000)	(0.420)	(0.195)
t=-1	0.0001	0.0002	-0.039	0.046	-0.0001	0.000	-0.246	0.047
	(0.0002)	(0.0001)	(0.188)	(0.310)	(0.0001)	(0.000)	(0.137)	(0.090)
t=0	0.0001	0.0001	-0.155	-0.197	0.0004	0.001	0.203	-0.058
	(0.0001)	(0.0001)	(0.163)	(0.235)	(0.0006)	(0.000)	(0.154)	(0.190)
t=1	0.0001	0.0001	-0.103	-0.263	0.0003	0.000	0.168	-0.022
	(0.0002)	(0.0001)	(0.146)	(0.161)	(0.0005)	(0.000)	(0.324)	(0.165)
t=2	-0.0004	-0.0001	-0.196	-0.272	-0.0005	0.000	-0.115	-0.288
	(0.0004)	(0.0002)	(0.053)	(0.224)	(0.0002)	(0.000)	(0.093)	(0.123)
t=3	-0.0003	0.0000	0.350	0.119	-0.0001	0.000	0.176	0.033
	(0.0003)	(0.0003)	(0.166)	(0.525)	(0.0003)	(0.000)	(0.175)	(0.160)
Clusters	24	24	24	24	81	86	80	85

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

Table 3.5: Calloway and Sant'Anna Difference-in-Differences Estimates, Same State Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	0.003	-0.002	0.056	0.177	-0.001	0.001	0.001	-0.107
	(0.002)	(0.001)	(0.063)	(0.149)	(0.003)	(0.001)	(0.067)	(0.100)
t=-1	0.001	0.000	-0.034	0.006	0.001	0.002	-0.037	0.020
	(0.001)	(0.001)	(0.026)	(0.099)	(0.001)	(0.001)	(0.025)	(0.035)
t=0	0.001	0.001	-0.043	0.034	0.001	0.001	-0.006	-0.025
	(0.002)	(0.001)	(0.024)	(0.022)	(0.001)	(0.001)	(0.018)	(0.051)
t=1	0.004	0.002	-0.030	0.088	0.000	-0.001	-0.027	0.038
	(0.002)	(0.002)	(0.026)	(0.156)	(0.001)	(0.001)	(0.027)	(0.043)
t=2	0.001	-0.001	0.011	-0.075	-0.001	-0.001	0.006	0.006
	(0.002)	(0.001)	(0.033)	(0.061)	(0.002)	(0.001)	(0.024)	(0.025)
t=3	0.000	-0.003	0.015	-0.011	-0.001	0.001	0.007	-0.007
	(0.002)	(0.001)	(0.041)	(0.058)	(0.002)	(0.002)	(0.027)	(0.036)
Clusters	53	52	53	51	338	337	337	335

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

Table 3.6: Calloway and Sant'Anna Difference-in-Differences Estimates, Different State Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	-0.001	-0.003	0.030	-0.039	0.001	0.002	0.032	0.060
	(0.001)	(0.002)	(0.078)	(0.106)	(0.000)	(0.002)	(0.091)	(0.082)
t=-1	0.001	0.000	-0.006	0.014	0.002	0.000	0.053	-0.049
	(0.001)	(0.001)	(0.061)	(0.033)	(0.000)	(0.001)	(0.075)	(0.056)
t=0	0.001	0.002	0.040	-0.016	0.001	0.003	-0.014	0.003
	(0.001)	(0.002)	(0.074)	(0.105)	(0.001)	(0.001)	(0.146)	(0.042)
t=1	0.002	0.000	-0.006	-0.004	-0.001	0.000	-0.124	0.025
	(0.002)	(0.001)	(0.037)	(0.037)	(0.001)	(0.001)	(0.072)	(0.039)
t=2	0.004	0.001	-0.035	-0.010	0.001	-0.002	-0.072	-0.010
	(0.002)	(0.002)	(0.068)	(0.050)	(0.001)	(0.001)	(0.071)	(0.040)
t=3	0.005	0.002	0.004	0.003	0.002	0.002	-0.025	0.015
	(0.002)	(0.002)	(0.056)	(0.043)	(0.002)	(0.002)	(0.071)	(0.044)
Clusters	53	52	52	52	328	327	321	325

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

Table 3.7: Calloway and Sant'Anna Difference-in-Differences Estimates, Neighboring Counties Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	0.000	-0.002	0.034	-0.005	-0.002	0.001	-0.050	-0.205
	(0.002)	(0.001)	(0.120)	(0.106)	(0.003)	(0.001)	(0.110)	(0.138)
t=-1	0.000	0.000	-0.067	-0.053	0.001	0.003	-0.042	0.009
	(0.001)	(0.001)	(0.033)	(0.033)	(0.001)	(0.001)	(0.030)	(0.034)
t=0	0.001	0.002	-0.067	0.003	0.001	0.001	0.017	-0.028
	(0.001)	(0.001)	(0.018)	(0.038)	(0.001)	(0.001)	(0.017)	(0.035)
t=1	0.002	0.000	-0.023	0.026	-0.001	-0.001	-0.008	-0.034
	(0.001)	(0.002)	(0.030)	(0.065)	(0.001)	(0.001)	(0.035)	(0.025)
t=2	0.001	-0.002	-0.014	-0.015	-0.001	-0.001	-0.004	0.006
	(0.001)	(0.001)	(0.041)	(0.028)	(0.001)	(0.001)	(0.018)	(0.025)
t=3	0.001	0.000	0.016	0.056	0.000	0.003	0.023	0.013
	(0.002)	(0.001)	(0.055)	(0.047)	(0.002)	(0.002)	(0.019)	(0.022)
Clusters	47	47	47	47	305	304	304	304

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

Table 3.8: Additional Calloway and Sant'Anna Difference-in-Differences Estimates

	Treated Counties			Adjacent Counties		
	Log Population	Non-Mover Log AGI	Log Home Value	Log Population	Non-Mover Log AGI	Log Home Value
t=-2	0.044	0.033	-0.004	-0.004	0.026	-0.021
	(0.030)	(0.044)	(0.006)	(0.008)	(0.019)	(0.019)
t=-1	0.028	-0.052	0.025	0.002	-0.022	0.010
	(0.018)	(0.024)	(0.031)	(0.004)	(0.009)	(0.014)
t=0	0.016	0.007	0.032	0.001	-0.005	0.018
	(0.004)	(0.018)	(0.017)	(0.005)	(0.008)	(0.023)
t=1	0.030	-0.021	0.054	0.006	-0.002	0.013
	(0.010)	(0.030)	(0.037)	(0.008)	(0.012)	(0.032)
t=2	0.032	0.010	0.063	0.015	0.016	0.005
	(0.014)	(0.033)	(0.060)	(0.010)	(0.015)	(0.039)
t=3	0.033	0.027	0.035	0.020	0.027	-0.021
	(0.019)	(0.036)	(0.057)	(0.012)	(0.018)	(0.041)
Clusters	53	53	45	338	338	287

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Standard errors are clustered at the county level. Home values come from the Zillow Home Value Index.

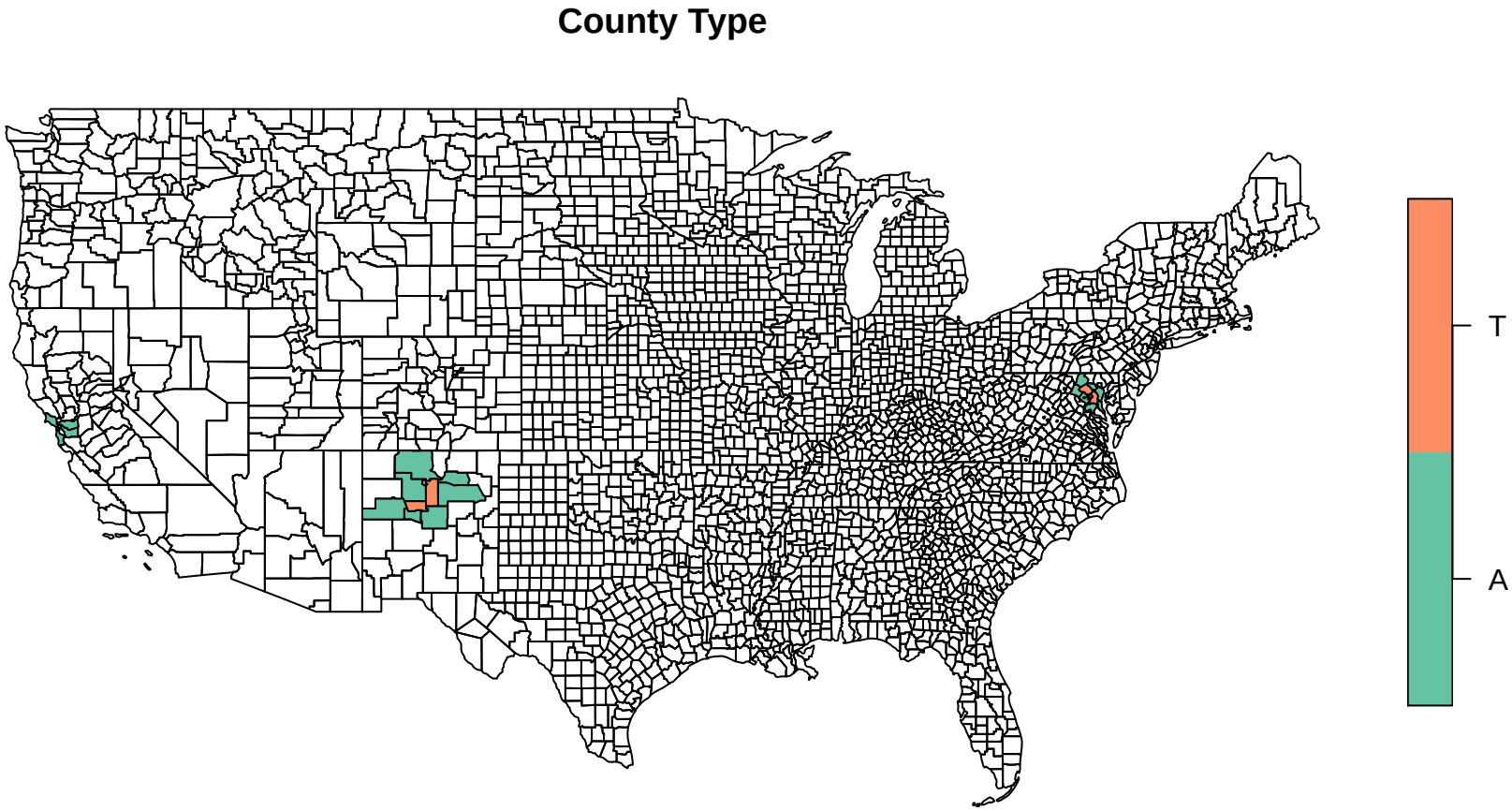


Figure 3.1: Maps of Treated and Neighboring Counties

Notes: "T" indicates a treated county that enacted a local minimum wage in 2013 or 2014, while "A" counties are their untreated adjacent neighbors.

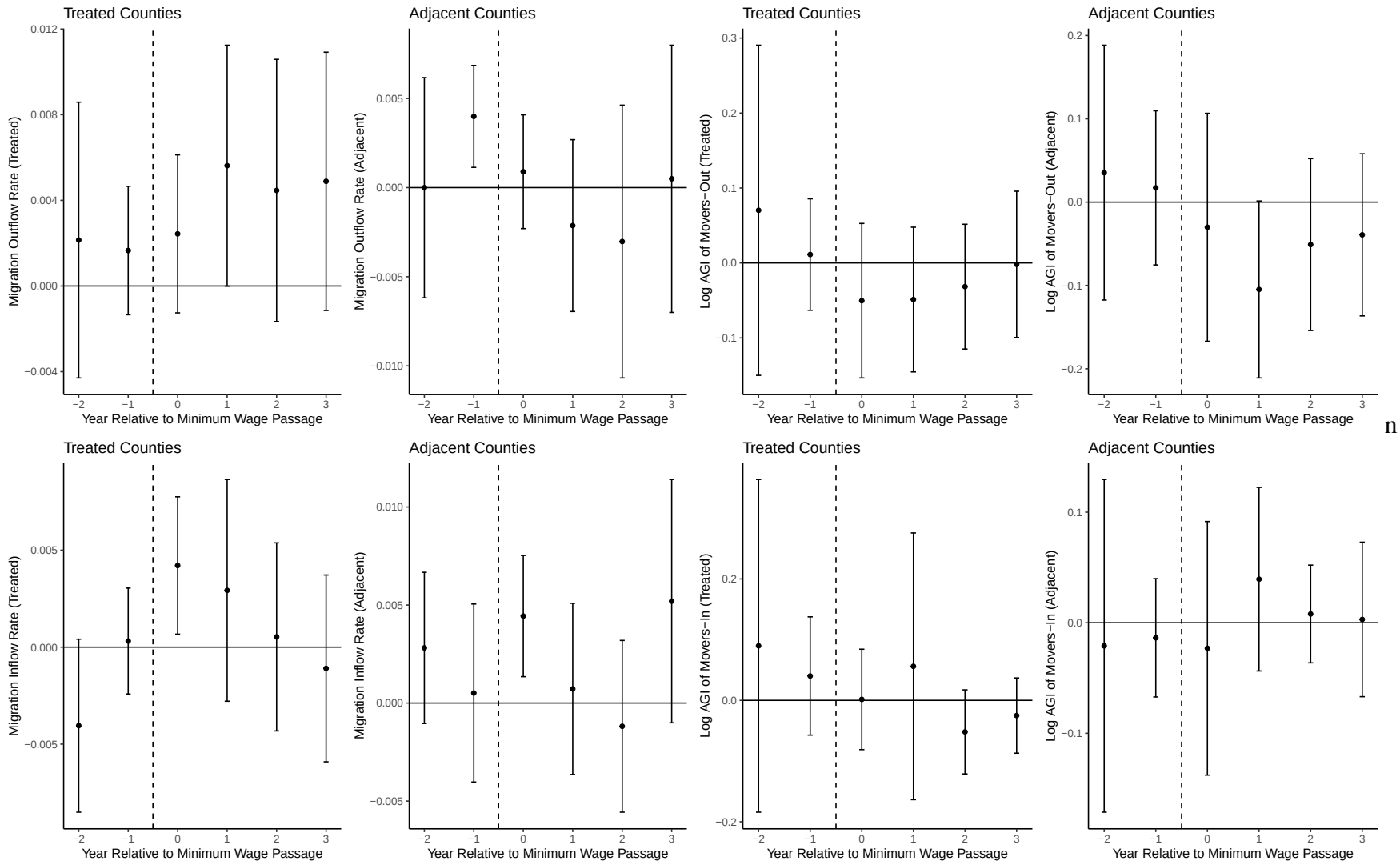


Figure 3.2: CS DiD Estimates: Totals

Notes: Figure presents CS difference-in-differences estimates.

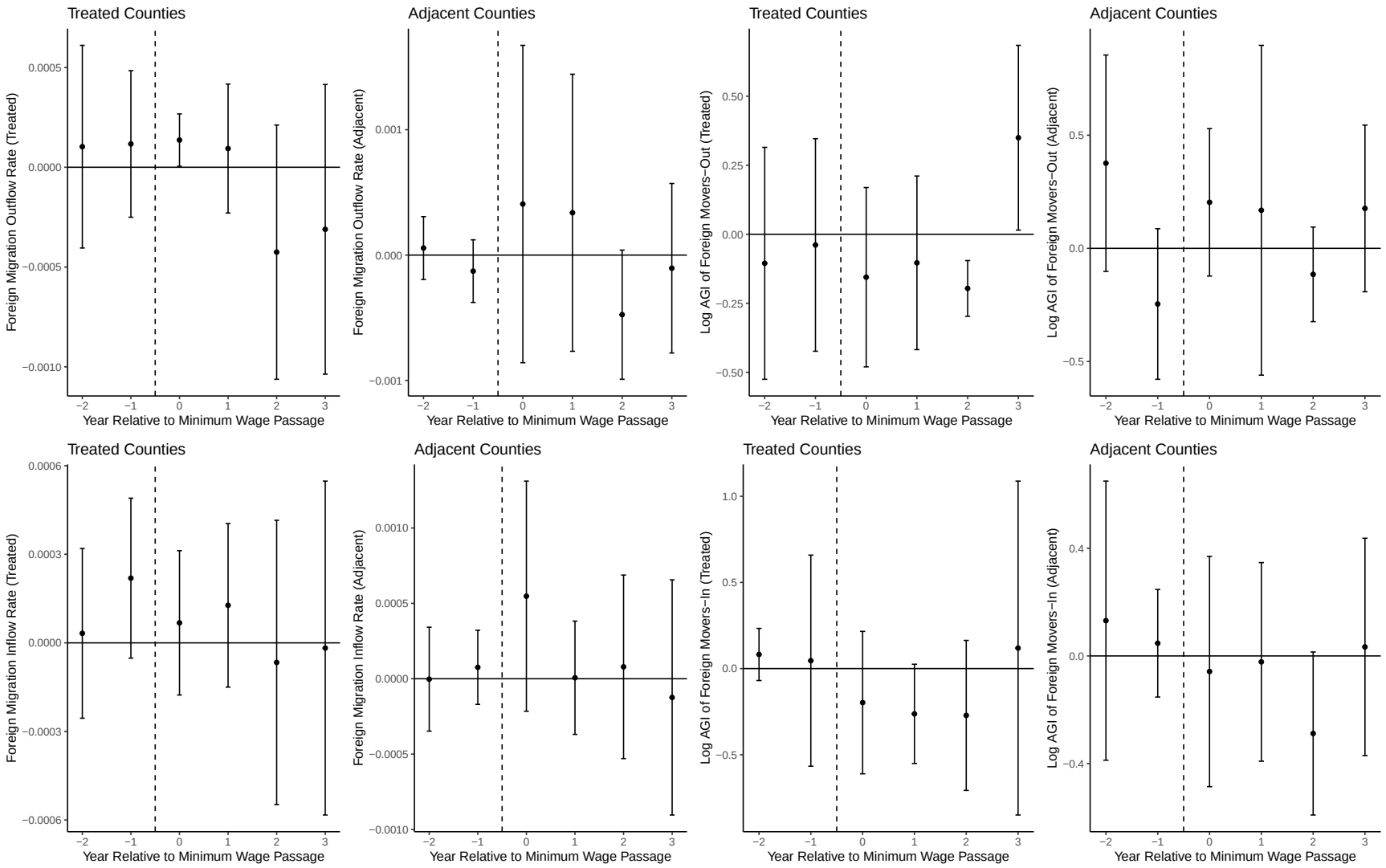


Figure 3.3: CS DiD Estimates: Foreign

Notes: Figure presents CS difference-in-differences estimates.

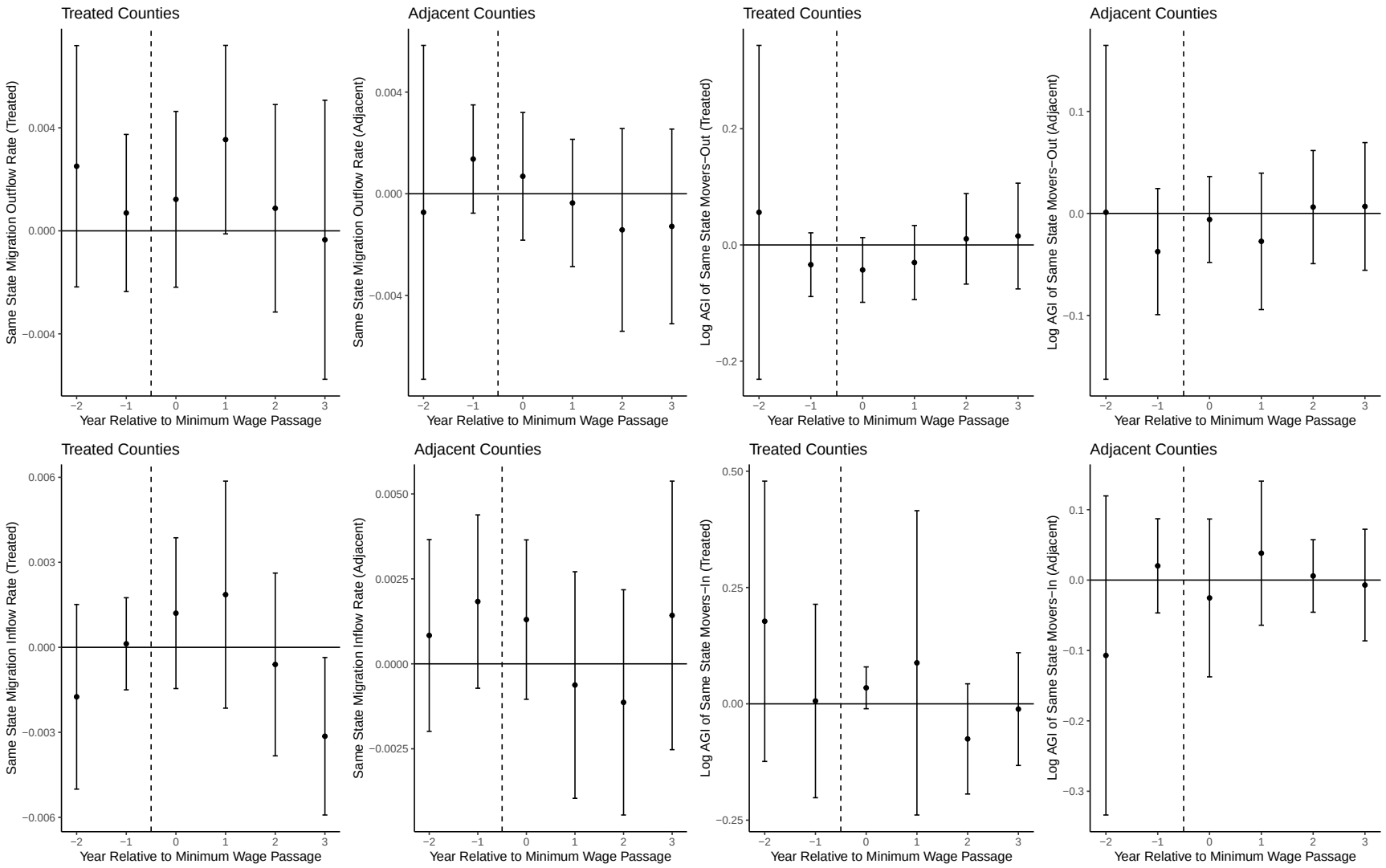


Figure 3.4: CS DiD Estimates: Same State

Notes: Figure presents CS difference-in-differences estimates.

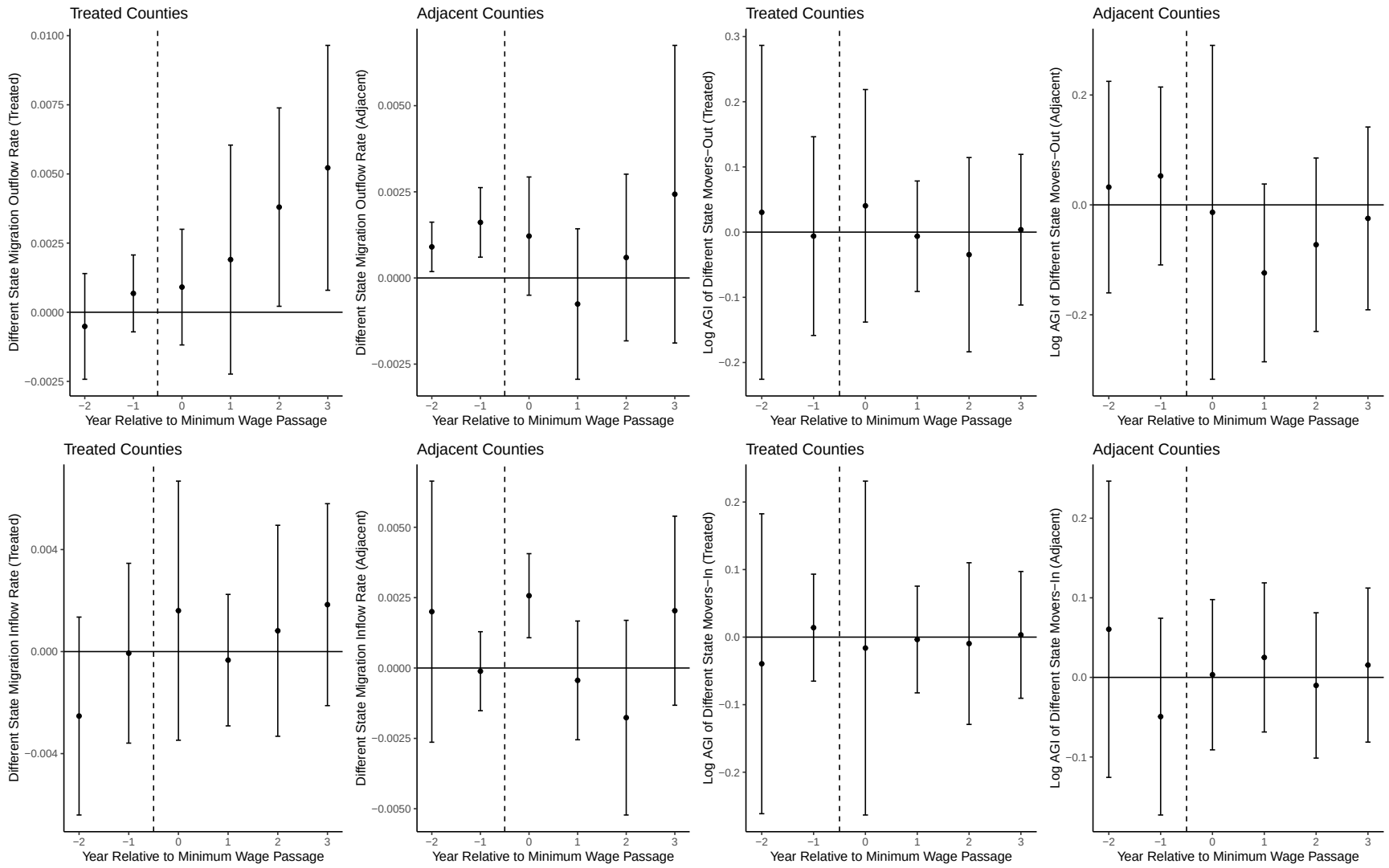


Figure 3.5: CS DiD Estimates: Different State

Notes: Figure presents CS difference-in-differences estimates.

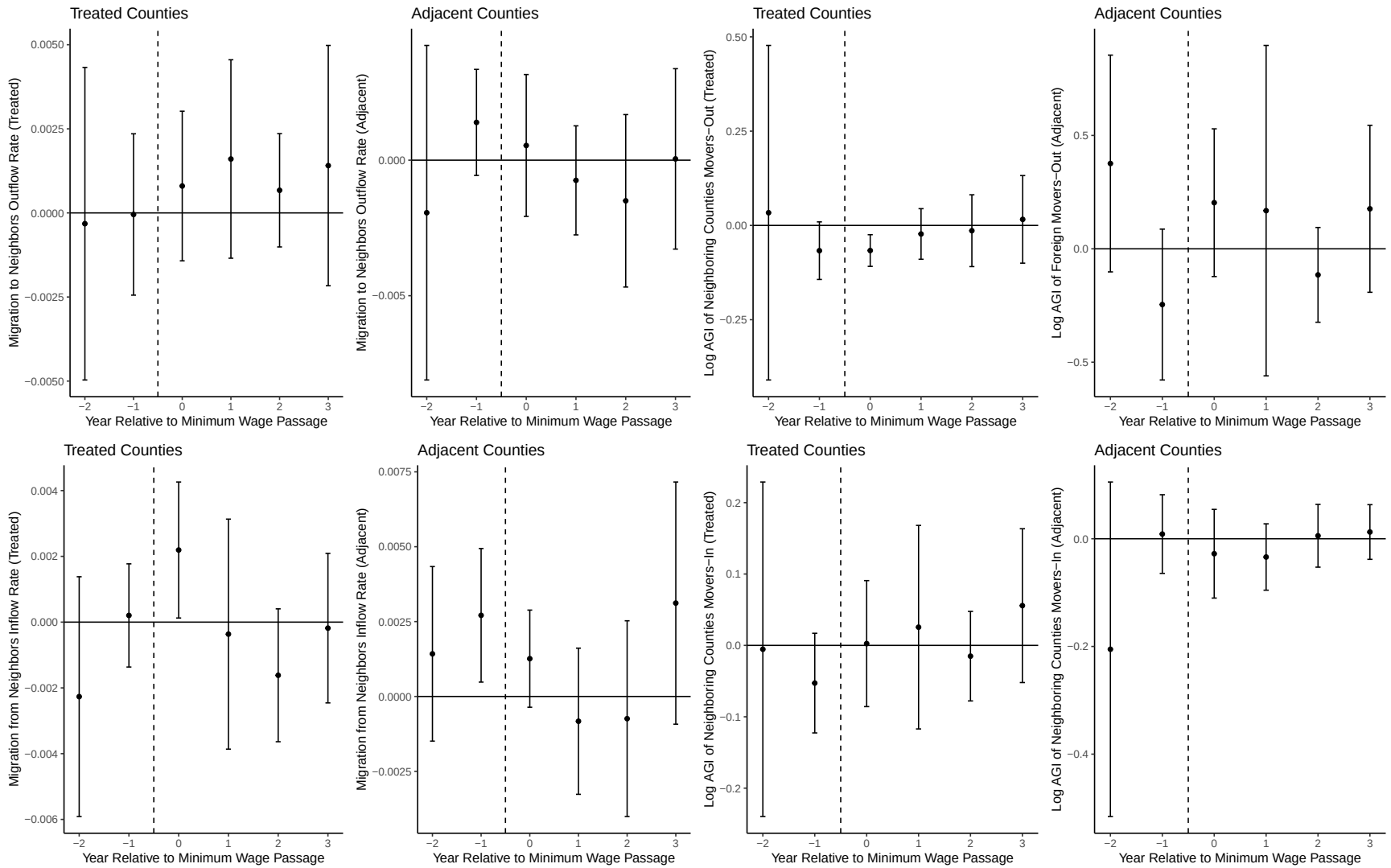


Figure 3.6: CS DiD Estimates: Neighboring Counties

Notes: Figure presents CS difference-in-differences estimates.

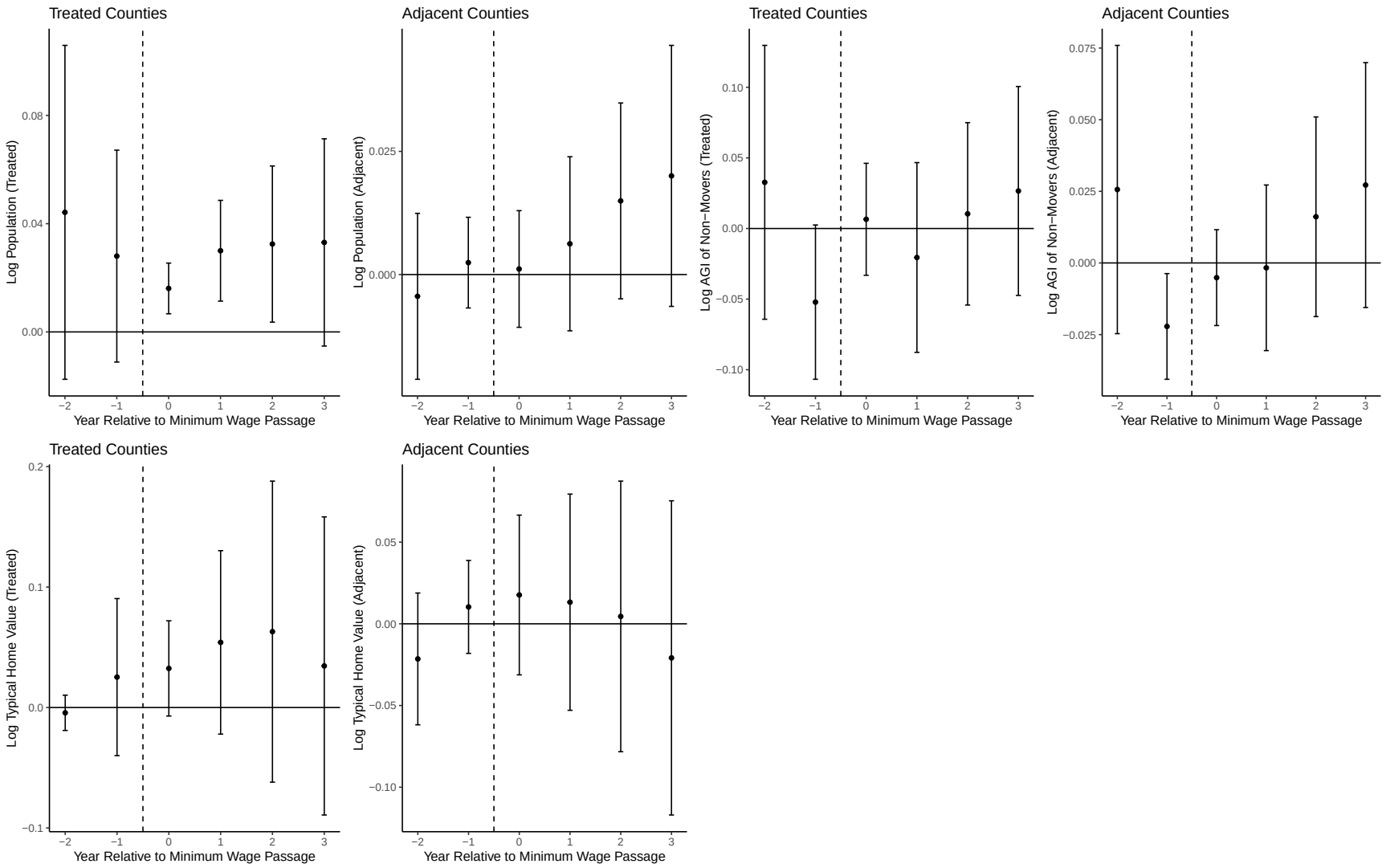


Figure 3.7: Matched DiD Estimates: Log Population, Non-Mover Income, and Home Prices

Notes: Figure presents matched difference-in-differences estimates.

Appendix A

Appendix for Chapter 1

A.1 Data Cleaning

Despite the substantial efforts by the teams at ProPublica and NJ Advance Media, the use of force dataset requires additional processing to be used for analysis. Here I outline the changes I make to the data.

The structure of each department's force reports and how many officers or subjects can be put on a single report vary. For consistency, I structure the data so that every observation is one subject who had force used against them by one officer. When there are multiple force reports for a single subject in an incident, I identify duplicates by computing similarity scores based on the Levenshtein distance between subject names, preventing situations such as a missing middle initial in one report or typos from creating repeat incidents. When officers from different departments use force against the same subject in the same incident as identified by the subject identifier and time of the incident, I continue to keep only the highest use of force from *any* department. For 167 force reports where multiple officers are recorded, I use the first officer listed. Because my empirical strategies are based on the most severe level of force used against a subject, for incidents where multiple officers use force against a single subject, I keep the officer who uses the greatest level of force,

choosing randomly in the event of ties.

Besides restructuring, the most significant modifications I make are for the types of force used. These are text strings and contain many irregularities. Some have typos, some do not directly correspond with force categories (for example, "Grabbed Gardening Tool Out Of Her Hand"), and some forms, notably from the New Jersey State Police, have their own names for force levels, such as "physical" and "mechanical." In keeping with my empirical strategy, I map these force levels to the highest type of force used in each incident.

Many force classifications require subjective judgments, outlined below. These decisions are further informed by officer narratives whenever possible. For instances where an officer uses a firearm as a blunt weapon ("pistol whipping"), I mark this as a use of a baton, the most similar type of force. Similarly, when blunt objects such as flashlights are used to strike a subject, I record them as batons. When batons are used for leverage in compliance holds, I count these incidents as compliance holds. I do not count restraints such as handcuffs or "the wrap" as force, but maneuvers to facilitate them may be, such as forcing one's arms behind their back. Unless there is an indication that a subject was punched, slapped, or otherwise struck, I classify hands on the subject as compliance holds rather than the hands/fists level. If an officer pushes a subject to get them to move, such as pushing them into a police vehicle, I classify it as a compliance hold. When the pushing is done to incapacitate a subject, such as pushing them off of a bicycle, I classify it as a use of hands/fists. I classify forcibly moving actively resisting subjects as compliance holds, but do not count moving passively resisting subjects, such as when a subject sits down and does not move.

Two other variables require manual cleaning: the unique officer identifier and the actions of the subject. The officer identifier present in the dataset does not track officers as they move across departments. Using available information on race, experience, rank, and geographic location, I manually create an officer identifier variable that follows officers

across departments. The subject actions variable is structured in the same manner as the force used variable and suffers from the same inconsistency problems. Again I manually map this variable to the "maximum" action a subject took that justified force. The lowest category is resisting, which includes physically resisting an officer's control, fleeing officer apprehension, and other actions that do not fit into another category. The next categories, physical, blunt, knife, and vehicle, cover both attacks and threats of attack. I split incidents involving firearms into threats with firearms and actually firing the weapons. I only count subject actions directed at humans, ignoring behaviors such as subjects attempting to kick out the windows of the police vehicle in which they are being held. I consider flailing and spitting or using other bodily fluids as projectiles to be a physical attack. When a subject attempts to disarm an officer, I mark it as a physical attack.

For the incident type variable, I use the categories present on the New Jersey Attorney General model form (Figure A1): crime in progress, domestic, other dispute, suspicious person, and other, reserving "other" for incidents that do not fit into another category. Some incidents have multiple types, which I allow, except for the "other" category.

I remove a small number of irregular or non-representative observations from the data. I drop 963 observations (over 600 of them would have been dropped anyway due to other missing variables) where no reason for the force was given, as it is more likely that the officer-stated reason is missing than nonexistent. I remove a handful of incidents involving nonlethal firearms ("riot guns"), as these are not standard police carry. I do not include the constructive authority category of force, incidents where a firearm or stun gun is drawn but not fired to induce compliant behavior, as no physical force is used in these situations. I remove 44 individuals whose indicated races do not fall within the categories of White, Black, Hispanic, or Asian/Pacific Islander, such as people marked as "mixed."

Table A1: Summary Statistics for Municipalities Represented in Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Population	454	18,424.43	25,402.93	10,373.00	296.00	277,140.00
Population/square mile)	454	4,020.84	5,616.57	2,612.12	39.13	55,880.00
Median household income	454	87,296.63	31,652.80	83,006.00	26,214.00	190,625.00
Gini coefficient	454	0.43	0.05	0.43	0.33	0.60
Land area (sq. miles)	454	11.16	16.50	3.67	0.10	111.13
Adjusted pop. % White	454	68.52	22.53	74.44	1.82	99.29
Adjusted pop. % Black	454	8.41	12.17	3.87	0.00	88.77
Adjusted pop. % Hispanic	454	14.16	14.30	9.19	0.00	82.94
Adjusted pop. % Asian/PI	454	7.66	9.15	4.51	0.00	58.60
Violent crimes per 1000	454	1.67	2.59	0.85	0.00	25.66
Romney vote share 2012 presidential election	454	45.20	14.79	47.74	1.31	81.82
County mean Black-White IAT *D*-score	21	0.35	0.05	0.35	0.24	0.41

Notes: Data cover all municipalities present in the force reports after cleaning and processing that are served by their own police department in the force reports data, i.e. not those served by New Jersey State Police. Data have been cleaned as described in Appendix A.1. Data come from the 2010 Census, 2012-2016 American Community Survey five-year estimates, FBI's Uniform Crime Reporting program, the New Jersey Division of Elections, and Project Implicit.

Table A2: Summary Statistics for Police Departments Represented in Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Avg. num. full-time police employees	455	42.75	76.41	23	0	1,088
% officers White	455	72.42	35.01	90.58	0.00	100.00
% officers Black	455	2.64	6.80	0.00	0.00	78
% officers Hispanic	455	2.04	5.69	0.00	0.00	53.49
% officers Asian/Pacific Islander	455	0.53	2.13	0.00	0.00	30

Notes: Data cover all police departments in New Jersey from 2012 through 2016 with force reports after cleaning and processing. Not all departments' racial breakdowns sum to 100%. Data from ProPublica and NJ Advance Media.

Table A3: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force (No Controls)

	Unenhanced	Enhanced	Fired Weapons
Intercept (White)	0.456^{***}	0.121^{***}	0.003^{***}
	(0.014)	(0.007)	(0.000)
Subject Black	0.098^{***}	0.052^{***}	0.003^{**}
	(0.014)	(0.012)	(0.001)
Subject Hispanic	0.088^{***}	0.032^{***}	0.003^{*}
	(0.024)	(0.012)	(0.002)
Subject Asian/PI	-0.011	-0.009	-0.003^{***}
	(0.028)	(0.018)	(0.000)
Clustering	Dept.	Dept.	Dept.
Outcome mean	0.505	0.146	0.004
R^2	0.009	0.005	0.000
Num. obs.	39322	39322	39322

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from regressions where the outcome is a binary measure of whether police used force of at least the specified severity on indicators for subject race with no other covariates.

Table A4: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Unenhanced		Enhanced		Fired Weapons	
Subject Black/Hispanic	0.015^{**}	0.020^{**}	0.028^{***}	0.033^{***}	0.001	0.001
	(0.006)	(0.008)	(0.005)	(0.006)	(0.001)	(0.001)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.505	0.505	0.146	0.145	0.004	0.004
R^2	0.180	0.515	0.118	0.466	0.180	0.554
Num. obs.	39184	39133	39184	39133	39184	39133

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ *Notes:* Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A5: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Unenhanced		Enhanced		Fired Weapons	
Subject Black	0.019^{***}	0.022^{***}	0.034^{***}	0.040^{***}	0.001	0.001
	(0.006)	(0.009)	(0.005)	(0.006)	(0.001)	(0.001)
Subject Hispanic	-0.002	0.006	0.004	0.003	0.002	0.001
	(0.010)	(0.013)	(0.007)	(0.010)	(0.002)	(0.002)
Subject Asian/PI	-0.022	-0.005	0.004	0.005	-0.005^{***}	-0.002
	(0.026)	(0.042)	(0.017)	(0.027)	(0.002)	(0.002)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.505	0.505	0.146	0.145	0.004	0.004
R^2	0.181	0.515	0.118	0.467	0.180	0.554
Num. obs.	39184	39133	39184	39133	39184	39133

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A6: Conditional Logit Odds Ratios on Race in Intensity of Force Used, Conditional on Force

	Unenhanced	Enhanced	Fired Weapons
Subject Black/Hispanic	1.046^{*}	1.247^{*}	1.439
	$(1.010, 1.083)$	$(1.154, 1.347)$	$(0.791, 2.619)$
Fixed effects	Dept.	Dept.	Dept.
Clustering	Dept.	Dept.	Dept.
Num. obs.	39322	39322	39322

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Notes: Table reports conditional logit estimates from Equation 1.2, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, and year fixed effects, stratified by department. Asymmetric 95% confidence intervals with clustering at the department level based on exponentiating the log odds confidence interval are in parentheses.

Table A7: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject and Subject at Least Physically Threatened/Attacked Officer or Another

	Unenhanced		Enhanced		Fired Weapons	
Subject Black/Hispanic	0.047^{***}	0.036	0.058^{***}	0.076^{**}	-0.001	0.001
	(0.013)	(0.032)	(0.014)	(0.030)	(0.002)	(0.004)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.677	0.677	0.245	0.245	0.009	0.009
R^2	0.189	0.728	0.140	0.708	0.276	0.841
Num. obs.	7778	7770	7778	7770	7778	7770

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions are run on a subset of the data where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night.

Table A8: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject

	Unenhanced		Enhanced		Fired Weapons	
Subject Black/Hispanic	0.026^{***}	0.029^{**}	0.033^{***}	0.044^{***}	0.000	-0.000
	(0.009)	(0.014)	(0.007)	(0.011)	(0.001)	(0.002)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.541	0.541	0.166	0.166	0.004	0.004
R^2	0.187	0.603	0.132	0.559	0.215	0.672
Num. obs.	19860	19840	19860	19840	19860	19840

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions are run on a subset of the data where a subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night.

Table A9: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Subject at Least Physically Threatened/Attacked Officer or Another

	Unenhanced		Enhanced		Fired Weapons	
Subject Black/Hispanic	0.024^{**}	0.031^{*}	0.047^{***}	0.058^{***}	0.001	0.001
	(0.010)	(0.018)	(0.010)	(0.017)	(0.002)	(0.003)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.642	0.642	0.218	0.218	0.009	0.009
R^2	0.174	0.665	0.125	0.627	0.227	0.733
Num. obs.	14580	14558	14580	14558	14580	14558

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions are run on a subset of the data where the subject at least physically threatened or attacked an officer or another person.

Table A10: Predicting 100 x (Department x Black/Hispanic Winsorized Empirical Bayes Estimates), RHS Numeric Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only

	Unenhanced	Enhanced	Fired Weapons
Log median household income	0.640	−0.555	0.023
	(0.998)	(1.062)	(0.139)
Gini coefficient	1.435	−0.789	−0.113
	(1.017)	(1.159)	(0.176)
Log population	3.700*	−0.732	−0.407
	(1.974)	(1.590)	(0.383)
Log population per square mile	1.677	0.831	0.055
	(1.242)	(1.166)	(0.192)
Log number of police	−4.781**	0.313	0.197
	(2.271)	(1.916)	(0.422)
% officers not White or Asian/PI	−1.754**	0.194	0.051
	(0.683)	(0.748)	(0.172)
% population 18-65 not White or Asian/PI	1.006	0.214	0.086
	(1.604)	(1.344)	(0.222)
Violent crime rate	0.918	0.582	0.021
	(0.756)	(0.923)	(0.107)
2012 presidential election Romney vote %	0.678	−0.329	−0.036
	(1.506)	(1.598)	(0.218)
County mean Black-White IAT *D*-score	−1.675*	1.489	−0.032
	(0.908)	(0.910)	(0.167)
Prob > F	0.11	0.71	0.70
R^2	0.134	0.069	0.045
Num. obs.	92	92	92

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each non-binary variable on the right-hand side of the regression has been divided by twice its standard deviation.

_____ POLICE DEPARTMENT

USE OF FORCE REPORT

A. Incident Information

Date	Time	Day of Week	Location	**INCIDENT NUMBER**

Type of Incident
☐ Crime in progress ☐ Domestic ☐ Other dispute ☐ Suspicious person ☐ Traffic stop
☐ Other (specify)

B. Officer Information

Name (Last, First, Middle)		Badge #	Sex	Race	Age	Injured Y / N	Killed Y / N
Rank	Duty assignment	Years of service		On-Duty Y / N		Uniform Y / N	

C1. Subject 1 (List only the person who was the subject of the use of force by the officer listed in Section B.)

Name (Last, First, Middle)	Sex	Race	Age	Weapon Y / N	Injured Y / N	Killed Y / N
☐ Under the influence ☐ Other unusual condition (specify)	Arrested Y / N		Charges			

Subject's actions (check all that apply)	Officer's use of force toward this subject (check all that apply)	
☐ Resisted police officer control	☐ Compliance hold	Firearms Discharge
☐ Physical threat/attack on officer or another	☐ Hands/fists	☐ Intentional
☐ Threatened/attacked officer or another with blunt object	☐ Kicks/feet	☐ Accidental
☐ Threatened/attacked officer or another with knife/cutting object	☐ Chemical/natural agent	
☐ Threatened/attacked officer or another with motor vehicle	☐ Strike/use baton or other object	Number of Shots Fired ____
☐ Threatened officer or another with firearm	☐ Canine	Number of Hits ____
☐ Fired at officer or another		[Use 'UNK' if unknown]
☐ Other (specify)	☐ Other (specify)	

C2. Subject 2 (List only the person who was the subject of the use of force by the officer listed in Section B.)

Name (Last, First, Middle)	Sex	Race	Age	Weapon Y / N	Injured Y / N	Killed Y / N
☐ Under the influence ☐ Other unusual condition (specify)	Arrested Y / N		Charges			

Subject's actions (check all that apply)	Officer's use of force toward this subject (check all that apply)	
☐ Resisted police officer control	☐ Compliance hold	Firearms Discharge
☐ Physical threat/attack on officer or another	☐ Hands/fists	☐ Intentional
☐ Threatened/attacked officer or another with blunt object	☐ Kicks/feet	☐ Accidental
☐ Threatened/attacked officer or another with knife/cutting object	☐ Chemical/natural agent	
☐ Threatened/attacked officer or another with motor vehicle	☐ Strike/use baton or other object	Number of Shots Fired ____
☐ Threatened officer or another with firearm	☐ Canine	Number of Hits ____
☐ Fired at officer or another		[Use 'UNK' if unknown]
☐ Other (specify)	☐ Other (specify)	

➤ If this officer used force against more than two subjects in this incident, attach additional USE OF FORCE REPORTS.

Signature:	Date:
Print Supervisor Name:	Supervisor Signature:

7/2001

Figure A1: New Jersey Model Use of Force Report

Notes: Figure obtained from the website of the New Jersey Attorney General.

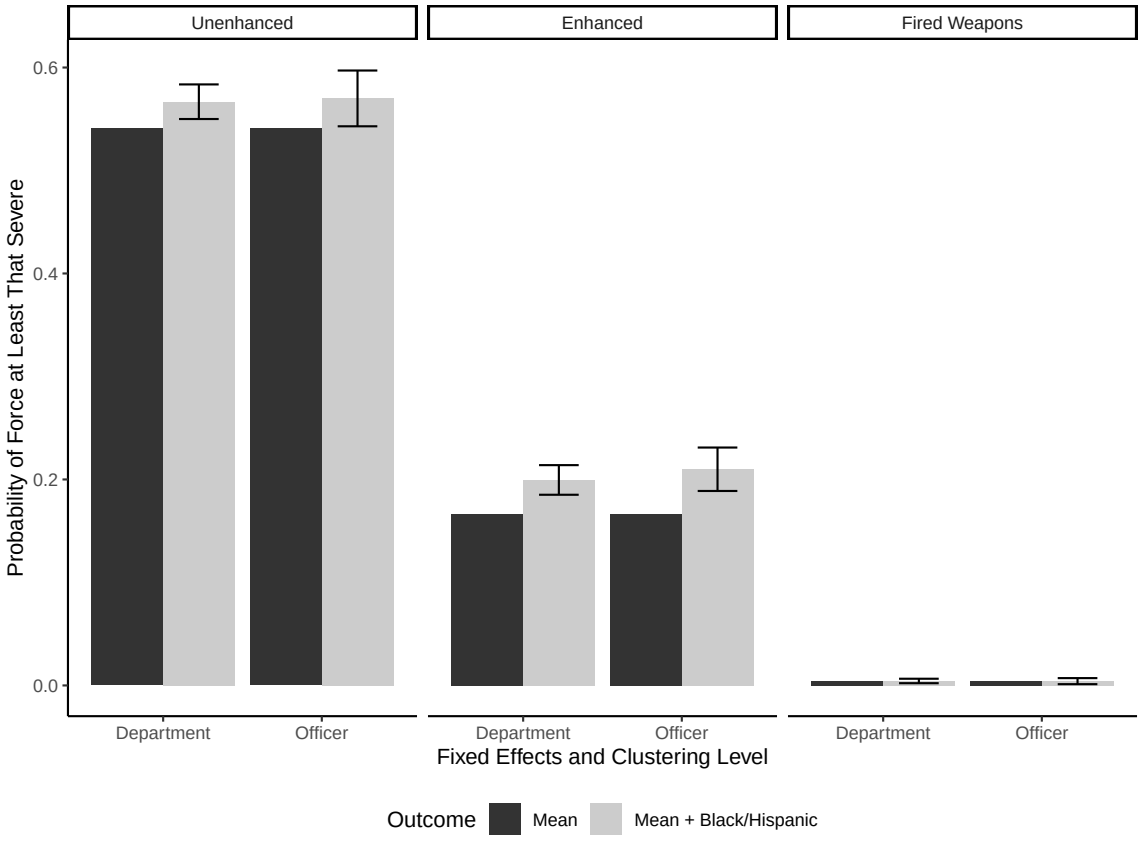


Figure A2: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject)

Notes: Figure presents results from a series of OLS models estimated on the subset of the data where a subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night. Regressions are fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

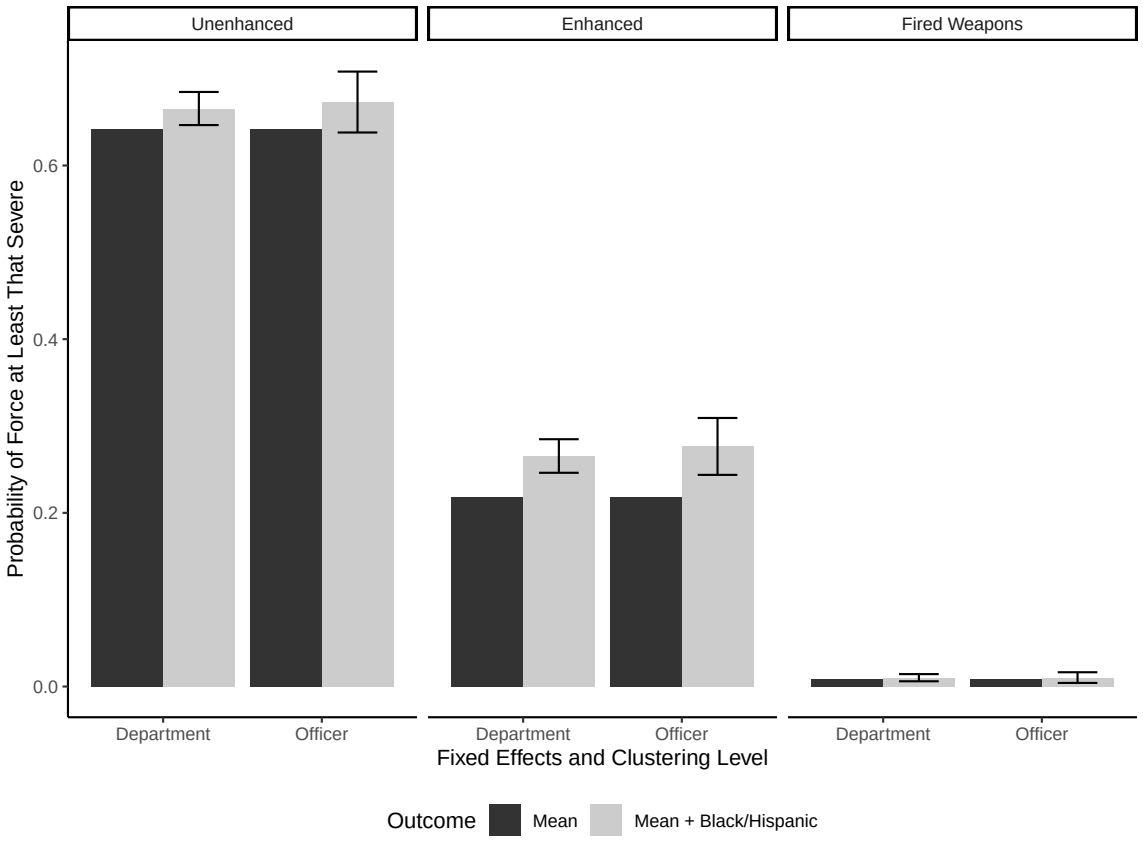


Figure A3: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Subject at Least Physically Threatened/Attacked Officer or Another)

Notes: Figure presents results from a series of OLS models estimated on the subset of the data where the subject at least physically threatened or attacked an officer or another. Regressions are fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

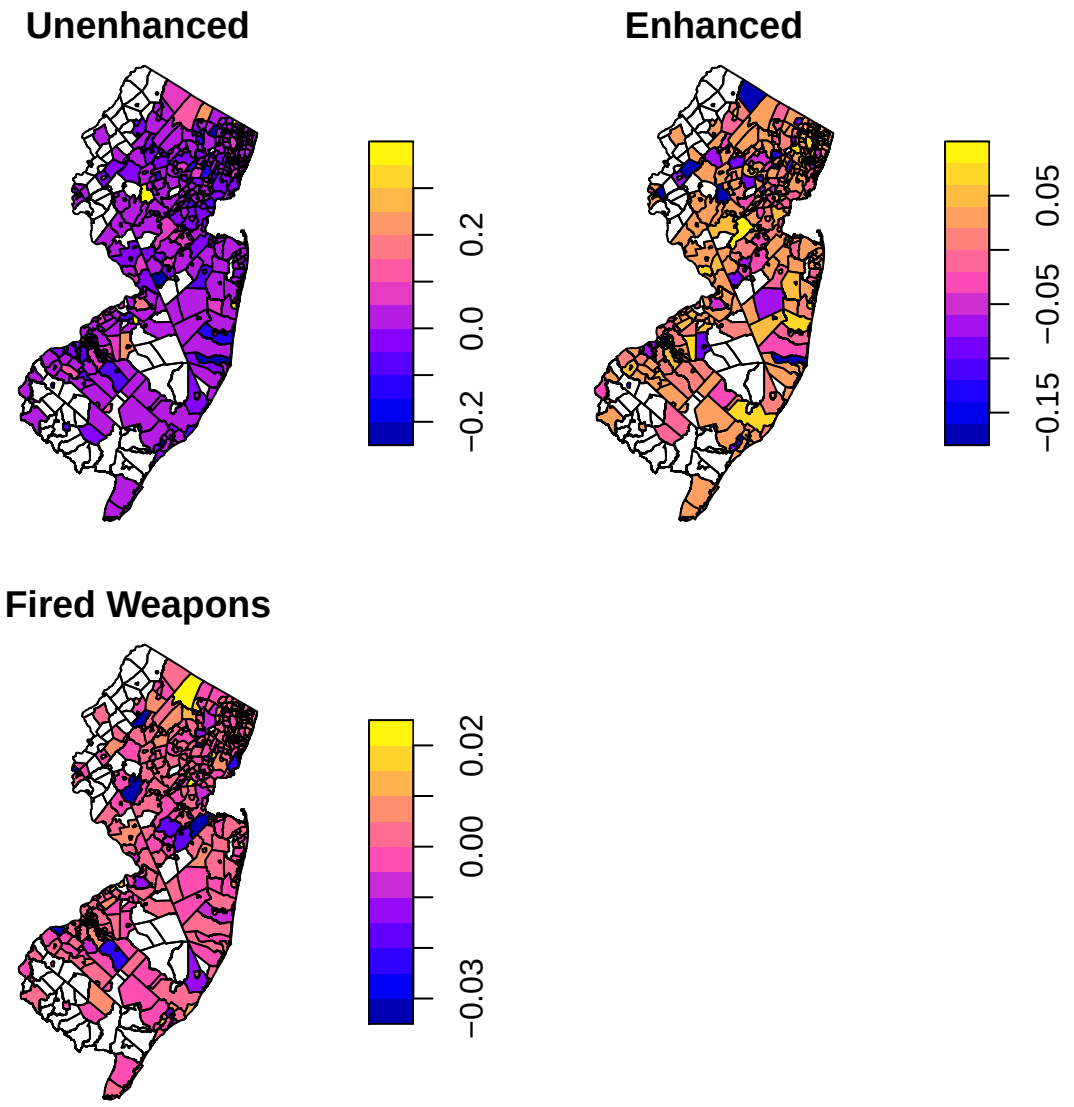


Figure A4: Heatmap of of Winsorized Empirical Bayes Estimates of Department-Black/Hispanic Interactions

Notes: Figure presents heatmaps of department-specific racial disparities β between White/Asian/-Pacific Islander subjects and Black/Hispanic subjects as estimated from the empirical Bayes estimator in Equation 1.5. Each subgraph shows results from regressions with the specified outcome outcome: whether, conditional on any force being used, force of at least the specified severity was used. Municipalities in white are not present in the data.

A.2 Results with Full Set of Force Outcomes

Table A11: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force (No Controls)

	Takedowns	Hands/Fists	Leg Strikes	Pepper Spray	Batons	Fired Weapons
Intercept (White)	0.456^{***}	0.401^{***}	0.162^{***}	0.121^{***}	0.029^{***}	0.003^{***}
	(0.014)	(0.012)	(0.007)	(0.007)	(0.002)	(0.000)
Subject Black	0.098^{***}	0.086^{***}	0.065^{***}	0.052^{***}	0.017^{***}	0.003^{**}
	(0.014)	(0.013)	(0.011)	(0.012)	(0.004)	(0.001)
Subject Hispanic	0.088^{***}	0.078^{***}	0.042^{***}	0.032^{***}	0.018^{***}	0.003^{*}
	(0.024)	(0.021)	(0.011)	(0.012)	(0.006)	(0.002)
Subject Asian/PI	−0.011	−0.009	−0.003	−0.009	0.003	-0.003^{***}
	(0.028)	(0.027)	(0.020)	(0.018)	(0.009)	(0.000)
Clustering	Dept.	Dept.	Dept.	Dept.	Dept.	Dept.
Outcome mean	0.505	0.444	0.193	0.146	0.038	0.004
R^2	0.009	0.007	0.006	0.005	0.002	0.000
Num. obs.	39322	39322	39322	39322	39322	39322

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from regressions where the outcome is a binary measure of whether police used force of at least the specified severity on indicators for subject race with no other covariates.

Table A12: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Takedowns		Hands/Fists		Leg Strikes		Pepper Spray		Batons		Fired Weapons	
Subject Black/Hispanic	0.015**	0.020**	0.019***	0.024***	0.031***	0.038***	0.028***	0.033***	0.010***	0.010***	0.001	0.001
	(0.006)	(0.008)	(0.006)	(0.008)	(0.005)	(0.007)	(0.005)	(0.006)	(0.002)	(0.003)	(0.001)	(0.001)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.505	0.505	0.444	0.444	0.193	0.192	0.146	0.145	0.038	0.038	0.004	0.004
R^2	0.180	0.515	0.177	0.521	0.114	0.452	0.118	0.466	0.075	0.422	0.180	0.554
Num. obs.	39184	39133	39184	39133	39184	39133	39184	39133	39184	39133	39184	39133

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A13: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Takedowns		Hands/Fists		Leg Strikes		Pepper Spray		Batons		Fired Weapons	
Subject Black	0.019^{***}	0.022^{***}	0.023^{***}	0.026^{***}	0.038^{***}	0.045^{***}	0.034^{***}	0.040^{***}	0.011^{***}	0.012^{***}	0.001	0.001
	(0.006)	(0.009)	(0.006)	(0.009)	(0.006)	(0.007)	(0.005)	(0.006)	(0.003)	(0.003)	(0.001)	(0.001)
Subject Hispanic	−0.002	0.006	0.002	0.010	0.003	0.006	0.004	0.003	0.004	−0.002	0.002	0.001
	(0.010)	(0.013)	(0.010)	(0.013)	(0.008)	(0.011)	(0.007)	(0.010)	(0.004)	(0.006)	(0.002)	(0.002)
Subject Asian/PI	−0.022	−0.005	−0.022	−0.012	0.010	0.007	0.004	0.005	−0.005	−0.002	-0.005^{***}	−0.002
	(0.026)	(0.042)	(0.025)	(0.040)	(0.019)	(0.031)	(0.017)	(0.027)	(0.008)	(0.014)	(0.002)	(0.002)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.505	0.505	0.444	0.444	0.193	0.192	0.146	0.145	0.038	0.038	0.004	0.004
R^2	0.181	0.515	0.177	0.521	0.115	0.452	0.118	0.467	0.076	0.422	0.180	0.554
Num. obs.	39184	39133	39184	39133	39184	39133	39184	39133	39184	39133	39184	39133

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. *Notes:* Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A14: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject and Subject at Least Physically Threatened/Attacked Officer or Another

	Takedowns		Hands/Fists		Leg Strikes		Pepper Spray		Batons		Fired Weapons	
Subject Black/Hispanic	0.047^{***}	0.036	0.044^{***}	0.035	0.055^{***}	0.066^{**}	0.058^{***}	0.076^{**}	0.024^{***}	0.036^{**}	-0.001	0.001
	(0.013)	(0.032)	(0.014)	(0.033)	(0.014)	(0.032)	(0.014)	(0.030)	(0.008)	(0.018)	(0.002)	(0.004)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.677	0.677	0.639	0.639	0.312	0.312	0.245	0.245	0.073	0.073	0.009	0.009
R^2	0.189	0.728	0.184	0.730	0.136	0.694	0.140	0.708	0.133	0.708	0.276	0.841
Num. obs.	7778	7770	7778	7770	7778	7770	7778	7770	7778	7770	7778	7770

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1.1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions are run on a subset of the data where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night.

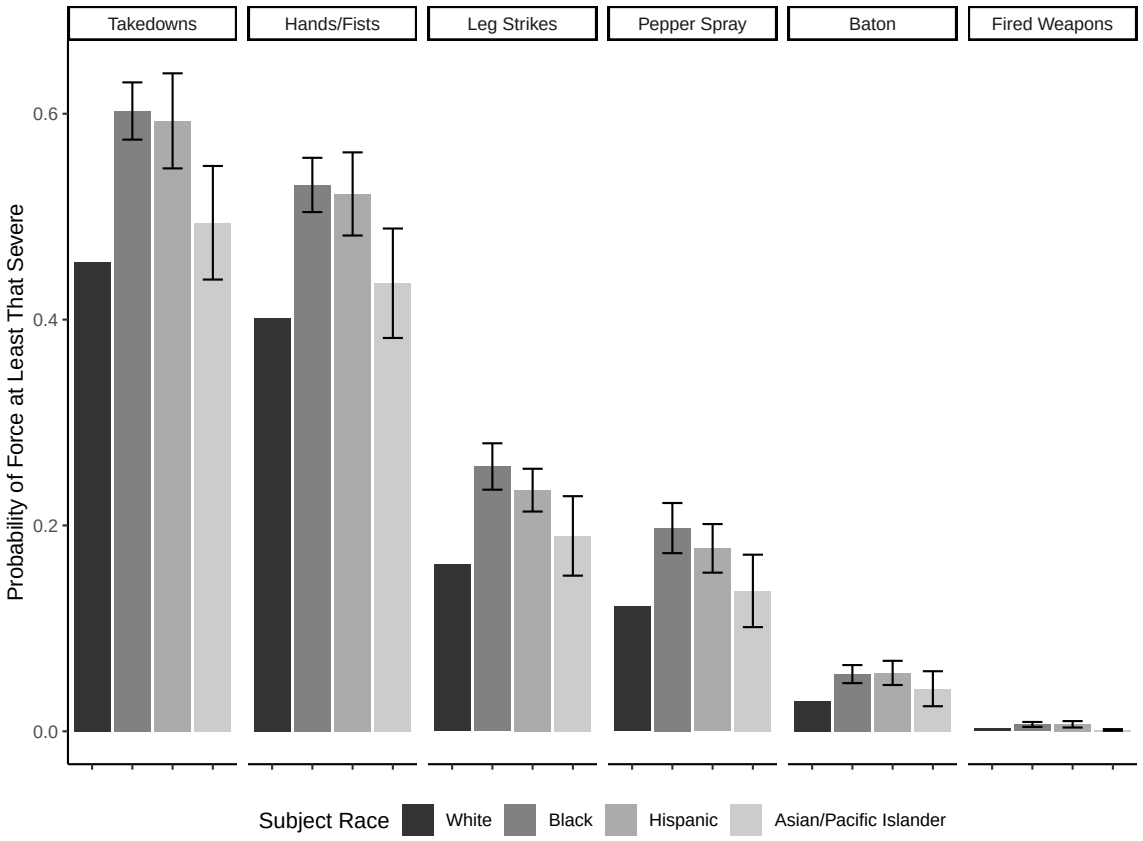


Figure A5: Overall Racial Disparities (No Controls)

Notes: Figure presents results from a series of OLS models regressing outcomes on a full set of racial dummies. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Non-White bars are obtained by taking the intercept (Whites) and adding the coefficient on that race. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the department level.

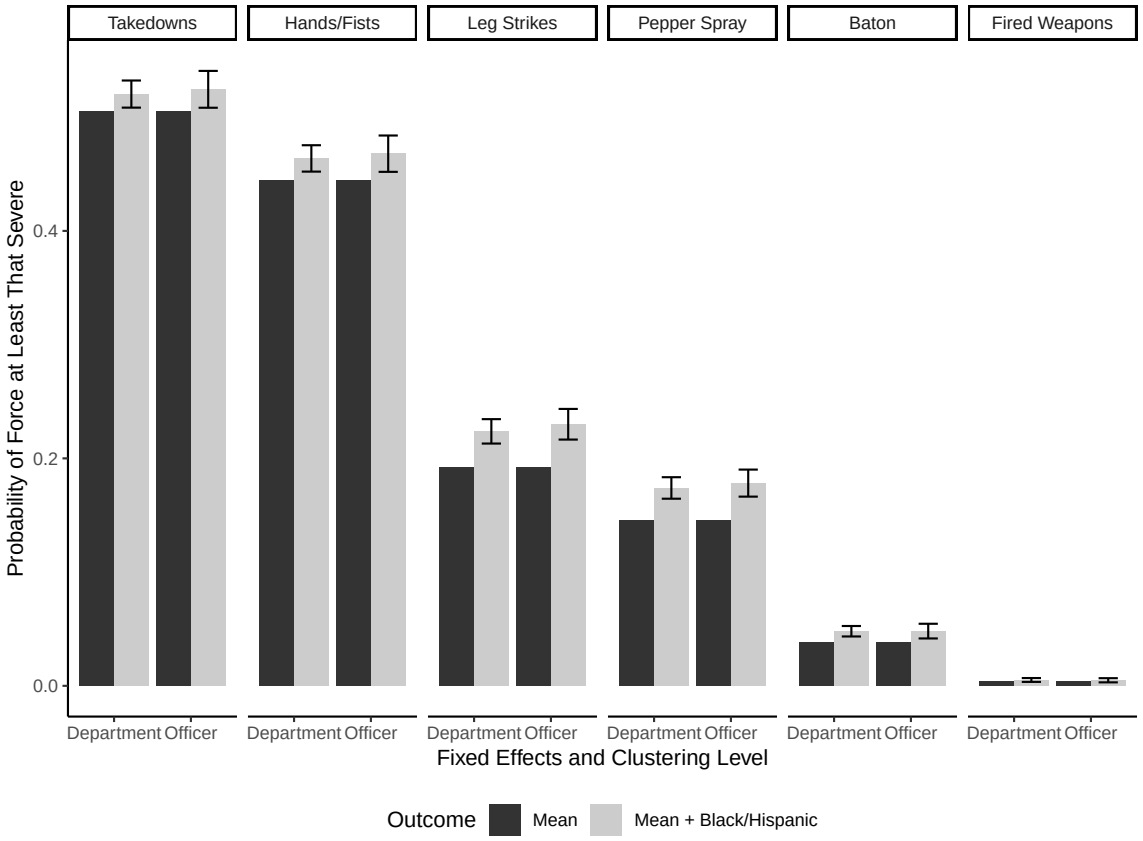


Figure A6: Overall Racial Disparities for Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

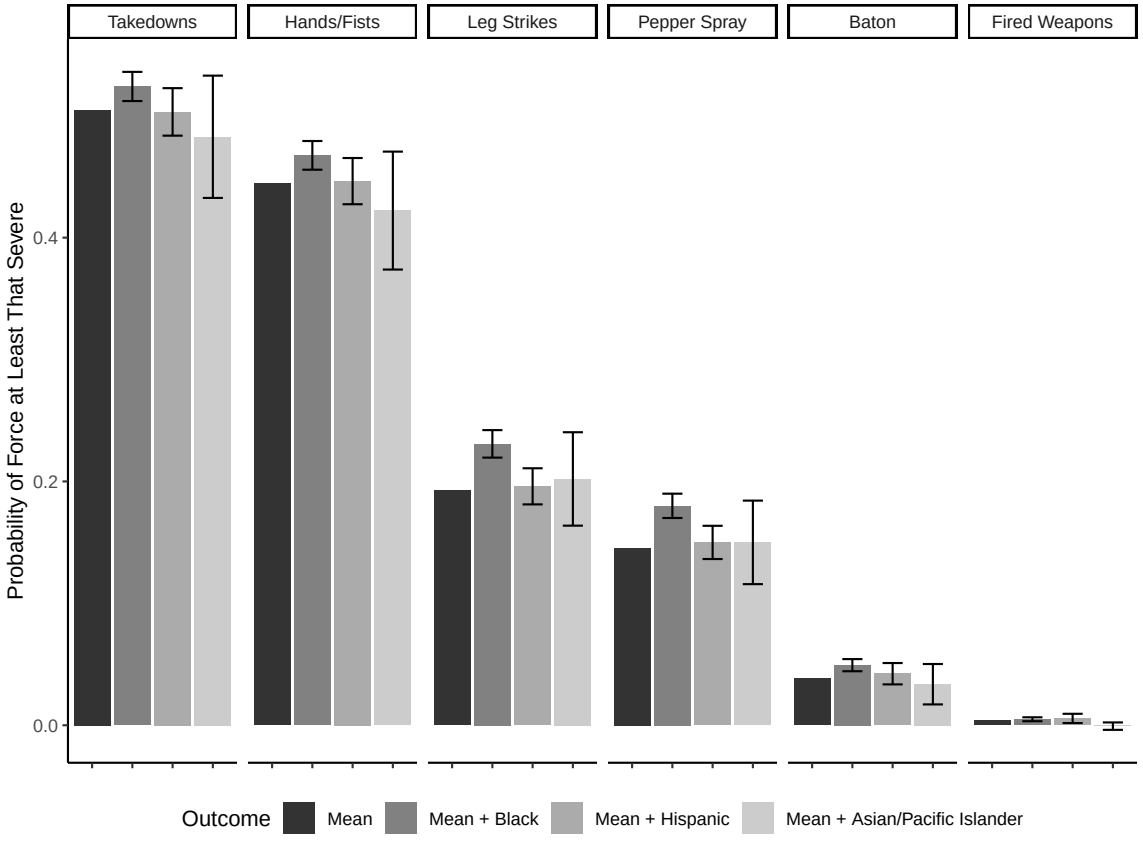


Figure A7: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Department Fixed Effects and Clustering)

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the department level.

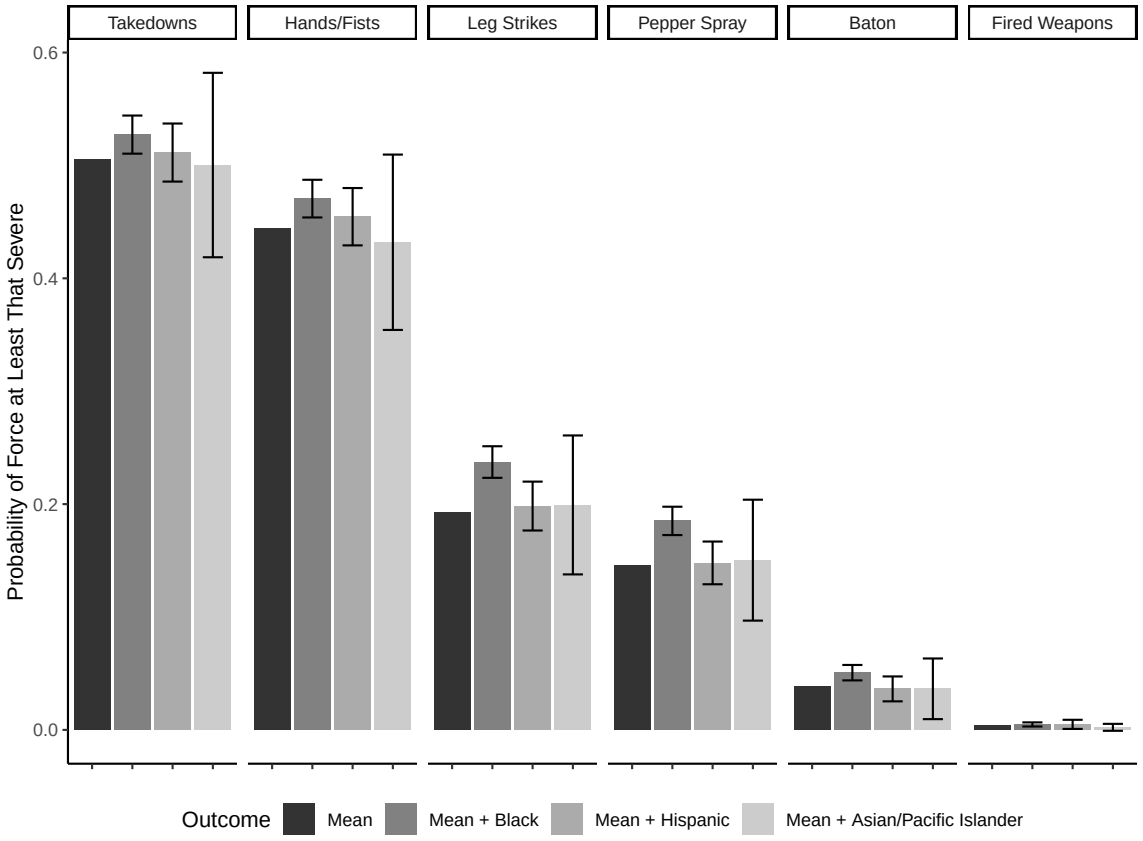


Figure A8: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity, Full Race Dummies (Officer Fixed Effects and Clustering)

Notes: Figure presents results from a series of OLS models fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the officer level.

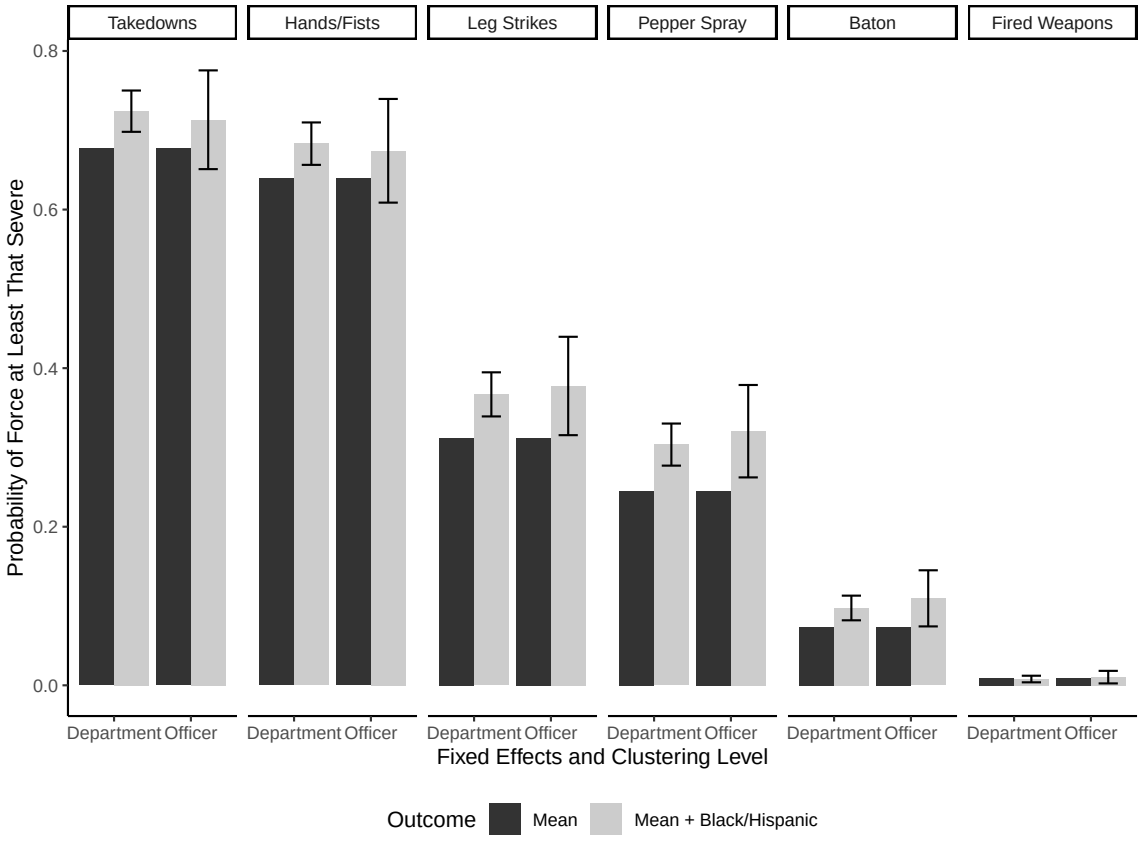


Figure A9: Overall Racial Disparities of Subject Being Black/Hispanic on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject and Subject at Least Physically Threatened/Attacked Officer or Another)
Notes: Figure presents results from a series of OLS models estimated on the subset of the data where a subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night, and where the subject at least physically threatened or attacked an officer or another. The latter restriction is equivalent to dropping observations where the most severe actions by the subject was resisting. Regressions are fit via Equation 1.1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black/Hispanic coefficient.

A.3 Results with OLS Estimates of Departmental Disparities

Table A15: Summary Statistics for OLS Estimates of Department x Black/Hispanic Subject Interactions

	Unenhanced	Enhanced	Fired Weapons
SD	0.263	0.182	0.027
Min	-0.968	-0.800	-0.215
P01	-0.518	-0.355	-0.070
P05	-0.359	-0.220	-0.020
P25	-0.094	-0.063	-0.002
Median	0.016	0.011	0.000
P75	0.144	0.082	0.004
P95	0.555	0.297	0.017
P99	0.809	0.827	0.065
Max	1.090	1.034	0.276
Mean	0.042	0.023	0.001
$\% \leq 0$	0.450	0.463	0.488

Notes: Table reports OLS estimates of departmental racial disparities from Equation 1.3, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race interacted with a department indicator, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department fixed effects.

Table A16: Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation

	Unenhanced	Enhanced	Fired Weapons
Log median household income	1.899	-5.135^{**}	0.244
	(3.875)	(2.590)	(0.281)
Gini coefficient	2.179	1.299	-0.191
	(3.590)	(2.642)	(0.277)
Log population	-0.665	1.041	0.348
	(9.311)	(8.050)	(0.598)
Log population per square mile	5.762^{*}	4.127^{*}	0.357^{*}
	(3.469)	(2.445)	(0.215)
Log number of police	-2.870	-0.002	-0.389
	(8.928)	(7.445)	(0.586)
% officers not White or Asian/PI	0.085	1.884	0.038
	(2.794)	(2.354)	(0.162)
% population 18-65 not White or Asian/PI	3.250	-2.748	-0.689
	(4.354)	(3.066)	(0.480)
Violent crime rate	2.195	-0.240	0.035
	(2.928)	(2.371)	(0.291)
2012 presidential election Romney vote %	11.241^{**}	3.389	-0.746
	(5.670)	(4.198)	(0.686)
County mean Black-White IAT *D*-score	-1.791	0.768	0.054
	(2.867)	(1.925)	(0.512)
Prob > F	0.54	0.20	0.45
R^2	0.024	0.023	0.015
Num. obs.	405	405	405

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each variable on the right-hand side of the regression has been divided by twice its standard deviation.

Table A17: Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation

	Unenhanced	Enhanced	Fired Weapons
Log median household income	21.273**	−9.137	−0.098
	(9.124)	(6.479)	(0.837)
Gini coefficient	−7.978	−6.949	−1.362
	(6.105)	(6.863)	(1.193)
Log population	13.507	2.103	−1.544
	(12.288)	(13.500)	(2.442)
Log population per square mile	8.177	7.808	−0.030
	(6.405)	(6.307)	(0.604)
Log number of police	−18.280	−5.013	0.712
	(12.633)	(14.300)	(2.155)
% officers not White or Asian/PI	−13.345**	−1.420	−0.393
	(5.316)	(4.648)	(0.723)
% population 18-65 not White or Asian/PI	14.858*	−1.228	−1.384
	(8.539)	(7.204)	(1.594)
Violent crime rate	10.006**	1.668	0.254
	(4.861)	(2.496)	(0.386)
2012 presidential election Romney vote %	8.342	2.838	−2.075
	(8.271)	(5.805)	(2.102)
County mean Black-White IAT *D*-score	3.123	1.684	−0.670
	(4.802)	(4.088)	(0.682)
Dept. has nonzero asset forfeiture	2.889	−5.102	−1.867
	(8.830)	(7.590)	(2.113)
Dept. uses written aptitude test in selecting recruits	−2.332	3.077	0.443
	(6.187)	(4.144)	(0.578)
Dept. maintains a written community policing plan	8.106	4.578	−0.303
	(5.718)	(3.733)	(0.626)
Dept. formally surveys local residents	1.690	−4.152	0.594
	(7.716)	(3.739)	(0.941)
Dept. has written policy on stop and frisk	−0.672	−2.697	0.728
	(6.014)	(5.676)	(1.249)
Dept. has written policy on foot pursuits	4.831	−1.573	−0.648
	(5.204)	(4.205)	(0.811)
Dept. has written policy on unbiased policing	−12.127	−1.351	1.771
	(9.089)	(6.691)	(2.186)
Dept. chief executive is female	−7.464	0.399	−0.245
	(10.654)	(9.714)	(2.019)
Dept. chief executive is non-Hispanic White	8.585	−1.983	0.270
	(11.985)	(4.686)	(0.805)
Prob > F	0.36	0.26	0.68
R^2	0.265	0.172	0.160
Num. obs.	83	83	83

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each non-binary variable on the right-hand side of the regression has been divided by twice its standard deviation.

Table A18: Predicting 100 x (Department x Black/Hispanic Interaction OLS Coefficients), RHS Variables Normalized by 2x Standard Deviation, 2016 LEMAS Departments Only

	Unenhanced	Enhanced	Fired Weapons
Log median household income	21.338**	−9.842*	−0.170
	(8.489)	(5.352)	(0.734)
Gini coefficient	−4.501	−5.705	−1.426
	(6.161)	(6.021)	(1.166)
Log population	15.334	2.742	−1.841
	(12.741)	(10.710)	(1.764)
Log population per square mile	10.764**	8.297	−0.465
	(5.360)	(5.710)	(0.743)
Log number of police	−20.995*	−7.563	0.898
	(12.523)	(10.728)	(1.404)
% officers not White or Asian/PI	−13.466***	−0.607	−0.168
	(4.859)	(4.137)	(0.520)
% population 18-65 not White or Asian/PI	15.025**	−2.217	−1.076
	(7.463)	(5.899)	(1.317)
Violent crime rate	10.929**	1.129	−0.037
	(4.351)	(2.474)	(0.294)
2012 presidential election Romney vote %	9.431	2.061	−1.671
	(8.061)	(5.266)	(1.697)
County mean Black-White IAT *D*-score	1.017	2.010	−0.701
	(3.747)	(3.391)	(0.547)
Prob > F	0.19	0.58	0.89
R^2	0.220	0.124	0.085
Num. obs.	83	83	83

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports coefficients from a regression of the department-specific racial disparities $\beta_{p,OLS}$ for Black/Hispanic subjects on municipal/departmental-level covariates as estimated by Equation 1.3. Each column corresponds to a regression with the outcome being whether, conditional on force being used, force of at least the specified severity was used. Each non-binary variable on the right-hand side of the regression has been divided by twice its standard deviation.

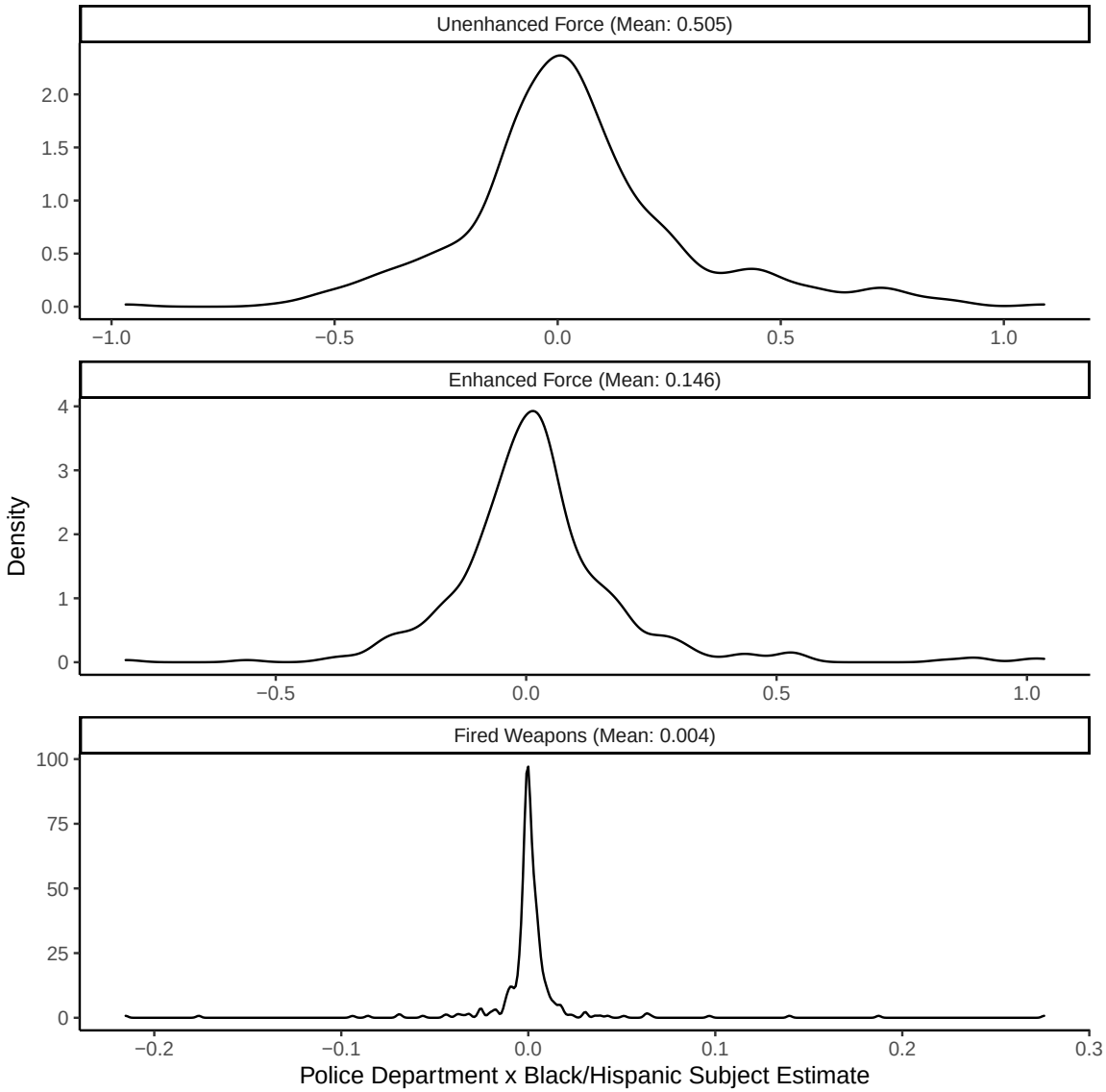


Figure A10: Distribution of Unshrunk Estimates of Department-Black/Hispanic Interactions

Notes: Figure presents kernel density estimates of department-specific racial disparities β between White/Asian/Pacific Islander subjects and Black/Hispanic subjects as estimated from the regression in Equation 1.3 with Gaussian kernels and the Silverman (1986) rule-of-thumb bandwidth. Each subgraph shows results from regressions with the specified outcome: whether, conditional on any force being used, force of at least the specified severity was used.

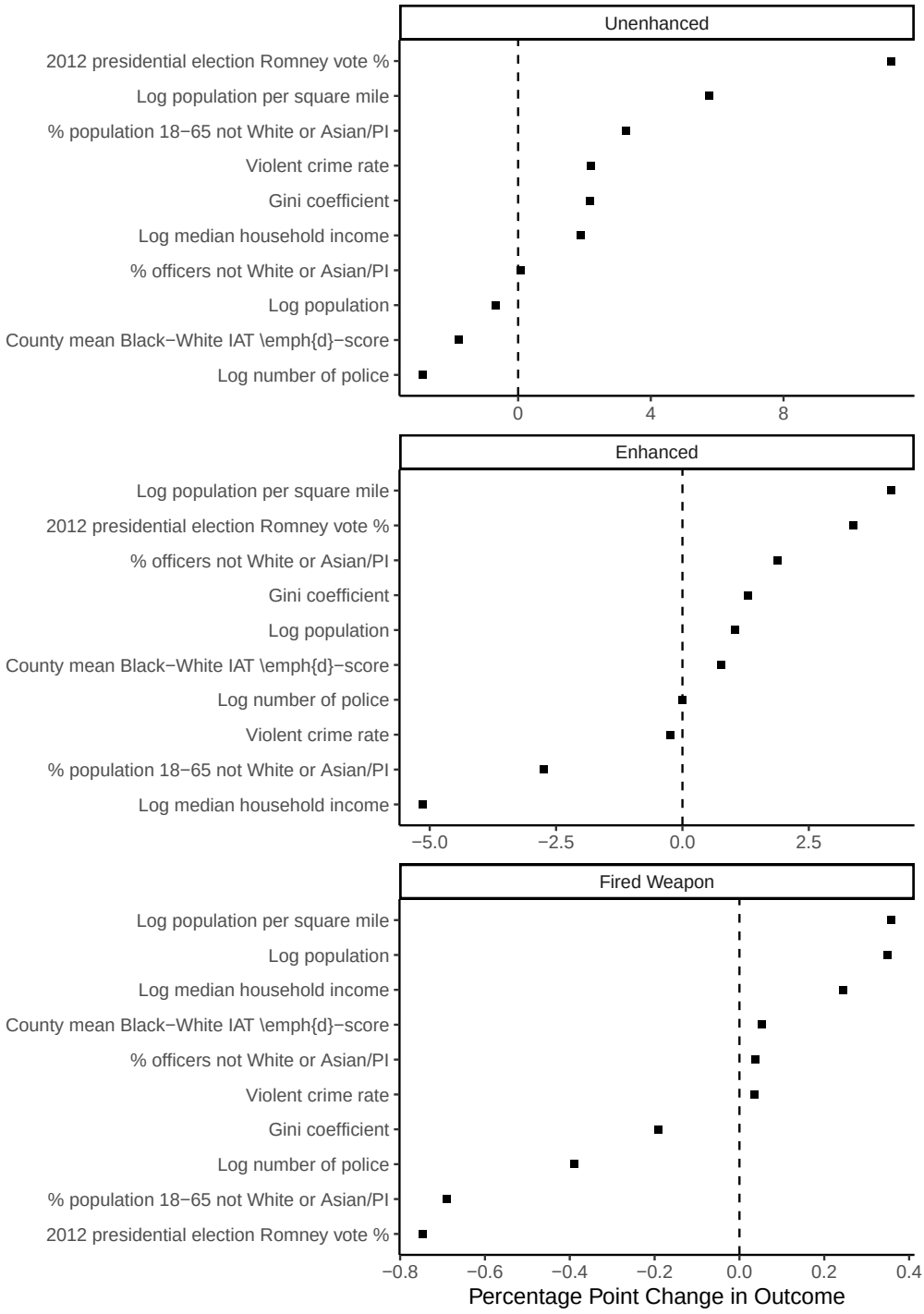


Figure A11: Coefficients from Regression of OLS Estimates of Department-Black/Hispanic Interactions on Possible Correlates

Notes: Figure presents coefficients from a regression of OLS estimates of departmental disparities on the listed variables as described in Equation 1.6. Each variable on the right-hand side of the regression has been divided by twice its standard deviation to aid in interpretation.

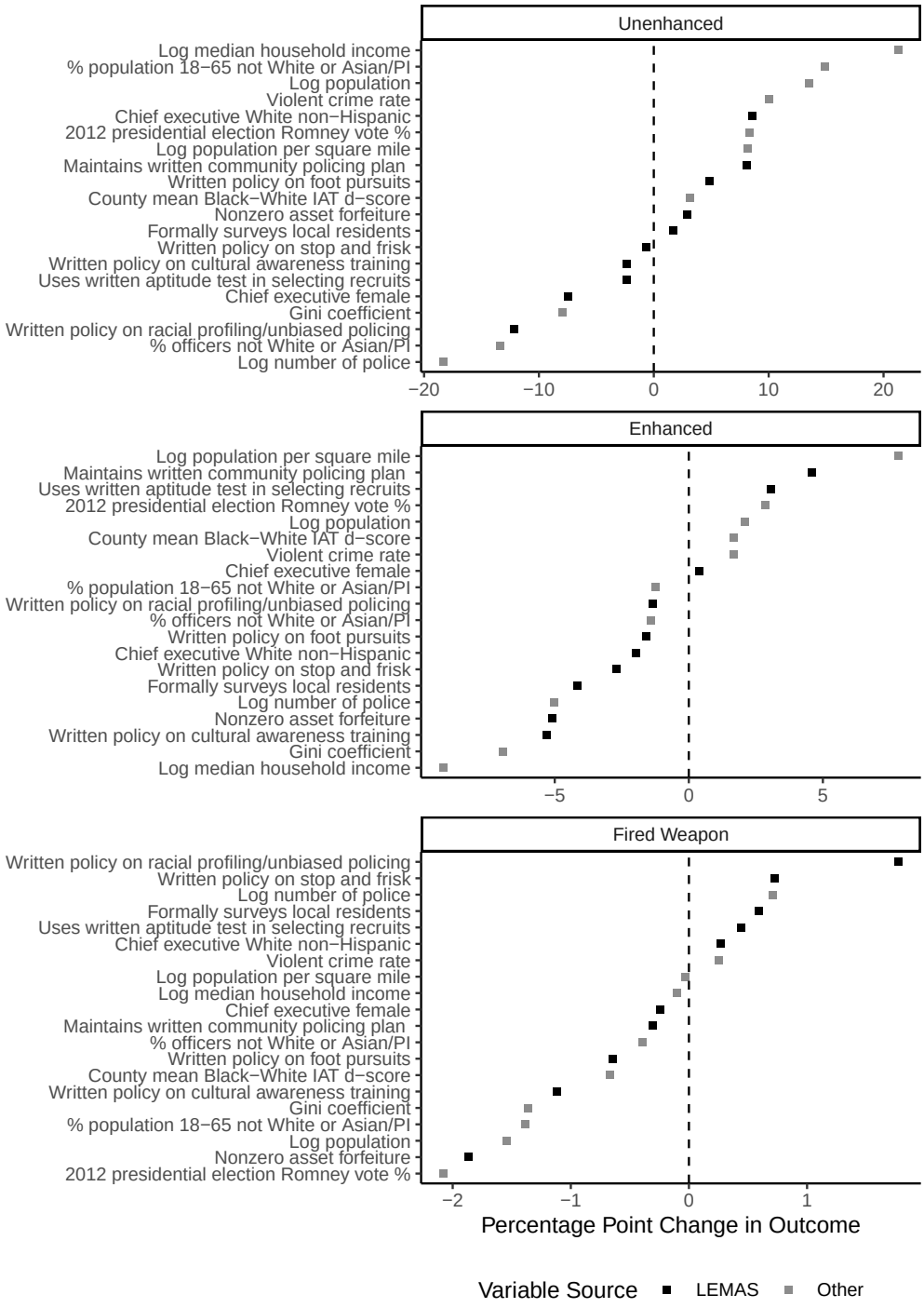


Figure A12: Coefficients from Regression of OLS Estimates of Department-Black/Hispanic Interactions on Possible Correlates, Departments in 2016 LEMAS Only
Notes: Figure presents coefficients from a regression of OLS estimates of departmental disparities on the listed variables as described in Equation 1.6. Non-binary variables on the right-hand side of the regression has been divided by twice its standard deviation to aid in interpretation.

Appendix B

Appendix for Chapter 2

B.1 Theory: Estimator Properties, Bin Selectors, and MSE Decompositions

B.1.1 Properties of the Type I and Type II Error Probability Estimators

Following Section 2.2, Proposition 1 states the properties of the g-specific estimator $\hat{p}(\gamma, g, d)$.

Proposition 1. *Under Assumption 1 and for a given DGP g*

1. $E[\hat{p}(\gamma, g, d)] = p(\gamma, g, d)$
2. $\hat{p}(\gamma, g, d) \xrightarrow{\mathbb{P}} p(\gamma, g, d)$ *as* $M \to \infty$
3. $M \cdot \hat{p}(\gamma, g, d) \sim Binomial(M, p(\gamma, g, d))$
4. $\sqrt{M}(\hat{p}(\gamma, g, d) - p(\gamma, g, d)) \Rightarrow N(0, p(\gamma, g, d) \cdot (1 - p(\gamma, g, d)))$.

The proof of the proposition follows trivially from Assumption 1.

Parts 1 and 2 of the proposition state that $\hat{p}(\gamma, g, d)$ is an unbiased and consistent estimator for $p(\gamma, g, d)$. Part 3 states that $\hat{p}(\gamma, g, d)$ has a scaled binomial finite sample distribution, for which methods such as the Clopper-Pearson confidence interval have been developed for finite sample inference, which allows us to construct confidence intervals on the g-specific type I and II error probabilities. The standard asymptotic normality result is provided in Part 4, but we do not invoke it in our analysis, as M is equal to eight in our experiments due to resource constraints.

Now we state the properties of the overall estimator $\hat{\bar{p}}(\gamma, d)$.

Proposition 2. *Under Assumptions 1 and 2*

1. $E[\hat{\bar{p}}(\gamma, d)] = \bar{p}(\gamma, d)$
2. $\hat{\bar{p}}(\gamma, d) \xrightarrow{\mathbb{P}} \bar{p}(\gamma, d)$ *as* $J \to \infty$
3. *Conditional on* $\{g_j\}_{j=1}^{J}$, $M \cdot J \cdot \hat{\bar{p}}(\gamma, d)$ *follows a Poisson binomial distribution*
4. *As* $J \to \infty, \sqrt{J}(\hat{\bar{p}}(\gamma, d) - \bar{p}(\gamma, d)) \Rightarrow N\left(0, var_{g \in \mathscr{G}}(\hat{p}(\gamma, g, d))\right)$, *with* $var_{g \in \mathscr{G}}(\hat{p}(\gamma, g, d)) = \frac{1}{M}\bar{p}(\gamma, d) + (1 - \frac{1}{M})E[p(\gamma, g, d)^2] - \bar{p}(\gamma, d)^2$.

Proof. Part 1:

$$\begin{aligned} E[\hat{\bar{p}}(\gamma, d)] &= \frac{1}{J}\sum_{j=1}^{J} E_{g_j \in \mathscr{G}}\left[E[\hat{p}(\gamma, g_j, d)|g_j]\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} E_{g_j \in \mathscr{G}}\left[p(\gamma, g_j, d)\right] \\ &= \bar{p}(\gamma, d) \end{aligned}$$

Part 2: Assumptions 1 and 2 imply that, for any J, $\{\hat{p}(\gamma, g_j, d)\}_{1 \le j \le J}$ is a set of independent and identically distributed random variables. Because $\hat{p}(\gamma, g_j, d)^2$ is uniformly bounded between zero and one, the weak law of large numbers for triangular arrays (see, for example, Durett, 2010) applies to the set $\{\hat{p}(\gamma, g_j, d)\}_{1 \le j \le J, J=1,2,\ldots}$. Because the expectation of $\hat{p}(\gamma, g_j, d)$ is $\bar{p}(\gamma, d)$ for each j (Part 1 of the proposition), we have the desired

result.[1]

Part 3: the statement simply follows the definition of a Poisson binomial distribution.

Part 4: Similar to the proof of part 2 of the Proposition, the result follows from an application of the Central Limit Theorem for Triangular Arrays (or the Lindeberg-Feller Central Limit Theorem):

$$\sqrt{J}(\hat{\bar{p}}(\gamma,d) - \bar{p}(\gamma,d)) \Rightarrow N\left(0, var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d))\right).$$

The law of total variance implies that

$$\begin{aligned}
&var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d)) \\
=&E_{g\in\mathscr{G}}[var(\hat{p}(\gamma,g,d)|g)] + var_{g\in\mathscr{G}}(E[\hat{p}(\gamma,g,d)|g]) \\
=&E_{g\in\mathscr{G}}\left[\frac{p(\gamma,g,d)\,(1-p(\gamma,g,d))}{M}\right] \\
&+ var_{g\in\mathscr{G}}(p(\gamma,g,d))
\end{aligned}$$

and the last equality follows from Proposition 1. Because

$$var_{g\in\mathscr{G}}(p(\gamma,g,d)) = E_{g\in\mathscr{G}}[p(\gamma,g,d)^2] - E_{g\in\mathscr{G}}[p(\gamma,g,d)]^2$$

we have

$$var_{g\in\mathscr{G}}(p(\gamma,g,d)) = \frac{1}{M}\bar{p}(\gamma,d) + (1-\frac{1}{M})E_{g\in\mathscr{G}}[p(\gamma,g,d)^2] - \bar{p}(\gamma,d)^2$$

following simple algebra. □

[1]The application of the law of large numbers for triangular arrays, as opposed to its standard counterpart, allows $\hat{p}(\gamma,g_j,d)$ to change as J increases. Take $\hat{p}(\gamma,g_1,d)$, for example. We accommodate the scenarios where M, the number of participants who see g_1, changes with J, or where we allocate different individuals to see g_1 as we increase the number of DGPs sampled.

Parts 1 and 2 of Proposition 2 establish unbiasedness and consistency of $\hat{\bar{p}}(\gamma,d)$ as an estimator for $\bar{p}(\gamma,d)$. Part 3 provides the finite-sample distribution of $\hat{\bar{p}}(\gamma,d)$; the Poisson binomial distribution is that of a sum of independent but potentially non-identically distributed Bernoulli random variables. Part 4 states the asymptotic normality result for $\hat{\bar{p}}(\gamma,d)$ as $J \to \infty$. In principle, we can consistently estimate the variance and use Part 4 to conduct inference on $\bar{p}(\gamma,d)$ if J is large. As with M, however, we choose a moderate J (11) in our experiments due to resource constraints, and the asymptotic normality statement is, therefore, more conceptual than practical. Nevertheless, we make the following observations on the variance expression in Part 4 to help understand the variation in the estimator $\hat{\bar{p}}(\gamma,d)$. In essence, the variance expression reflects the block assignment nature of the DGPs to the participants. When $M = 1$, each DGP is only assigned to a single participant, the R_i's are i.i.d., and $\hat{\bar{p}}(\gamma,d)$ simply follows a (scaled) binomial distribution with variance $\bar{p}(\gamma,d) - \bar{p}(\gamma,d)^2$. When $M > 1$, each DGP is assigned to a block of participants, the R_i's are no longer independent within the same block, and the distribution of $\hat{\bar{p}}(\gamma,d)$ generally deviates from a (scaled) binomial. When M is large, we can precisely estimate $p(\gamma,g,d)$ for each DGP g, and the variance of $\hat{\bar{p}}(\gamma,d)$ reduces to that of $p(\gamma,g,d)$ over the space of DGPs.

In the actual analysis of our experimental data, we use the normal approximation of the binomial random variable $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ when $M \cdot J$ is large. Based on a result from Hoeffding (1956) and summarized by Percus and Percus (1985), using $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ leads to conservative inference on the Poisson binomial random variable $M \cdot J \cdot \hat{\bar{p}}(\gamma,d)$ when conditioning on $\{g_j\}_{j=1}^{J}$.[2] Therefore, conditioning on the set of selected DGPs, using the normal distribution $N\left(\hat{\bar{p}}(\gamma,d), \hat{\bar{p}}(\gamma,d)(1-\hat{\bar{p}}(\gamma,d))/(M \cdot J)\right)$ provides (approximately) conservative inference for $\hat{\bar{p}}(\gamma,d)$ and only relies on the product $M \cdot J$ being large, as opposed to J being large by itself.

[2]For a given (γ,d), inference for $\hat{\bar{p}}(\gamma,d)$ using $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ is exact when $p(\gamma,g_j,d)$ is the same for all j.

B.1.2 Calonico, Cattaneo, and Titiunik (2015) Bin Selection Algorithms

This section describes the two bin width algorithms from Calonico, Cattaneo, and Titiunik (2015): the IMSE algorithm, which minimizes the integrated mean squared error of the bin-average estimators of the CEF and results in fewer, larger bins, and the MV algorithm, which aims to approximate the variability of the underlying data and results in more, smaller bins.

Formally, consider without loss of generality the number of bins above the threshold. We denote the relevant quantities with a "+" subscript. The IMSE bin selector chooses the number of bins $J_{+,n}$ to minimize the integrated mean squared error

$$\int_0^{\bar{x}} E[(\hat{\mu}_+(x) - \mu_+(x))^2 | X = x] f(x) dx = \frac{J_{+,n}}{n} Var_+\{1 + o_p(1)\} + \frac{1}{J_{+,n}^2} Bias_+\{1 + o_p(1)\}$$

where μ_+ denotes the CEF, $\hat{\mu}_+$ denotes the bin-average estimator of μ_+, $f(x)$ is the density of X, n is the overall sample size, and the constants are equal to

$$Var_+ = \frac{1}{\bar{x}} \int_0^{\bar{x}} \sigma_+^2(x) dx$$

$$Bias_+ = \frac{\bar{x}^2}{12} \int_0^{\bar{x}} (\mu_+'(x))^2 f(x) dx,$$

with $\bar{x}$ being the upper bound of the support of X, μ_+' the first derivative of μ_+, and $\sigma_+^2(x)$ the conditional variance of Y given $X = x$.

Solving the minimization problem, we obtain the IMSE-optimal number of bins

$$J_{+,n,IMSE} = \left\lceil \left(\frac{2Bias_+}{Var_+} \right)^{\frac{1}{3}} n^{\frac{1}{3}} \right\rceil$$

with an analogous $J_{-,n,IMSE}$ for the number of bins below the threshold.

The goal of the mimicking variance bin selector, on the other hand, is to "choose the number of bins so that the binned sample means have an asymptotic (integrated) variability approximately equal to the amount of variability of the raw data" (Calonico, Cattaneo, and Titiunik, 2015). Consider again the problem of choosing the number of bins above the threshold. The MV criterion translates to setting

$$\frac{J_{+,n}}{n} Var_+ = Var(Y|X \geqslant 0) \equiv V_+.$$

It follows that the number of bins should be

$$J_{+,n} = \frac{V_+}{Var_+} n.$$

However, this number of bins grows too quickly for the rate conditions in Calonico, Cattaneo, and Titiunik (2015) that ensure the consistency of $\hat{\mu}_+(x)$. Instead, the authors modify the formula slightly and use the following bin number

$$J_{+,n,MV} = \left\lceil \frac{V_+}{Var_+} \frac{n}{\log(n)^2} \right\rceil.$$

B.1.3 Mean Squared Error Decompositions

For each meta-graph (γ, g, d), let $d_i(\gamma, g, d, \varepsilon_i)$ be the discontinuity estimate by participant i. The "human estimate" of the discontinuity for this meta-graph is then given by the average $\hat{d}(\gamma, g, d) = \frac{1}{M}\sum_{i=1}^{M} d_i(\gamma, g, d, \varepsilon_i)$. Meanwhile, let $d_{\hat{\theta}}(g, d, \varepsilon_i)$ and $\hat{d}_{\hat{\theta}}(g, d) = \frac{1}{M}\sum_{i=1}^{M} d_{\hat{\theta}}(g, d, \varepsilon_i)$ be the corresponding estimates given by the econometric estimator $\hat{\theta}$. In order to better compare the performance of the humans versus the econometric estimator, we further decompose the mean squared error (MSE) of the above estimates into the

square of the bias and the variance, by using the following identities respectively:

$$\frac{1}{M}\sum_{i=1}^{M}(d_i(\gamma,g,d,\varepsilon_i)-d)^2=(\hat{d}(\gamma,g,d)-d)^2+\frac{1}{M}\sum_{i=1}^{M}(d_i(\gamma,g,d,\varepsilon_i)-\hat{d}(\gamma,g,d))^2 \quad \text{(B.1)}$$

$$\frac{1}{M}\sum_{i=1}^{M}(d_{\hat{\theta}}(g,d,\varepsilon_i)-d)^2=(\hat{d}_{\hat{\theta}}(g,d)-d)^2+\frac{1}{M}\sum_{i=1}^{M}(d_{\hat{\theta}}(g,d,\varepsilon_i)-\hat{d}_{\hat{\theta}}(g,d))^2. \quad \text{(B.2)}$$

Note that the interpretation of the quantities in the first identity differs slightly from the conventional MSE decomposition, as the randomness here not only comes from the realization (ε_i), but also from the participant (d_i). Nevertheless, the second term can still be conveniently understood as the variance, as $d_i(\gamma,g,d,\varepsilon_i)$ remains independent and identically distributed for different i. In Figure B20, we plot the square root of the expression (B.1) and the expressions (B.2) for various $\hat{\theta}$. We also decompose the difference between the left-hand sides of (B.1) and (B.2) into differences in bias and variance for the case where $\hat{\theta}$ is the piecewise quintic estimator.

B.2 Additional Details on DGP Specification and Graph Creation

In conjunction with Section 2.3.2 in the main text, this appendix describes the process by which we specify DGPs and create graphs for our experiments.

For RDD, we identify 110 empirical papers drawn from top economics journals including the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal of Business and Economic Statistics*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics and Statistics.* We then randomly sample 11 papers from this list that have replication data available.

For the specification of the running variable for each DGP, we directly use the empirical distribution of X from the corresponding paper and normalize the running variable to lie in $[-1,1]$. If the support of the running variable is asymmetric, this normalization may create

a discontinuity in the density at $x = 0$, which violates the assumptions in the identification framework of Lee (2008) (but not the minimally sufficient identifying assumptions of Hahn, Todd, and Klaauw, 1999). The original running variable in seven of our 11 papers fails the McCrary (2008) test for continuity of the running variable density. Among our normalized running variables, eight fail the test, with two passing only in their non-normalized version and one passing only after normalizing. In our testing, visual inference performance does not vary significantly across DGPs whether the support of their original (unnormalized) running variables is symmetric. Following Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014), we remove any observations where $|X| > 0.99$.

Next, we use the resulting datasets and further follow Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) to specify the DGP's CEF via a piecewise global quintic regression. Denoting the left and right intercepts of the CEFs by α^- and α^+, we shift the right arm of the CEF upwards/downwards by the amount $|\alpha^+ - \alpha^-|$ to make the functions $E[\tilde{Y}|X = x]$ continuous at the policy threshold.

Our specification of the distribution of the error term u as i.i.d. normal similarly follows Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014). u has mean zero and standard deviation σ, which is specified as the root mean squared error of the global quintic regression in the CEF specification above. We generate eight draws of u for each DGP, enough so that every graph we generate for our experiments is seen by no more than one participant. Most of our RD papers use a continuous running variable, but two feature semi-discrete variables where the running variable takes on many unique values but these points have multiple observations each. In these two cases, we add small amounts of noise from a $N(0, (\frac{1}{\min\{N_-, N_+\}})^2)$ distribution, where N_- and N_+ are the number of observations below and above the policy threshold, to match the continuous running variable condition assumed in Calonico, Cattaneo, and Titiunik (2015). We add this noise prior to fitting the piecewise quintic CEF. Perturbing the data prior to fitting the quintics introduces "measurement error" that can attenuate estimated discontinuities. Alternatives

include fitting the CEF prior to adding the noise or drawing the noise from a uniform distribution to prevent observations from crossing the policy threshold. For a thorough discussion of measurement error in RD, see Pei and Shen (2017).

In practice, the difference between adding the noise before and after fitting the CEF is minor. Figure B22 presents lineup protocols for these two DGPs. To test the null hypothesis that the DGP resulting from fitting the piecewise quintic prior to adding noise is indistinguishable from the DGP where noise is added first, we randomly place one graph from the former distribution among 19 from the latter. All graphs use a continuous running variable generated by adding noise to the original semi-discrete one. The solutions, i.e. the graphs from the DGP fit on the original data, are in the footnote at the end of this sentence.[3] We present both sets of CEFs in Figure B23, with the scale of the y-axis chosen to match that of the corresponding graph in the original paper. Although the differences in the tails for the CEF on the right of Figure B23 underscore the sensitivity of the piecewise quintic specifications, the difficulty of these lineup protocols and the general similarity of the CEFs suggest that the infusion of measurement error is not a major concern.

Specifying the RKD DGPs requires several modifications to our process. As mentioned in Section 2.3.2, we generate the running variable by sampling from an estimated kernel density due to data confidentiality issues. We again normalize the support of the running variable to $[-1,1]$ with observations where $|X| > 0.99$ trimmed. Because the X variable in all of the RKD applications is income, its distribution is skewed and its support quite asymmetric around 0. Therefore, we select the largest subsample such that the support of X therein is symmetric around 0.

While one can create arbitrary discontinuities that preserve the shape of the CEF just by vertically shifting one of the DGP's "arms," kinks are not as simple. How do we create a zero-kink graph from the seven kinked CEFs in Figure B7? (The four global quintic CEFs already have no kink.) We decide to rotate the right arm of the CEF until the left and right

[3]The solution for the left lineup is row $5 \cdot 2 - 6$ and column $2^2 - 1$. The solution for the right lineup is row $-3 \cdot 2 + 10$ and column $9/3 - 2$.

derivatives match. Following that logic, we create graphs with kinks by rotating the right arm from the zero-kink graph at nine different levels. As explained in Section 2.3.2, we create the four smooth global quintic DGPs out of concern that only using the seven piecewise quintic DGPs may be too favorable to visual inference. Specifically, several of the zero-kink CEFs of the piecewise quintic DGPs obtained by rotating their right arms appear similar to a linear function, at least locally around $x = 0$. In comparison, the curvature in the global quintic DGPs, especially in combination with the function's peaking at $x = 0$, is likely much harder to distinguish from a kink for visual inference.

Because we generate the RK graphs by rotations, we express the kink magnitude in terms of the angle between the left and right arms of the CEF. A kink is defined as the difference between the right derivative (denoted by κ^+) and the left derivative (denoted by κ^-), $\kappa^+ - \kappa^-$, and the corresponding angle is simply $\psi \equiv \arctan(\kappa^+) - \arctan(\kappa^-)$. Our nine rotations are $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$ radians, where ξ is the asymptotic standard deviation of the angle estimator between the left and right arms $\hat{\psi} \equiv \arctan(\hat{\kappa}^+) - \arctan(\hat{\kappa}^-)$ and $\hat{\kappa}^+$ and $\hat{\kappa}^-$ are estimators of κ^+ and κ^- from a piecewise quintic regression on the data used to calibrate the CEF. A complication of this rotation-based approach is that the right arm of a function can be rotated only a certain amount before it bends back over itself and ceases to be a function. For each RK DGP, we calculate the maximum angle such that its rotated CEF remains a function, and this maximum prevents us from specifying the largest kink magnitude as $\pm 1.5\xi$ for some of the DGPs. While the maximum kink magnitude in our RK DGPs is $\pm 0.9\xi$, we cannot create kinks that large for all of the DGPs. We replace two meta-graphs that would go too far, $+0.54\xi$ and $+0.9\xi$ for Card, Lee, and Pei (2009), with their non-offending negative counterpart rotation.

Specifying the distribution of the error term u is more complex in RKD for two reasons. First, adding u prior to rotation may result in points changing order along the x-dimension post-rotation, exacerbating the maximum rotation problem. Second, unlike the vertical shift technique used in creating RD datasets and graphs of different discontinuity levels, data

rotation changes the standard deviation of the error term. To circumvent these problems, we rotate the right arm of the CEF first and then add the error term to generate the outcome variable. The standard deviation of u depends on the degree of rotation and is calculated such that ξ, the asymptotic standard deviation of the angle estimator $\hat{\psi}$ in the original fitted CEF, remains the same after rotation. Like in the RD case, we use eight draws of u for each DGP.

B.3 Design of Experiments and Studies

This section outlines the structure of the RDD and RKD experiments. We first describe the general sequence of events and survey structure before discussing the sequential rollouts of the RDD and RKD experiments. All parts of this study were approved by the Institutional Review Boards (IRBs) of Cornell University, Princeton, Columbia Teachers College, and the University of Waterloo and were pre-registered on Open Science Framework and at the American Economic Association's RCT Registry. Expert studies run as part of seminars were additionally approved by their respective local IRBs. Before beginning any of the studies, participants completed consent forms. Participants had the opportunity to exit the study at any time.

B.3.1 Non-Expert Experiment

Each experiment consisted of three parts. In Part 1, participants watched a video tutorial on RDD (RKD) outlining the graph construction (normalization of the running variable, binning the data, etc.) and stating the objective as classifying a discontinuity (kink) in the "true" underlying relationship at $x = 0$. The video tutorial lasted between three to four minutes depending on the participant's graphical representation treatment. For example, participants in treatment arms featuring fit lines received a brief supplementary segment informing them that fit lines only serve as approximations of the true underlying relationship

between the running variable and the outcome variable. To ensure participant engagement, the video tutorial featured an attention check about halfway through the video. This took the form of a colored bird (see example in Figure B24); the voiceover on this slide informed participants that when asked about the color of our "bird of interest," they should report a specific color that did not match the color of the bird in the picture. For example, participants seeing Figure B24 below might be asked to report that the bird is red, not blue. Participants had the opportunity to watch the video several times or to pause and rewatch specific sections as desired. However, participants were not able to return to the video tutorial once they had moved on to the next section.

The video tutorial was followed by a sequence of example questions. Participants were asked to classify tutorial graphs and received feedback on their answers. Participants could navigate between these graphs using a panel of buttons on the right-hand side of the screen (see Figures B25 and B26).

Part 2 of the experiment consisted of a series of paid classification tasks. Participants were asked to assess whether a given graph featured a discontinuity/kink or not and could choose between two potential bonus options for each graph. We discuss this bonus system in detail in the next subsection. Figure B27 shows a typical classification screen. Participants were not given any feedback regarding their accuracy between tasks.

Part 3 of the experiment was an exit survey. We solicited basic sample demographics such as age, gender, occupation, education level, and prior experience with statistics. For all phases after the initial pilot, we also asked participants to report whether they had participated in earlier iterations of the study. At the end of the experiment, participants were informed of the total number of graphs they classified correctly as well as their corresponding bonus earnings (see Figure B28 for an example). Experiment participants were paid in Amazon gift cards after each phase was complete.

Analysis of Bonus Payments

Participants in the experiment had the option to choose between earning a bonus of 40 cents if their discontinuity/kink classification was correct and nothing otherwise, or to earn a bonus of 20 cents irrespective of their classification. Note that participants will choose a given answer if their perceived probability that the answer is correct is at least 50%. Denoting the utility function by u, the expected utility under the wager for a perceived probability of $p \equiv \mathbb{P}(correct)$ is given by $pu(\$0.40) + (1-p)u(\$0)$, which is equal to $pu(\$0.40)$ if we normalize $u(0)$ to zero. Assuming expected utility maximization, participants will choose the sure amount whenever $pu(\$0.40) \leq u(\$0.20)$, i.e. whenever $p \in [0.5, u(\$0.20)/u(\$0.40)]$. Further assuming that subjects are approximately risk neutral at small stakes, participants would only choose the sure amount whenever they believe their chance of being correct to be a coin flip (because $u(\$0.20)/u(\$0.40)$ is approximately equal to 0.5 in this case).

If we instead allow for loss aversion, participants prefer the sure amount under a wider range of (perceived) success probabilities. Under loss aversion, the structure of the model is similar except that the utility function is replaced with a value function for gains and losses with respect to a reference point (Kahneman and Tversky, 1979). Prospect theory is based on the assumption that losses loom larger than gains, which is commonly operationalized through a loss-aversion parameter $\lambda > 1$ multiplying outcomes in the loss domain. Assuming that participants view the sure amount of 20 cents as their reference point, prospect theory suggests that they would consider the gamble as an opportunity to win 20 cents or lose 20 cents depending on their answer. The expected payoff of the wager in this case is given by $pu(\$0.40 - \$0.20) - \lambda(1-p)u(\$0.20 - \$0) = u(\$0.20)(p - \lambda(1-p))$, which would be compared to a certain outcome of $u(\$0.20 - \$0.20) = 0$ under the fixed payment option. Participants would therefore choose the wager whenever $p/(1-p) \geq \lambda$. Estimates for λ commonly range between 1.5 and 2.5, though significant heterogeneity may exist with respect to this preference parameter. See Chapman et al. (2018) for a review.

RDD

The rollout of the RDD experiments was staggered across four phases which were completed between November 2018 and October 2019. The first phase, phase RDD1, served as a pilot and was not part of the pre-registration package. The goal of this pilot was to gauge effect sizes in the case of the bin width and axis scaling treatments in order to determine the required sample size for subsequent (pre-registered) phases. Table B2 outlines the graphical representation treatments implemented in each phase. Each participant's classification task featured 11 graphs, one from each of the 11 specified DGPs. The order of graphs was randomized within subjects. The discontinuity levels were chosen as follows: two graphs featured no discontinuity, one graph featured the extreme discontinuity levels of $\pm 1.5\sigma$, and the remaining eight graphs were based on each of the eight discontinuity levels $\pm 0.1944\sigma, \pm 0.324\sigma, \pm 0.54\sigma, \pm 0.9\sigma$. In total, each treatment arm contained 968 unique graphs split across 88 unique participants. Figure B29 outlines the sequence of events for the RDD experiments.

RKD

The rollout of the RKD experiments consisted of a pilot followed by a single experiment, both completed between April and June of 2019. Table B3 outlines the graphical representation treatments implemented in each phase. Each classification task was split into two distinct sections. One section corresponded to RKD graphs presented in levels and required participants to assess the presence of a kink. The other section instead plotted estimated derivatives and required participants to assess the presence of a discontinuity. The order of the two sections was randomized between participants.

In the RKD pilot, all graphs were based on 7 piecewise quintic DGPs. Participants saw one graph from each DGP in each of the two sections, for a total of 14 graphs. In the final experiment (Phase RKD), we added four additional global quintic DGPs that featured smooth peaks instead of kinks at $x = 0$ and asked participants to classify a total of

22 graphs (11 graphs in each section). The test kink levels (measured in radians) were $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$, for a total of nine possible levels. The pilot featured 504 unique graphs for each treatment arm split across 72 unique surveys, and Phase RKD featured 792 unique graphs per treatment arm split across 72 unique surveys. Figure B30 outlines the sequence of events for the RKD experiments.

B.3.2 Expert Study

The expert study was run both online using members of the National Bureau of Economic Research in applied microeconomic fields (aging, children, development, education, health, health care, industrial organization, labor, and public) and members and affiliates of the Institute of Labor Economics (IZA), and during three seminar presentations between May and October of 2019. Table B11 shows the timeline and number of participants for each session. Figure B31 outlines the sequence of events for the expert study.

Unlike in the non-expert experiments, participants in the expert study were not asked to watch a video tutorial or complete practice questions at the start. Instead, Part 1 of the expert study was a classification task paralleling the one non-expert participants completed (see Figure B33). In addition, the graphical representation choices in this study were not randomized between subjects. Participants had the opportunity to navigate between the 11 graph classifications using a panel of navigation buttons at the bottom of the screen. This feature was introduced to address feedback we received during the testing of the survey. Figure B33 shows an example of the task screens for participants in the RD and RK treatments.

Part 2 of the expert study asked participants to rank four alternative graphical representation options for RDD or two for RKD. The four treatment combinations in the RD case corresponded to small or large bins interacted with both fit line treatments, as in Phase RDD4. The two treatment combinations in the RK case were small and large bins. The instructions for Part 2 of the expert survey are outlined in Figures B34 and B35. Participants

were asked to specify which graphing treatment they believe researchers should use to present evidence of the main treatment effect of a study, as well as which graphical options they believe would perform best and worst in our non-expert sample. Participants were also able to report having "no preference" across the different graphing options. These three decisions were repeated three times in the context of a zero, small, and large discontinuity or kink.

Part 3 of the expert study was an exit survey. We asked about basic sample demographics such as age, gender, geographic region, main area of research, and prior experience with RD/RK. After all sessions of the expert study were complete, we randomly selected four participants to receive a participation fee of $450 plus $50 for every correctly classified graph in Part 1 of the study. We did not take the accuracy of the discontinuity magnitude estimate into account to determine payments in the RD case, only the binary classification decision.

Table B11 highlights the different participant pools for the sessions of the expert study and lists the graphical representation choices in the classification task (Part 1 of the expert study). All graphs in Part 1 used the default Stata axis scaling with no fit lines and a vertical line at $x = 0$. All but one session of the expert study used small bins for the RD case. The exception was the session run at Princeton's Quantitative Social Science Colloquium in October 2019 in which we used large bins to compare the effect of bin widths on experts and non-experts.

The distribution of graphs seen by the experts and non-experts is slightly different due to the methods used to randomize participants' graphs. The non-expert experiment's randomization is based around the order in which the 88 participants for each treatment arm begin the survey. We use the same randomization mechanism for experts, but repeating the survey in different seminars and online phases results in more experts at the "beginning" of the graph assignments. Reweighting the data to match the distribution of graphs seen by non-experts does not meaningfully change any of our results.

B.3.3 Eyetracking Study

We chose the 11 graphs for the eyetracking study participants to classify based on information available when designing it. Specifically, we had data from the first three phases of experiments on RD graphical parameters as well as expert responses from two seminars. We saw the experts in these early samples attain both lower type I and type II error rates than the non-experts, and we chose the discontinuity-DGP pairs based on the maximum difference in classification accuracy between experts and non-experts. The final set of 11 graphs features four graphs with no discontinuity, three graphs with a discontinuity magnitude of 0.1944σ, two graphs with a discontinuity magnitude of 0.324σ, and two graphs with a discontinuity magnitude of 0.54σ. Like those shown to most of the participants in the expert study, all graphs have equally spaced small bins and default Stata axis scaling with a vertical line indicating the policy threshold and no fit lines.

The eyetracking study consisted of three parts for non-experts and two for experts (see Figure B36 for details). For non-experts, Part 1 was the same video- and example-based tutorial used in the RDD experiments. To avoid eye fatigue, this tutorial was completed on a regular computer. After completing the tutorial, participants were moved to a Tobii XL60 eyetracker to complete Part 2, which consisted of 11 graph classification tasks. Each task was broken up into three separate screens as illustrated in Figure B37: a white screen with a cross around the perimeter, a screen showing the graph of interest, and a screen asking participants to report whether or not they believe the graph featured a discontinuity. Presenting a white screen with a cross avoids "central fixation bias": by diverting participants' gaze towards the perimeter of the screen before showing a graph, we can rule out that time spent looking at the cutoff is simply an artifact of people's tendency to fixate their gaze at the center of the screen when a new stimulus appears. When looking at the graph of interest, participants could spend as much time as desired before proceeding to the next screen to answer the classification question. Participants could advance between pages by

pressing the space bar in order to avoid having participants look down at the keyboard or having them use a button on the screen, which might confound their viewing patterns. Part 3 consisted of verbal feedback on the number of correct classifications and the corresponding bonus payment. Experts moved directly to the classification task at the eyetracking terminal without having to complete the tutorial.

To minimize distractions, participants completed the study one-by-one. Participants were paid in Amazon gift cards after each session.

B.4 Subject Characteristics and Performance Predictors

B.4.1 Demographic Balance and Effects

For our non-expert experiment, we check the validity of our randomization by testing for the balance of covariates across treatment groups. For each phase, we regress the covariates on treatment group indicators and test whether all coefficients are equal. Specifically, we test for the balance of sex, education, age categories, statistical knowledge, passing the attention check, being a first-time participant, and not completing the experiment. The results are in Tables B12 through B15. Across these 44 hypotheses, we cannot reject balance in all but two instances (4.5% of tests) when using a 5%-level test: college completion and fraction of participants aged 23-49 in Phase RDD4. In addition, p-values for joint tests of significance for all covariates within each phase based on Pei, Pischke, and Schwandt (2019) are 0.18, 0.50, 0.65, and 0.14 for Phases RDD1-4 and 0.54 for Phase RKD. The evidence is consistent with successful randomization.

As can be seen in Tables B17 and B18, participant demographics such as gender, age, college completion, statistical knowledge, and being a repeat participant are generally not predictive of performance in the experiment for both discontinuities and kinks, with insignificant point estimates of a few percentage points. Out of the 34 factors we test, two have a statistically significant effect on the probability of classifying a graph correctly

(5.9% of tests). In Phase RDD1, failing the attention check is associated with a 5.5pp decrease in the probability of classifying a graph correctly. This effect is smaller and insignificant in Phases RDD2 (4.3pp), RDD3 (-1.2pp), and RDD4 (-1.2pp). In Phase RDD3, being 50 or older correlates with a 9.4pp increase in the probability of classifying a graph correctly. This lack of predictive strength makes any imbalances in demographics across phases an even smaller concern. It also suggests that comparisons of treatments across phases are unlikely to be confounded by differing subject pools.

B.4.2 RD DGP Characteristics

While our results in Section 2.4 focus on the treatment effects of each graphical parameter, in this section we focus on other factors of inference performance. In particular, we explore how the characteristics of an RD DGP affect visual inference.

Figure B38 plots the density of each RD DGP along its running variable. These distributions are varied. Some are approximately uniform, some are more normal, and others, such as DGPs 6 and 8, are highly skewed. Some DGPs show clear changes in density at the policy threshold, some of which were present in the original data and some of which were introduced by having asymmetric supports prior to our rescaling the running variable. Figure B39 compares power functions for the eight DGPs with symmetric supports prior to normalization with the three with asymmetric supports in Phase RDD4. There is a noticeable flatness or dip in the share of participants reporting a discontinuity between 0 and 0.1944σ. This pattern appears in every RD phase, but with only three asymmetric DGPs, this may also be due to the effects of the individual DGPs rather than something about the effects of normalizing the running variable when asymmetry is present.

When the distribution of the running variable is uniform, there is no difference between graphs generated with even spacing and quantile spacing. When the distribution is not uniform, we may expect quantile spacing to perform better, for example by preventing outlier bins with very few observations in more sparsely populated regions of the support. To

quantify a DGP's deviation from a uniform distribution, we can compute Gini coefficients based on the distribution of observations within evenly spaced bins.[4] We calculate Gini coefficients for both the IMSE and MV bin width algorithms and take the average across ε for each RD DGP. Figure B40 explores how the Gini coefficient interacts with bin spacing to test whether quantile spacing outperforms equal spacing when the distribution is more skewed by averaging the probability of a correct response across treatment arms within each DGP. There is no relationship between a DGP's Gini coefficient and performance across spacing types. Among DGPs with higher distributional inequality, as well as for those with intermediate and lower values, neither spacing dominates the other.

In Table B19, we examine how the first derivatives of each arm of the DGP interact with the direction of the discontinuity in impacting power. Specifically, we ask whether visual inference performs better if the discontinuity is positive or negative within a combination of left and right slope signs. Looking only at graphs with nonzero discontinuities, with which we measure power, we regress the binary classification decision on dummy variables for the negative discontinuity sign within each slopes combination (omitting the positive discontinuity dummy) with DGP fixed effects and clustering at the participant level. These regressions show that visual inference performs better when the discontinuity direction matches the signs of both the left and right derivatives. For example, when the left and right first derivatives are both positive, power is 11 percentage points worse when the discontinuity is negative than positive (see Figure B41 for an illustration). Table B20 repeats a similar exercise for size over the no-discontinuity graphs. We cannot include DGP fixed effects here, as they would eliminate the variation needed to identify impacts. Unlike with power, we do not find any meaningful results for size.

B.4.3 Dynamic Visual Inference

In this subsection, we investigate participants' performance over the course of the 11 graphs

[4] We do not compute Gini coefficients with quantile spacing, as all bins have about the same number of observations and the Gini coefficients would all be approximately zero.

they see. This exercise provides evidence to alleviate the concern that participants have a preconceived notion that about half of the graphs feature a discontinuity, which may bias our results in favor of good size control by RD visual inference.

First, we trace out the fraction of participants reporting a discontinuity over their 11 graphs. If there is a consistent upward or downward trend, it would be indicative of participants having a fixed number of continuous graphs in mind. As shown in Figure B42, however, the trends are flat in the treatment arm (small bins and no fit lines) used to compare visual and econometric inference procedures over the four RDD non-expert phases and in the expert sample.[5]

We also implement an AR(1) regression to more closely examine the dynamic discontinuity classification over the 11 graphs, which Figure B42 may miss:

$$R_{is} = \rho R_{i(s-1)} + c_i + u_{is}. \tag{B.3}$$

R_{is} is the binary variable indicating whether participant i reports a discontinuity in graph s, c_i is the individual fixed effect, and u_{is} is the population residual from the projection of R_{is} on $R_{i(s-1)}$ and c_i. We expect a negative ρ if participants believe ex-ante that half of the graphs will feature a discontinuity. In fact, we do see negative estimates of ρ ranging between -0.1 and -0.2 as reported in column (1) of Table B21. However, these estimates are subject to a large-N asymptotic bias of order $O(1/S)$ per Nickell (1981) where S is the number of graphs. Inverting equation (18) from Nickell (1981), we solve for the bias-corrected estimates of ρ and report them in column (2) of Table B21. Although these estimates are still negative, they are considerably closer to zero. For the four phases of the non-expert Phases RDD1-4, three bias-corrected estimates are statistically insignificant at the 5% level, while the other one is marginally significant.

[5]Because we randomize the graph order, the flat discontinuity classification rate is consistent with a flat correct classification rate, which we also verify in the data. That is, no evidence supports "learning" over the 11 graphs in terms of visual inference success, which is concordant with our design in which participants find out the total number of correct discontinuity classifications only after they complete the study.

Although the estimate of ρ in the expert sample is still significant after bias correction, the fact that it is the largest in magnitude of the five estimates in Table B21 is reassuring. Because non-experts appear to exhibit better size control than experts and are less experienced with RDD graphs, we may be more concerned that the non-experts' performance is driven by having a fixed number of continuous graphs in mind. However, this is refuted by the four estimates of ρ from the non-expert sample being smaller. Therefore, we conclude that the evidence does not support the hypothesis that participants expect half of the graphs to be continuous, and the good size control of RD visual inference is not a result of it.

B.5 What *t*-Statistic Can Eyes Detect?

Using our experimental results, we can provide an evidence-based answer to the question of how large a t-statistic needs to be in order for readers to visually detect a discontinuity (a question that was also posed by Kirabo Jackson on Twitter). We do so by using an alternative scaling of the discontinuity. Instead of specifying the discontinuity d as a multiple of the error standard deviation σ as in Section 2.4, we specify the x-axis of the power functions in Figures B43 and B44 as $d/\sigma\sqrt{(X'X)^{-1}_{dd}}$, where X is a regressor matrix containing the 12 polynomial terms of the running variable for a piecewise quintic regression and the dd subscript denotes the entry of the matrix corresponding to the discontinuity estimator. Because the denominator is the large-sample error of the discontinuity estimator from a (correctly specified) piecewise quintic regression, the x-axes of these figures have a t-statistic interpretation. Note that because the rescaling results in a distinct set of six discontinuity magnitudes for each of our 11 DGPs, we now have 66 discontinuity magnitudes. Therefore, we bin points along the x-axis, and consequently, these power functions are no longer monotone, as the composition of DGPs changes across bin points.

Our analysis shows that the power of visual inference approaches 0.8, i.e. 80% of the participants correctly report a discontinuity, across all phases of our RD experiment as the t-

statistic from the correctly specified regression exceeds 4. There are caveats to this already nuanced answer. First, this threshold t-statistic is specific to the range of parameters we test. For example, one could substantially increase the y-axis scale, which would be sure to make any discontinuity we test in this paper disappear before the human eye. Second, as we see in the RK results, this threshold is specific for detecting discontinuities, and a much smaller threshold applies to kink detection.

One might conjecture that the t-statistic threshold is sensitive to the sample size. In particular, it seems possible to keep the plot fixed and increase the sample size arbitrarily by adding observations in each bin, which would increase the t-statistic without altering visual inference. The pitfall of this argument is that for the plot to look the same while increasing sample size, one needs to change σ as well. In fact, σ needs to increase proportionally to $\sqrt{n}$, leading to an invariant t-statistic.

B.6 Combining Visual and Econometric Inferences

In section 2.4.4, we present "marginal" power functions for visual and econometric inference procedures. We now discuss the joint distribution of visual and econometric discontinuity/kink tests. Knowing the joint distribution i) informs us of their complementarity, i.e. whether visual and econometric inferences tend to be wrong on different data, and ii) is useful for combining inferences.

We begin by considering empirical results on i). From the data that underlie the power function comparisons in Figures 2.4 and 2.5, we produce a series of two-by-two contingency tables for the binary visual inference decision R and the binary econometric inference decision based on an estimator $\hat{\theta}$, $R_{\hat{\theta}}$. More specifically, for each normalized discontinuity or kink magnitude $|d|$ ($|d|$ takes on the values 0, 0.1944, 0.324, etc. per the design of our experiment), we characterize the joint distribution of R and $R_{\hat{\theta}}$ over the set of graphs presented to the expert participants. We conduct Fisher's exact test for independence and

present one-sided p-values in Table B22 by estimator $\hat{\theta}$ and discontinuity/kink magnitude $|d|$.[6]

Several patterns emerge from Table B22. For RDD (Panel A), the first row shows no strong support for an association between econometric and visual inference decisions when the true discontinuity is zero. But when the true discontinuity is nonzero (rows two to six), there appears to be strong evidence in support of association (though not reported in Table B22, all correlations are positive in these cases). In other words, type II errors by experts are predictive of type II errors by various econometric inference methods, but this is not true for type I errors. For RKD (Panel B), the evidence for an association is much weaker. However, consistent with our interpretation of Figure 2.5, the p-values generally become smaller and the evidence of positive association becomes stronger when we move from the default `rdrobust` inference (labeled CCT1) to conventional local linear inference with a relatively large bandwidth (labeled CCT4).

To addresses point ii), we provide a conceptual discussion underscoring the possibility of combining visual and econometric inference. That inferences can complement one another is not unique to visual inference, and it harkens back to the introduction of Fisher's method (Fisher, 1925 and Mosteller and Fisher, 1948) which combines independent statistical tests by summing the logarithms of their p-values. Although we do not explicitly ask our participants to state a p-value, we can impute it from their visual inference decisions. For example, we can follow Habiger and Peña (2011): for each $R_i = 1$ when seeing a graph that is generated with parameter γ and discontinuity d, we uniformly randomly draw a p-value, p_i, from the interval $[0, \hat{p}(\gamma, d)]$; for each $R_i = 0$, draw p_i from $[\hat{p}(\gamma, d), 1]$. We can then combine the individual p-values along with that from econometric inference via the

[6]Because the maximum of the standard correlation measure depends on the marginal distributions of R and $R_{\hat{\theta}}$ (for example, if $\Pr(R = 1) \neq \Pr(R_{\hat{\theta}} = 1)$, then their correlation cannot be 1), which vary across $|d|$ and $\hat{\theta}$, correlation is difficult to interpret. Although we can follow the spirit of the proposals by Cohen (1960) and Davenport and El-Sanhurry (1991) and present correlations scaled relative to their maximum as determined by the marginal distributions, we choose not to do so for brevity. Instead, we only show one-sided p-values from Fisher's exact test—two-sided p-values are method-dependent because of the ambiguity in classifying contingency tables as extreme in the opposite direction (see for example Agresti, 1992).

proposed method by Wilson (2019), which allows the p-values to be dependent unlike in Fisher's method.

To see that using the joint distribution of R and $R_{\hat{\theta}}$ can refine the inference on the discontinuity/kink d, it is easiest to use a Bayesian approach for illustration. Suppose researchers have an ex-ante sense of the discontinuity magnitude in an empirical application based on their read of the literature, which is described by prior density $\psi(\cdot)$. They then analyze the data and conduct visual and econometric inferences and reject no-discontinuity in both cases, i.e. $R = 1$ and $R_{\hat{\theta}} = 1$. The posterior probability that the discontinuity magnitude falls in a given set A is

$$\frac{\int_{z \in A} \Pr(R = 1, R_{\hat{\theta}} = 1|z) d\psi(z)}{\int_{z} \Pr(R = 1, R_{\hat{\theta}} = 1|z) d\psi(z)}.$$

When R and $R_{\hat{\theta}}$ always agree, i.e. they are perfectly correlated, knowing R does not change the posterior. As the association between R and $R_{\hat{\theta}}$ lessens, knowing R provides more information.

Although this is a conceptual demonstration, it does illustrate—through both frequentist and Bayesian lenses—that combining visual and econometric inference can be useful. This is true especially when the two inference decisions are not strongly associated, which is generally the case for RKD and when there is no discontinuity in RDD as shown in Section B.6.

B.7 Eyetracking Study: Details and Additional Analyses

As outlined in Section 2.3.5 and Appendix B.3.3, the eyetracking study consists of a fixed sequence of 11 graphs which we show to both expert and non-expert participants on a Tobii XL60 eyetracking terminal that records the location of the subject's gaze every 17 milliseconds (60 times per second).

Figure B45 is a visual representation of the eyetracking microdata from one participant

looking at a particular RD graph. It plots a sequence of fixations in connected circles against the RD graph seen by the participant. The number inside each circle represents the order in which the fixations took place. In total, this participant fixated on 22 points in the graph over approximately 9 seconds. The time lengths vary across the fixations and are not represented in the figure. The initial fixation occurred at the upper right corner of this particular graph for every participant, which was induced by the placement of the cross on the previous screen (see Figure B37). We therefore remove the first two fixations from the microdata for every participant-graph combination. We also remove all saccades before the third fixation. Further, we remove all eye gazes that lie outside the outermost axis ticks because they mostly coincide with axis labels.

In the resulting data set, we transform the eyegaze pixel coordinate data into units of the (x, y) variables of each graph and compute a set of summary measures from the eyetracking microdata. We present the means of these measures in Table B10. As mentioned in Section 2.4.5, the experts perform better than the non-experts. The 15 percentage point expert-non-expert success gap translates to 1.7 graphs out of the 11 in total. As seen in the second and third rows of the table, experts spend more than twice as long on each graph and have more than twice as many fixations than non-experts. Next, we attempt to measure the "visual bandwidth" by calculating the horizontal range and horizontal standard deviation of the eyegaze coordinates. Out of a maximum horizontal range of 2 (recall that the support of the x-variable is normalized to be between -1 and 1), the overall average is 1.19, with the experts covering a 30% longer distance. The overall average standard deviation of the eyegazes is much smaller at 0.24, and again we see experts using a larger ocular bandwidth.

The next set of measures assesses the extent to which participants visually trace out the conditional expectation function. First, we compute the square root of the eyegaze IMSE, the RIMSE. The IMSE is the average squared deviation of the eyegaze from the true underlying CEF of the DGP, and we scale the RIMSE by σ. The average RIMSE is 0.25σ and shows little difference between experts and non-experts. Second, we fit RD

regression models to the eyegaze data using CCT's `rdrobust` package and record the conventional and bias-corrected local linear estimates. We then compute the absolute deviation of these estimates from the true discontinuity in the DGP and, consistent with our discontinuity specification in Section 2.3.2, scale this deviation by σ. One complication is that `rdrobust` cannot compute the bandwidths when there are few distinct eyegaze points, and the resulting conventional and bias-corrected local estimates are both missing. Not surprisingly, there are more missing estimates for non-experts (13%, or 13 out of 99 graphs) than experts (2%, or 1 out of 66 graphs) given that experts spend more time on each graph and exhibit a larger number of fixations. In these instances, we replace the conventional and bias-corrected estimates by global linear and global quadratic estimates, respectively.[7] The average deviation is 0.23σ for conventional estimates and 0.24σ for bias-corrected estimates. The average deviation is slightly lower among experts for conventional estimates and more so for bias-corrected estimates. However, this difference is driven by the global regression estimates—within the 151 graphs for which local RD estimates can be computed by `rdrobust`, the difference in the deviations between experts and non-experts virtually disappears.

We test for differences in the summary measures between experts and non-experts, accounting for the potential covariance for the same individual across graphs by clustering at the participant level. Because of the small number of participants, we implement the test via wild bootstrap (see Cameron and Miller, 2015 for a review). As highlighted in Section 2.4.5, the only measures that are significantly different between the experts and non-experts at the 5% level are time spent on each graph and whether the local RD eyegaze estimates are available. The difference in one other eyegaze measure, horizontal range, is statistically significant at the 10% level, and so is the difference in the outcome variable, correct discontinuity classification.

[7]The bias-corrected local linear estimate obtained by using the same main and pilot bandwidth from Calonico, Cattaneo, and Titiunik (2014) is numerically equivalent to the local quadratic estimate with the same bandwidth. Therefore, we can alternatively think of these global regressions as producing conventional and bias-corrected local linear estimates by using a large bandwidth.

For our prediction exercises, we split the data into a training sample and a testing sample, use the training sample to train the models, and compute the out-of-sample prediction error in the testing sample.[8] To gauge the robustness of the prediction results, we use two schemes to split our training and testing samples. In scheme A, we randomly select data from two-thirds of the 165 graphs to be the training sample. In scheme B, we randomly select all data from two-thirds of the participants (six of the nine non-experts and four of the six experts) to be the training sample. In both schemes, the out-of-sample prediction error rates range between 25% and 40%. In scheme A, the respective error rates are 25%, 40%, 33%, and 33% for penalized logit, pruned tree, bagging, and random forest. They are 31%, 29%, 33% and 29% in scheme B.

The penalized logit coefficient path plots in Figure B46 trace out the coefficients on the predictors as functions of the penalty parameter (Lambda). In scheme A, the first nine predictors to turn on (become nonzero) are eight of the graph identifiers and the expert indicator. In scheme B, the graph identifiers are also among the first to turn on, except the coefficients on RIMSE and the absolute deviation of the bias-corrected RD estimate also become nonzero at high values of the penalty parameter. In fact, both coefficients are negative, indicating that higher RIMSE and absolute deviation of eyegaze RD estimates are associated with a lower probability of successful inference, which makes intuitive sense. However, these are not robust associations: in scheme A, the RIMSE coefficient is instead positive, and although the coefficient on the absolute deviation of the bias-corrected estimate is still negative, the model selects this variable only when the penalty parameter is low, and the variable appears to counteract the *positive* coefficient on the absolute deviation of the conventional estimate. We see this offsetting pattern of the conventional and bias-corrected RD estimates in scheme B as well. Analogously, the horizontal eyegaze standard deviation and horizontal range also have opposite signs from the penalized logit model, and it is unclear whether a larger or smaller visual bandwidth is conducive to successful

[8] We select the tuning parameter in penalized-logit and pruned tree by five-fold cross validation.

inference.

The tree-based methods paint a broadly similar picture on the importance of the various predictors as penalized logit. In both training-testing schemes, the pruned tree is a stump and only splits by the Graph1 indicator (graph numbers also correspond to DGP numbers, i.e. Graph1 is from RD DGP1). Figure B47 presents the importance plots from bagging and random forest models, where each variable's importance is measured by the decrease in prediction accuracy if the variable is omitted from the model (see Friedman, Hastie, and Tibshirani, 2001 for details). The relative importance of the eyegaze measures varies by method and scheme, but the Graph1 and Graph4 indicators consistently stand out.

We also try re-running the prediction exercise after removing the four graphs that are "too easy" or "too hard," i.e. for which at least 14 out of 15 participants classify discontinuity correctly/incorrectly (Graph 1 is too hard, while Graphs 4, 10 and 11 are too easy). The resulting prediction error rates increase to between 40% and 50%, but the most important predictors still appear to be several graph identifiers. Our visual attention summary measures continue to lack robust predictive potency within the remaining seven graphs.

Based on these findings, we cannot pinpoint ocular patterns that robustly predict successful visual inference. We conclude that performance heterogeneity in visual inference is likely driven by the processing of visual signals and not by their acquisition as quantified by our eyegaze summary measures.

B.8 Appendix Tables

Table B1: Data-Driven Global Polynomial Orders for Fit Lines

DGP	AIC	BIC	Lee and Lemieux Dummy F-Test
DGP1	5	0	0
DGP2	0	0	0
DGP3	5	5	0
DGP4	5	1	0
DGP5	0	0	0
DGP6	5	5	3
DGP7	2	2	0
DGP8	4	1	3
DGP9	0	0	0
DGP10	2	2	1
DGP11	3	0	0

Notes: Table presents optimal polynomial orders based on three algorithms: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the procedure from Lee and Lemieux (2010) based on a joint F-test of significance for the bin dummies. We restrict all algorithms to polynomial orders zero through five.

Table B2: Timeline of RDD Experiments

Phase	Holding fixed	Treatments	Date	# participants recruited	# completions
RDD1	bin spacing: ES fitlines: no vertical line: yes	binwidth: large vs small # axis scaling: default vs 2x default	November 13-16, 2018	4*88=352	330 (94%)
RDD2	scaling: default fitlines: no vertical line: yes	binwidth: large vs small # bin spacing: ES vs QS	February 11-12, 2019	4*88=352	325 (92%)
RDD3	binwidth: small bin spacing: ES scaling: default	fitlines: no; center line: yes fitlines: no; center line: no fitlines: yes; center line: yes	February 27, 2019	3*88=264	248 (94%)
RDD4	bin spacing: ES scaling: default vertical line: yes	binwidth: large vs small # fitlines: yes vs no	October 28-29, 2019	4*88=352	340 (97%)

Table B3: Timeline of RKD Experiments

Phase	Holding fixed	Treatments (between subject)	Treatments (within subject)	Date	# participants recruited	# completions
Pilot	bin spacing: QS fit lines: no vertical line: yes scaling: default	binwidth: large vs small # detrending: yes vs no	derivative vs level	April 24-25, 2019	4*72=288	282 (98%)
RKD	bin spacing: QS fit lines: no vertical line: yes	binwidth: large vs small # detrending: yes vs no	derivative vs level	June 26-28, 2019	4*72=288	250 (87%)

Table B4: Treatment Effects: Phase RDD1

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; Default y-axis	0.197	0.209	0.177	0.140	-0.024	-0.012
	(0.039)	(0.045)	(0.042)	(0.046)	(0.030)	(0.019)
Small bin; Large y-axis	-0.024	-0.018	-0.056	0.008	-0.034	-0.035
	(0.029)	(0.043)	(0.042)	(0.043)	(0.029)	(0.026)
Large bin; Large y-axis	0.142	0.186	0.120	0.093	0.013	-0.001
	(0.036)	(0.040)	(0.042)	(0.042)	(0.025)	(0.017)
Small bin; Default y-axis (Mean)	0.054	0.181	0.367	0.669	0.922	0.988
Number of graphs	660	660	660	660	660	330
Number of players	330	330	330	330	330	330

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B5: Treatment Effects: Phase RDD2

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; Equal spacing	0.256	0.207	0.204	0.155	0.048	0.011
	(0.042)	(0.039)	(0.045)	(0.039)	(0.033)	(0.029)
Small bin; Quantile spacing	0.034	0.009	-0.012	-0.008	-0.030	-0.013
	(0.033)	(0.039)	(0.046)	(0.046)	(0.036)	(0.033)
Large bin; Qunatile spacing	0.157	0.185	0.217	0.138	0.030	0.024
	(0.038)	(0.040)	(0.046)	(0.039)	(0.031)	(0.023)
Small bin; Equal spacing (Mean)	0.054	0.169	0.355	0.657	0.892	0.964
Number of graphs	650	650	650	650	650	325
Number of players	325	325	325	325	325	325

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B6: Treatment Effects: Phase RDD3

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Fit line; Vertical line	0.142	0.215	0.172	0.051	0.057	-0.002
	(0.038)	(0.046)	(0.044)	(0.045)	(0.029)	(0.025)
No fit line; No vertical line	0.017	0.012	-0.007	-0.052	-0.026	-0.023
	(0.024)	(0.043)	(0.045)	(0.043)	(0.031)	(0.025)
No fit line; Vertical line (Mean)	0.037	0.159	0.348	0.634	0.860	0.976
Number of graphs	496	496	496	496	496	248
Number of players	248	248	248	248	248	248

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B7: Treatment Effects: Phase RDD4

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; No fit line	0.145	0.283	0.223	0.239	0.074	0.027
	(0.038)	(0.044)	(0.047)	(0.042)	(0.033)	(0.030)
Small bin; Fit line	-0.019	0.088	0.047	0.059	0.024	0.002
	(0.030)	(0.042)	(0.049)	(0.046)	(0.036)	(0.034)
Large bin; Fit line	0.230	0.300	0.239	0.207	0.058	0.003
	(0.045)	(0.046)	(0.047)	(0.044)	(0.034)	(0.034)
Small bin; No fit line (Mean)	0.074	0.154	0.395	0.611	0.870	0.951
Number of graphs	680	680	680	680	680	340
Number of players	340	340	340	340	340	340

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B8: Treatment Effects: Phase RKD (Levels)

	Dependent variable: player reports kink				
	True kink magnitude =				
	0	0.1944 ξ	0.324 ξ	0.54 ξ	0.9 ξ
Raw; Big bin	-0.008	0.054	0.107	-0.003	-0.084
	(0.040)	(0.046)	(0.052)	(0.057)	(0.064)
Residual; Small bin	0.013	0.002	0.046	0.111	-0.031
	(0.041)	(0.048)	(0.057)	(0.053)	(0.058)
Residual; Big bin	0.018	0.242	0.291	0.175	-0.015
	(0.041)	(0.051)	(0.055)	(0.054)	(0.059)
Raw; Small bin (Mean)	0.147	0.248	0.352	0.543	0.827
Number of graphs	614	608	607	614	307
Number of players	250	250	250	250	250

Notes: We have a stratified randomized experiment, where each of the 9 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B9: Treatment Effects: Phase RKD (Derivatives)

	Dependent variable: player reports kink				
	True kink magnitude =				
	0	0.1944 ξ	0.324 ξ	0.54 ξ	0.9 ξ
Derivative; Big bin	0.143	-0.004	0.089	0.153	0.302
	(0.040)	(0.041)	(0.045)	(0.042)	(0.058)
Derivative; Small bin (Mean)	0.235	0.312	0.294	0.306	0.335
Number of players	250	250	250	250	250
Number of graphs	614	608	607	614	307

Notes: We have a stratified randomized experiment, where each of the 9 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table B10: Summary Measures of Eyegaze Data

	Average across graphs			
	(1)	(2)	(3)	(4)
	Experts	Non-experts	Overall	p-value of difference between (1) and (2)
Number of Correct Classifications	0.68	0.53	0.59	0.078
Time Spent per Graph in Seconds	16.8	6.9	10.9	0.046
Number of Fixations per Graph	45.8	20.5	30.6	0.112
Horizontal Range of Eyegazes	1.40	1.05	1.19	0.090
Horizontal Standard Deviation of Eyegazes	0.26	0.22	0.24	0.268
RIMSE of Eyegazes (scaled)	0.25	0.26	0.25	0.615
Conventional Local Linear RD Estimate from Eyegazes				
Absolute Deviation from True Discontinuity (scaled)	0.22	0.23	0.23	0.136
Discontinuity Estimate Unavailable	0.02	0.13	0.08	0.032
Bias-corrected Local Linear RD Estimate from Eyegazes				
Absolute Deviation from True Discontinuity: Bias-corrected RD Estimate	0.22	0.25	0.24	0.103
Discontinuity Estimate Unavailable	0.02	0.13	0.08	0.032
Number of Participants	6	9	15	
Number of Graphs	66	99	165	

Notes: The statistical test in column (4) accounts for correlation across the same participant via clustering, and we use wild bootstrap due to the small number of clusters.

Table B11: Timeline of Expert Study

Participant pool	Binwidth choice in Part 1	Date	# participants recruited	# completions
Cornell Econometrics Workshop (RD only)	RD: small binwidth, even spacing	May 7, 2019	-	12
UC Irvine Econometrics Series (RD only)	RD: small binwidth, even spacing	May 21, 2019	-	15
Princeton Quantitative Social Science Colloquium (RD only)	RD: large binwidth, even spacing	October 11, 2019	-	48
NBER + IZA (RD only, Pilot)	RD: small binwidth, even spacing	July 28 - August 2, 2019	101	7 (7%)
NBER + IZA Email list (RD and RK, Pilot)	RD: small binwidth, even spacing RK: big binwidth, quantile spacing	August 12 - 19, 2019	200	21 (11%)
NBER + IZA Email list (RD and RK)	RD: small binwidth, even spacing RK: big binwidth, quantile spacing	August 26 - September 3, 2019	1,000	94 (9%)

Table B12: Covariate Means and Balance across Treatment Arms: Phase RDD1

	T1	T2	T3	T4	p-value
Female	0.687	0.738	0.759	0.688	0.649
Completed college	0.518	0.512	0.482	0.625	0.268
Age 18 to 22	0.422	0.571	0.554	0.425	0.088
Age 23 to 49	0.494	0.381	0.398	0.487	0.318
Age 50 or older	0.060	0.048	0.048	0.087	0.742
Graduate stats knowledge	0.145	0.119	0.133	0.050	0.091
Passed attention check	0.867	0.857	0.843	0.912	0.516
Attrition rate	0.057	0.045	0.057	0.091	0.690
Participants	83	84	83	80	

Notes: The joint p-value for the significance of the covariates is 0.175.

Table B13: Covariate Means and Balance across Treatment Arms: Phase RDD2

	T1	T2	T3	T4	p-value
Female	0.614	0.623	0.617	0.726	0.325
Completed college	0.494	0.429	0.543	0.512	0.526
Age 18 to 22	0.518	0.584	0.481	0.488	0.548
Age 23 to 49	0.386	0.364	0.481	0.429	0.447
Age 50 or older	0.084	0.039	0.037	0.071	0.485
Graduate stats knowledge	0.072	0.143	0.099	0.119	0.499
Passed attention check	0.928	0.870	0.877	0.940	0.306
First time player	0.711	0.805	0.802	0.762	0.466
Attrition rate	0.057	0.125	0.080	0.045	0.264
Participants	83	77	81	84	

Notes: The joint p-value for the significance of the covariates is 0.496.

Table B14: Covariate Means and Balance across Treatment Arms: Phase RDD3

	T1	T2	T3	p-value
Female	0.683	0.725	0.616	0.323
Completed college	0.659	0.650	0.558	0.339
Age 18 to 22	0.439	0.463	0.523	0.529
Age 23 to 49	0.488	0.512	0.419	0.449
Age 50 or older	0.073	0.025	0.047	0.345
Graduate stats knowledge	0.159	0.113	0.070	0.184
Passed attention check	0.902	0.900	0.907	0.988
First time player	0.488	0.500	0.453	0.823
Attrition rate	0.068	0.091	0.023	0.087
Participants	82	80	86	

Notes: The joint p-value for the significance of the covariates is 0.653.

Table B15: Covariate Balance across Treatment Arms: Phase RDD4

	T1	T2	T3	T4	p-value
Female	0.654	0.782	0.733	0.744	0.327
Completed college	0.543	0.655	0.442	0.535	0.040
Age 18 to 22	0.568	0.471	0.663	0.570	0.084
Age 23 to 49	0.420	0.506	0.291	0.407	0.031
Age 50 or older	0.012	0.023	0.035	0.023	0.797
Graduate stats knowledge	0.160	0.115	0.070	0.047	0.067
Passed attention check	0.914	0.897	0.953	0.907	0.427
First time player	0.556	0.563	0.628	0.581	0.771
Attrition rate	0.080	0.011	0.023	0.023	0.185
Participants	81	87	86	86	

Notes: The joint p-value for the significance of the covariates is 0.1365.

Table B16: Covariate Means and Balance across Treatment Arms: Phase RKD

	T1	T2	T3	T4	p-value
Female	0.758	0.721	0.705	0.710	0.903
Completed college	0.515	0.492	0.541	0.500	0.952
Age 18 to 22	0.667	0.574	0.557	0.597	0.588
Age 23 to 49	0.288	0.361	0.393	0.387	0.554
Age 50 or older	0.030	0.066	0.016	0.016	0.522
Graduate stats knowledge	0.030	0.115	0.131	0.113	0.052
Passed attention check	0.909	0.967	0.951	0.919	0.466
First time player	0.455	0.410	0.557	0.468	0.417
Attrition rate	0.250	0.307	0.307	0.295	0.804
Participants	66	61	61	62	

Notes: The joint p-value for the significance of the covariates is 0.545.

Table B17: Correlations of Correct Classification with Subject Covariates for RD

	Phase I	Phase II	Phase III	Phase IV
Female	-0.001	-0.007	-0.019	0.017
	(0.014)	(0.014)	(0.018)	(0.017)
Completed college	0.004	-0.018	-0.025	0.015
	(0.019)	(0.018)	(0.019)	(0.023)
Age 23 to 49	-0.006	0.003	-0.012	0.009
	(0.021)	(0.019)	(0.021)	(0.023)
Age 50 or older	0.004	0.037	0.094	0.055
	(0.026)	(0.031)	(0.033)	(0.073)
Graduate stats knowledge	-0.013	0.030	0.053	-0.048
	(0.022)	(0.022)	(0.028)	(0.029)
Passed attention check	0.055	0.043	-0.014	-0.011
	(0.023)	(0.026)	(0.029)	(0.031)
First time player	NA	0.000	-0.001	0.018
	(NA)	(0.016)	(0.017)	(0.015)
Joint test p-value	0.321	0.460	0.009	0.389

Notes: This table presents the coefficients from regressing whether study participants are correct in classifying an RD graph on their characteristics. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table B18: Correlations of Correct Classification with Subject Covariates for RK

	RK Phase I
Female	-0.009
	(0.018)
Completed college	-0.030
	(0.021)
Age 23 to 49	-0.003
	(0.023)
Age 50 or older	-0.047
	(0.035)
Graduate stats knowledge	0.017
	(0.033)
Passed attention check	-0.013
	(0.039)
Joint test p-value	0.347

Notes: This table presents the coefficients from regressing whether study participants are correct in classifying an RK graph on their characteristics. Standard errors are clustered at the participant level.

Table B19: Predictions of Power on DGP Characteristics

	Phase RDD1	Phase RDD2	Phase RDD3	Phase RDD4
LHS deriv-, RHS deriv-, discont-	0.244	0.262	0.267	0.308
	(0.032)	(0.033)	(0.039)	(0.031)
LHS deriv+, RHS deriv-, discont-	-0.005	0.029	-0.020	0.096
	(0.039)	(0.042)	(0.047)	(0.041)
LHS deriv-, RHS deriv+, discont-	0.024	0.103	0.173	0.102
	(0.058)	(0.057)	(0.068)	(0.054)
LHS deriv+, RHS deriv+, discont-	-0.111	-0.169	-0.103	-0.110
	(0.022)	(0.022)	(0.026)	(0.022)

Notes: This table presents the coefficients from regressions of whether the participant is correct in classifying discontinuous graphs on the interactions of the signs of left and right derivatives of the CEF with the sign of discontinuity. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table B20: Predictions of Size on DGP Characteristics

	Phase RDD1	Phase RDD2	Phase RDD3	Phase RDD4
LHS deriv-, RHS deriv-	0.052	0.049	-0.003	0.093
	(0.035)	(0.037)	(0.029)	(0.036)
LHS deriv+, RHS deriv-	-0.040	-0.018	0.072	-0.048
	(0.032)	(0.035)	(0.037)	(0.032)
LHS deriv-, RHS deriv+	-0.124	-0.168	-0.032	-0.074
	(0.026)	(0.022)	(0.033)	(0.039)
LHS deriv+, RHS deriv+ (Mean)	0.141	0.168	0.077	0.154
	(0.020)	(0.022)	(0.018)	(0.020)

Notes: This table presents the coefficients from regressions of whether the participant is correct in classifying continuous graphs on the interactions of the signs of left and right derivatives of the CEF. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table B21: Dynamic Discontinuity Classification: AR(1) Regression

	(1)	(2)
Phase	Estimate of ρ from Eq. (B.3)	Bias-Corrected Estimate of ρ
RDD1	-0.127	-0.041
	(0.038)	(0.046)
RDD2	-0.170	-0.088
	(0.035)	(0.039)
RDD3	-0.099	-0.009
	(0.029)	(0.036)
RDD4	-0.126	-0.034
	(0.035)	(0.041)
Expert	-0.205	-0.126
	(0.032)	(0.037)

Notes: Standard errors are in parentheses. Standard errors in both columns are clustered at the participant level, and those in column (2) additionally corrects the $O(1/S)$ bias from regression (B.3) by inverting equation (18) of Nickell (1981).

Table B22: Fisher's Exact Test of Association: p-values (One-Sided)

Panel A: Expert Visual vs Econometric Inferences (RDD)

	Estimator for Econometric Inference			
Discontinuity $\lvert d\rvert$	PQ	IK	CCT	AK
0	0.727	0.231	0.616	0.695
0.1944σ	0.000	0.000	0.000	0.000
0.324σ	0.000	0.000	0.000	0.000
0.54σ	0.000	0.000	0.001	0.000
0.9σ	0.073	.	0.042	0.112
1.5σ	.	.	.	.

Panel B: Expert Visual vs Econometric Inferences (RKD)

	Estimator for Econometric Inference				
Kink $\lvert d\rvert$	PQ	CCT1	CCT2	CCT3	CCT4
0	0.289	0.707	.	0.613	0.004
0.1944ξ	0.124	0.418	0.432	0.571	0.076
0.324ξ	0.456	0.212	0.600	0.080	0.194
0.54ξ	0.087	0.158	0.521	0.571	0.070
0.9ξ	0.191	0.572	0.626	0.258	0.601

Notes: Missing values indicate that a test always accepts or rejects the null hypothesis. The five columns correspond to the five estimators underlying the power functions in Figure 2.5. PQ is "Piecewise Quintic"; CCT1 is "CCT Robust"; CCT2 is "CCT Rob. Lin. w/ Reg. Tri."; CCT3 is "CCT Conv. Quad. Unif."; CCT4 is "CCT Conv. Lin. Unif.".

Table B23: Timeline of Eyetracking Study

Participant pool	Date	# participants recruited	# completions
Nonexperts	July 15 & 17, 2019	11	10
Experts	July 30, 2019	6	6

B.9 Appendix Figures

Figure B1: Same Data with Different Axis Scales for Correlation Detection

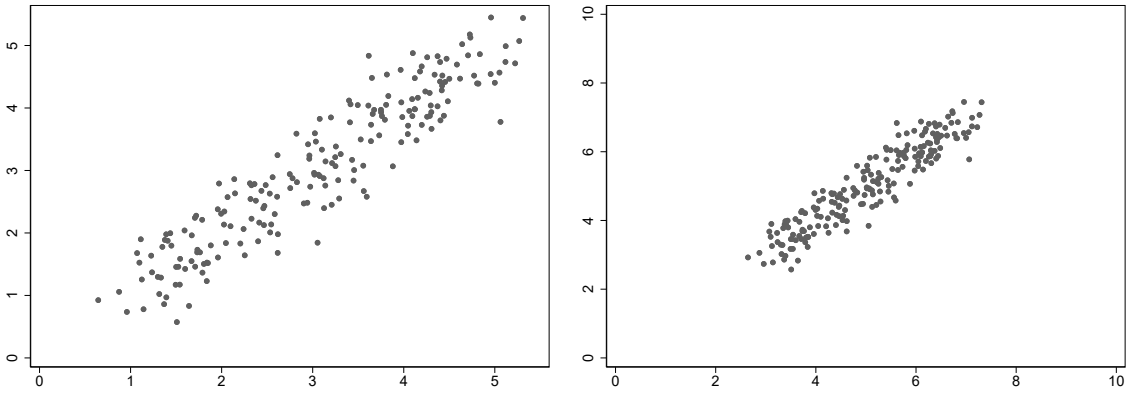


Figure B2: Original Plot vs Plot with Residualized Y

(a) Original

(b) Residualized

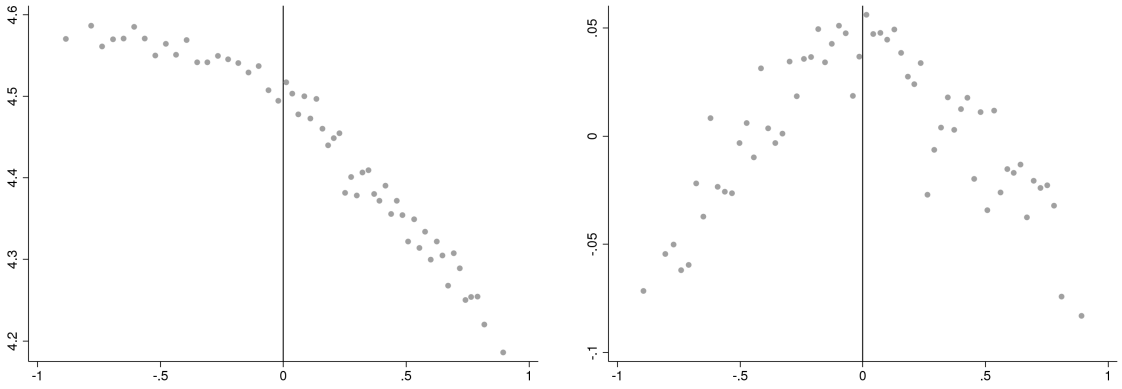
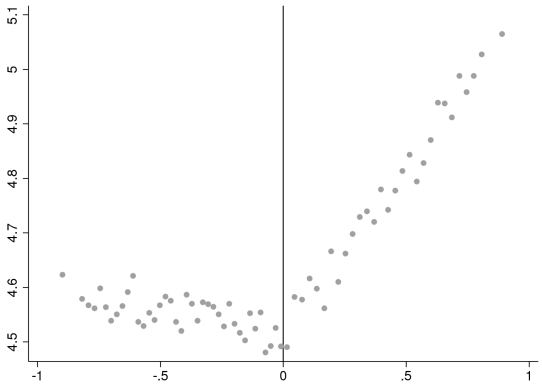


Figure B3: Original Plot vs Plot of Estimated First Derivatives

(a) Original



(b) Estimated First Derivatives

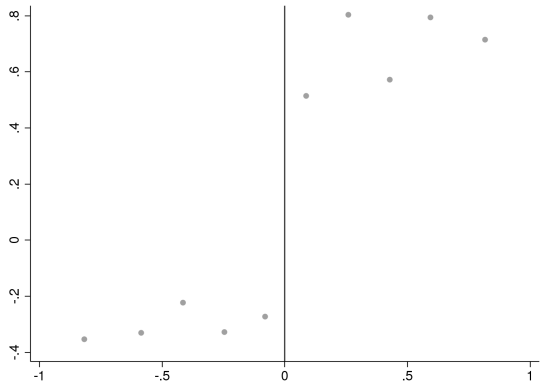
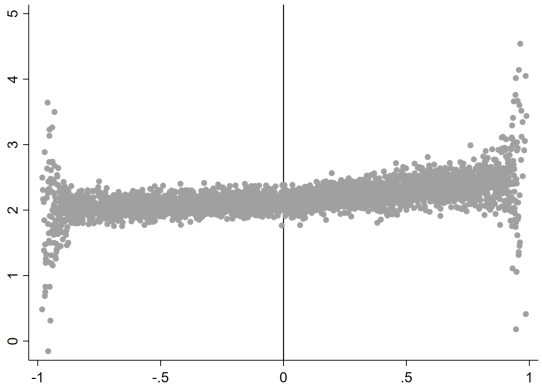


Figure B4: RKD: Even Spacing vs Quantile Spacing Example

(a) Even Spacing



(b) Quantile Spacing

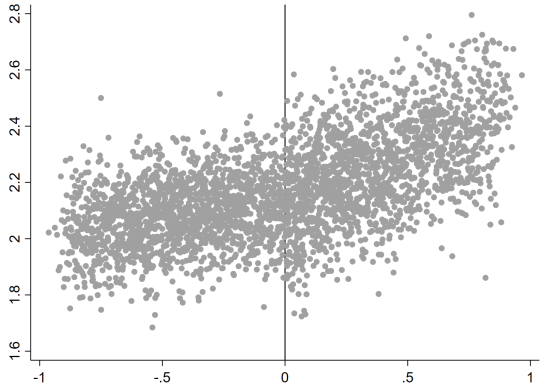


Figure B5: RDD DGP Conditional Expectation Functions

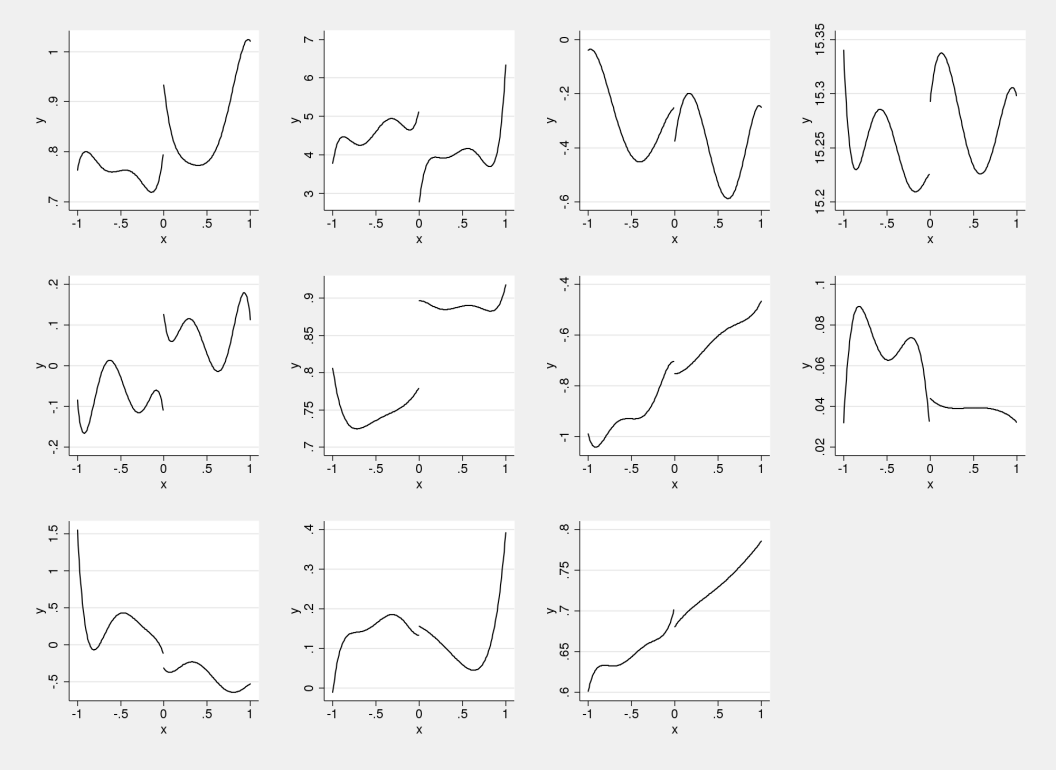
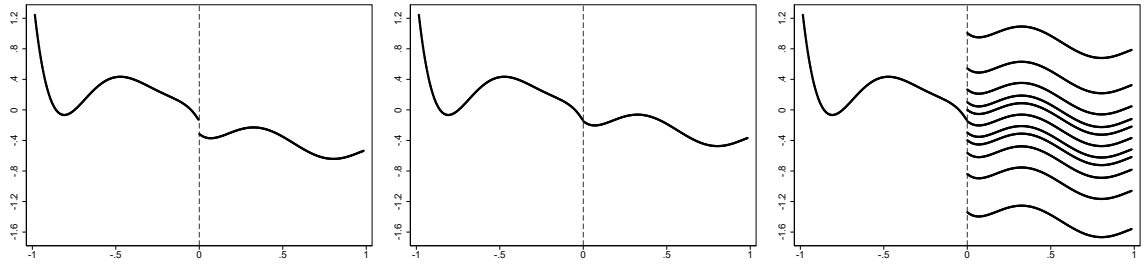


Figure B6: Creation of RDD DGP Conditional Expectation Functions, DGP9



Notes: The leftmost figure plots the piecewise quintic CEF from the original microdata of the paper from which we calibrated DGP9. The central figure removes the discontinuity by setting the right intercept to equal the left intercept. The rightmost figure plots the final 11 CEFs for DGP9 corresponding to different levels of discontinuity by further changing the right intercept.

Figure B7: RKD DGP Conditional Expectation Functions

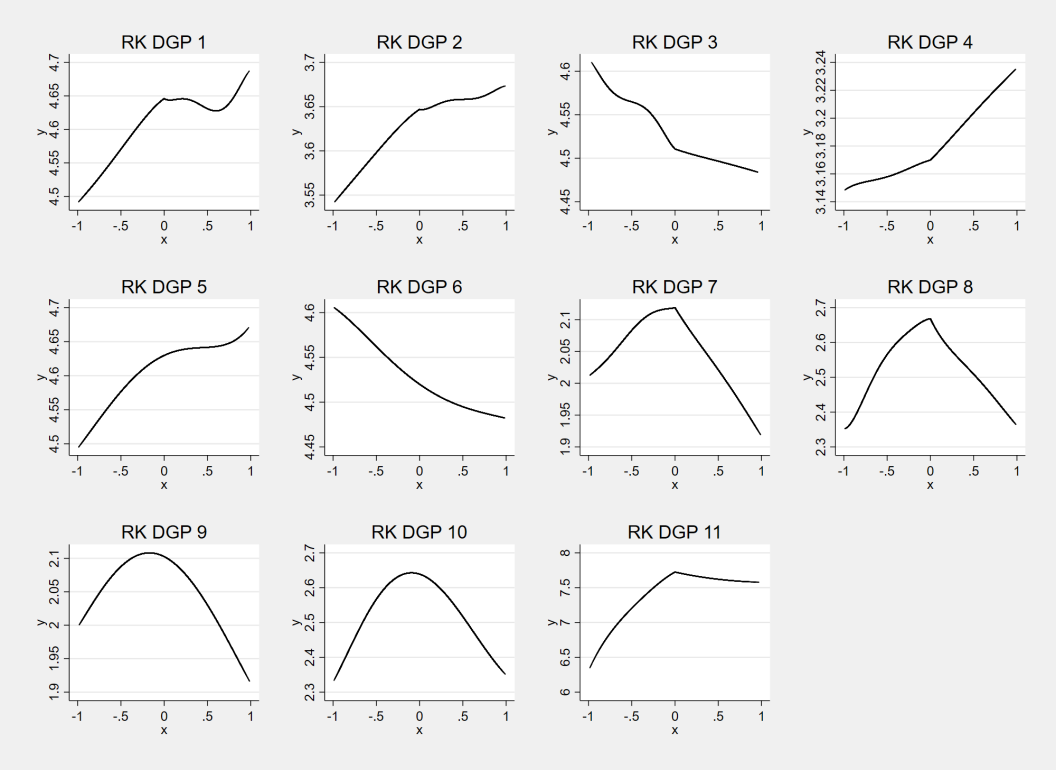
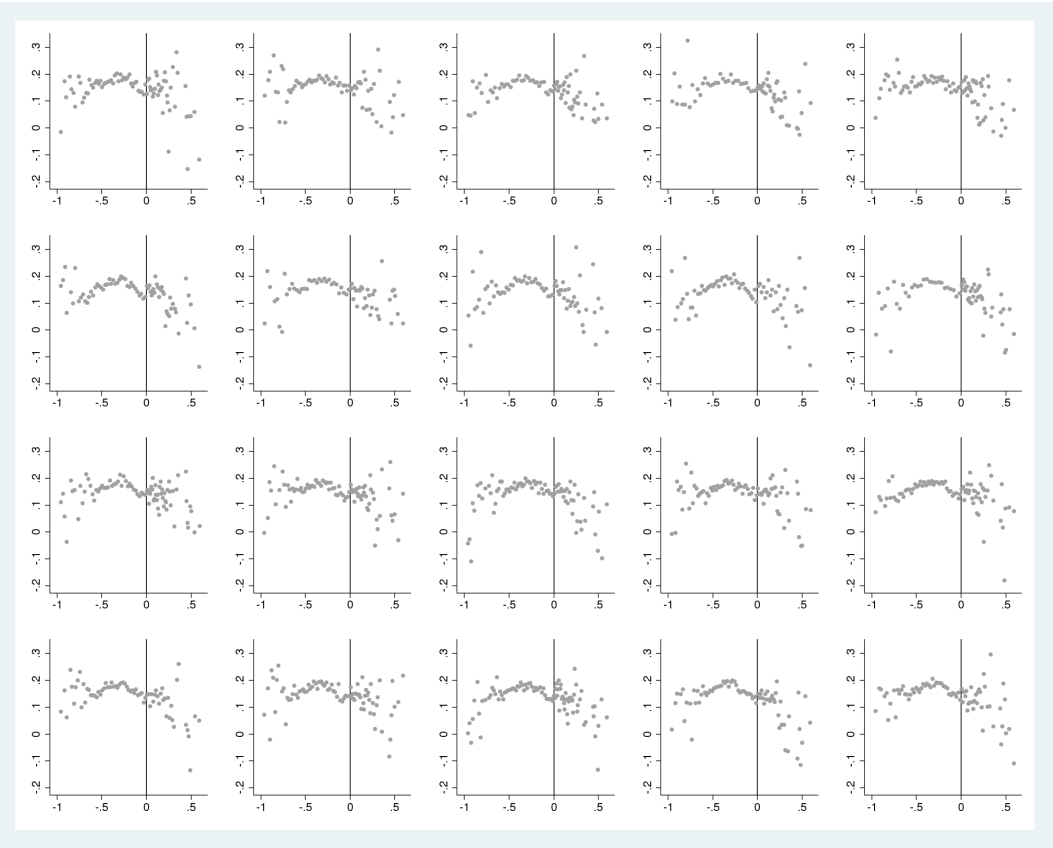
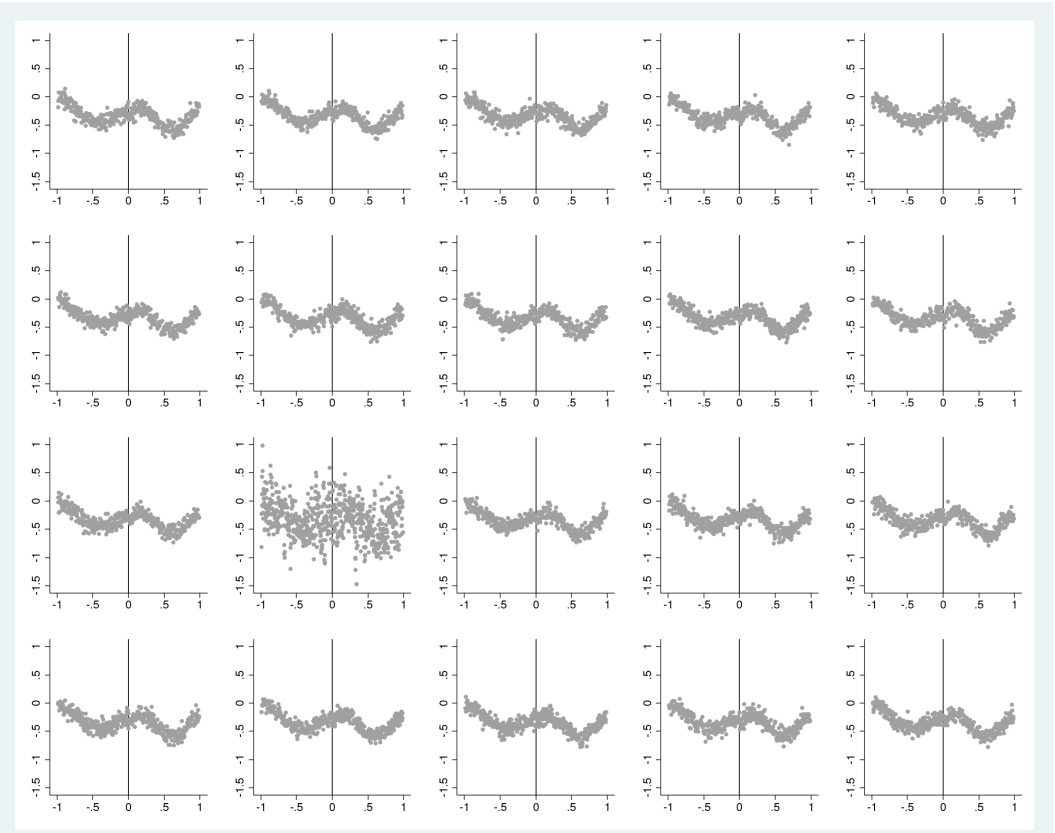


Figure B8: Lineup Protocol Graph: RD DGP10



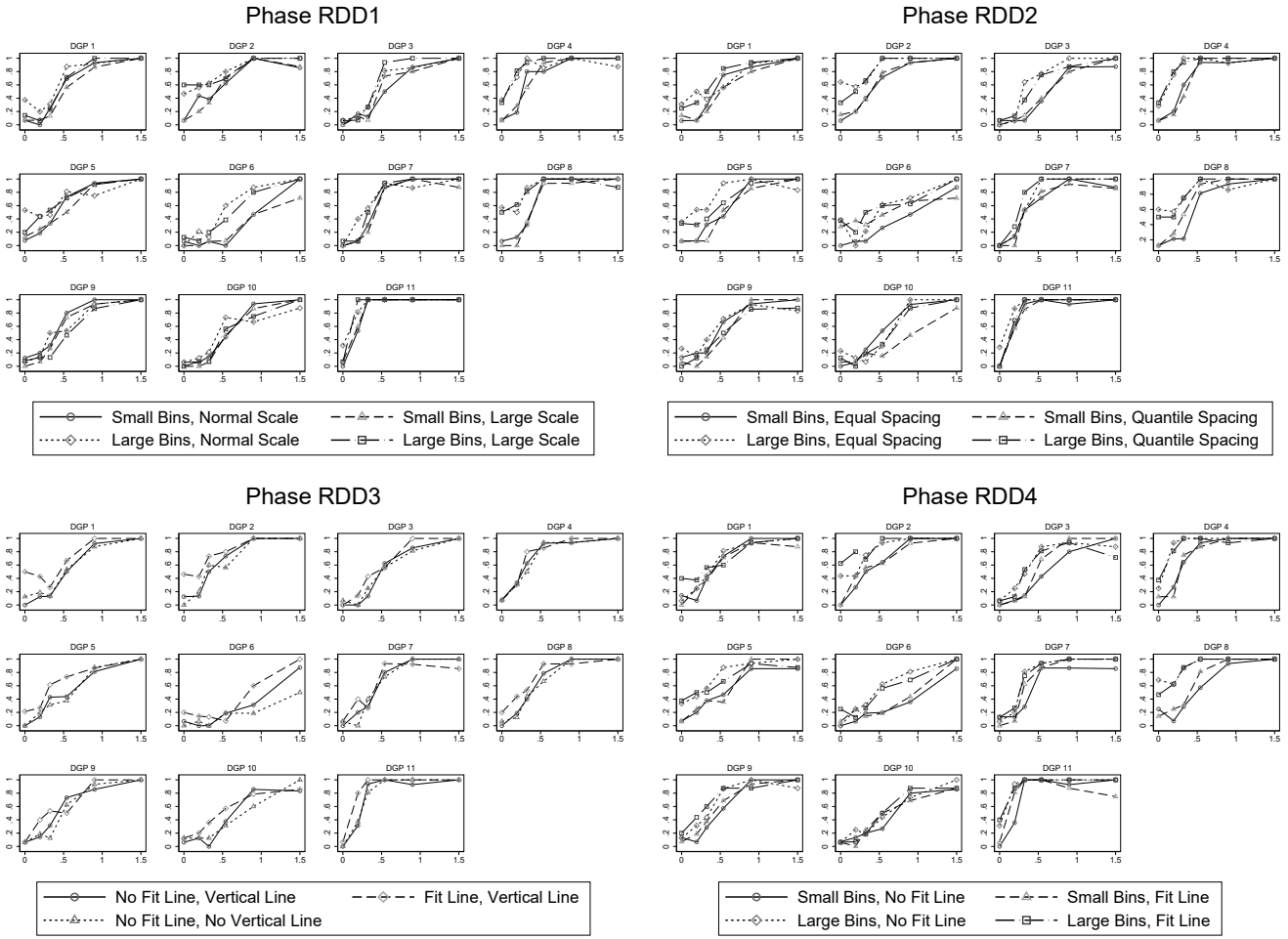
Notes: One of the 20 graphs is produced from the real data. The other 19 are produced from simulated data drawn from the DGP calibrated to the real data.

Figure B9: Lineup Protocol Graph: RD DGP3



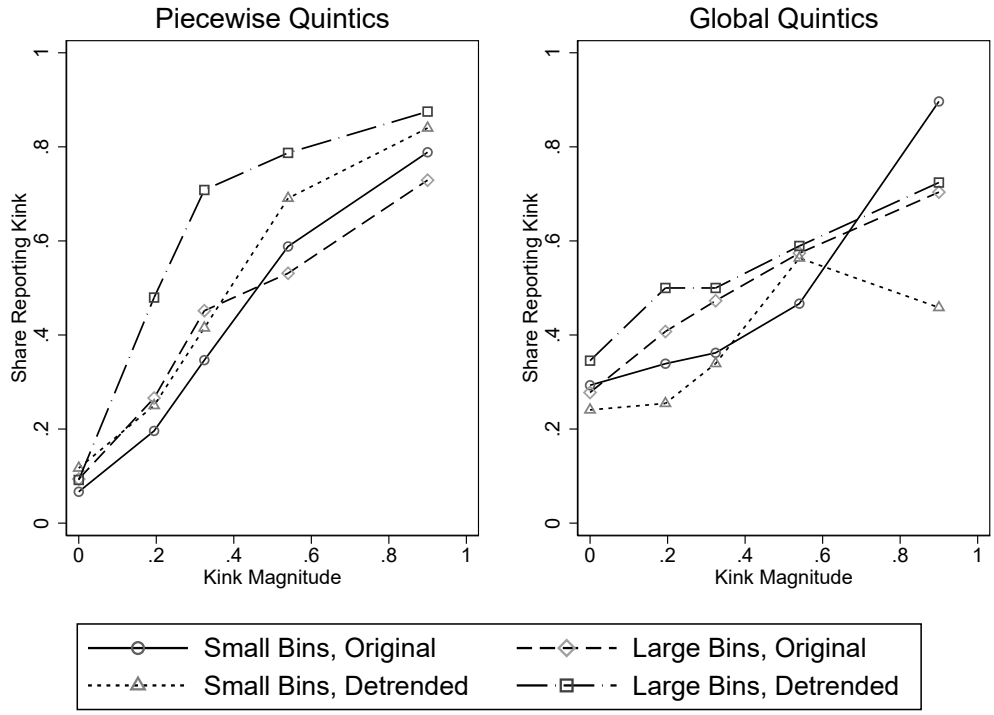
Notes: One of the 20 graphs is produced from the real data. The other 19 are produced from simulated data drawn from the DGP calibrated to the real data.

Figure B10: RDD Phases: Power Functions by DGP



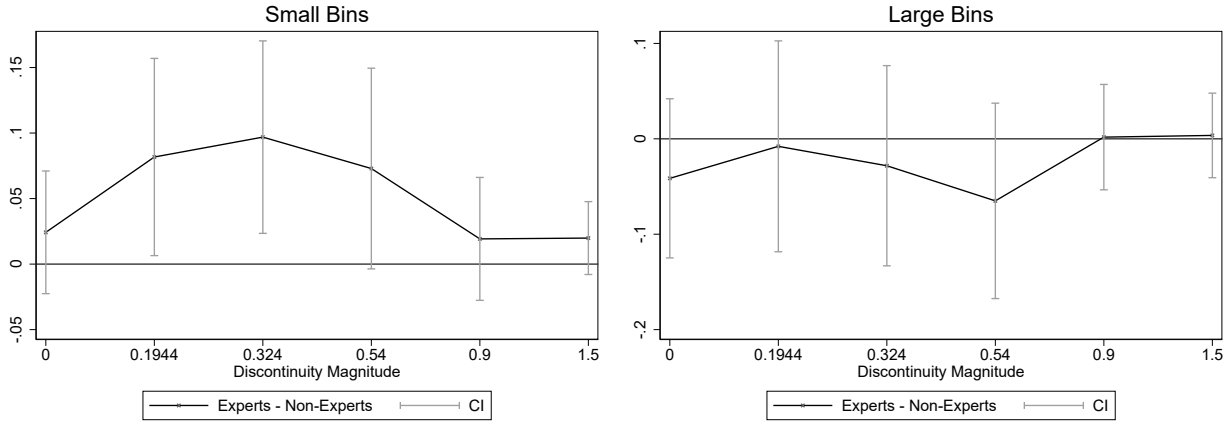
Notes: The figures here break up the power functions plotted in Figure 2.1 by DGP.

Figure B11: RKD Phase: Power Functions By Quintic Type



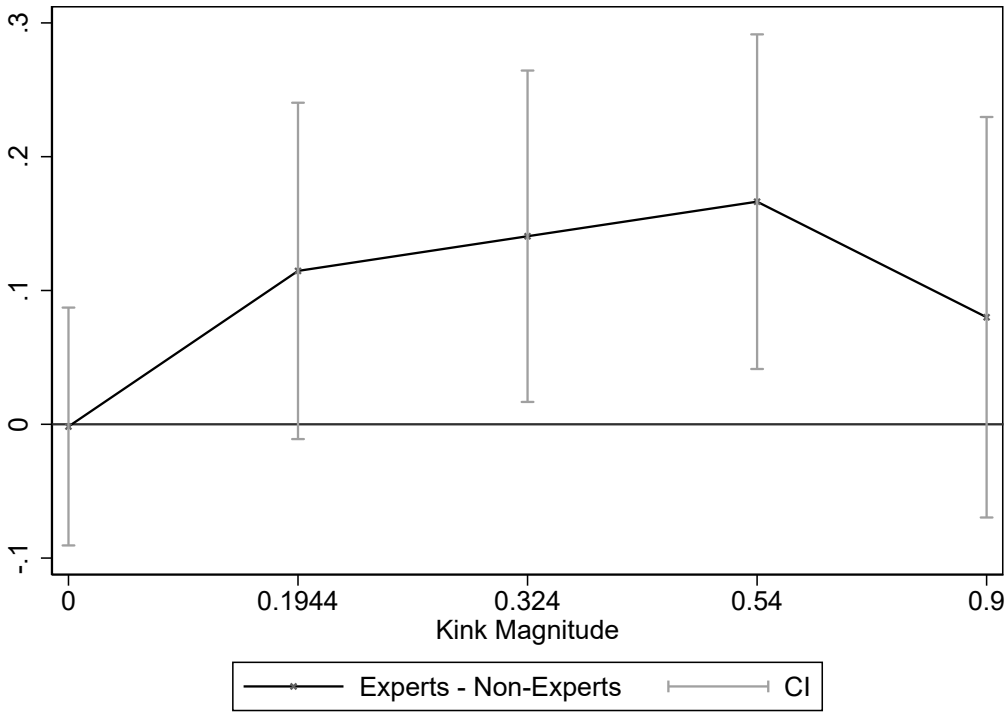
Notes: The figures here break up the power functions plotted in Figure 2.2 by two groups of DGPs—those specified with piecewise quintics and those specified with global quintics.

Figure B12: RD: Expert vs Non-Expert Performance Differences



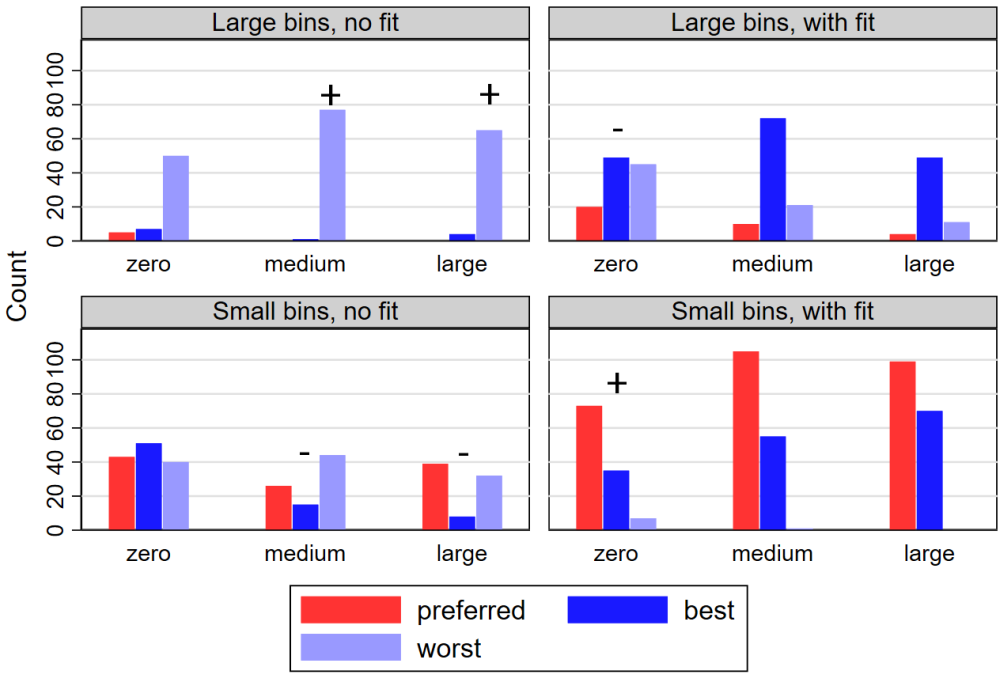
Notes: These figures plot the difference in the RD power functions between experts and non-experts from the left panel of Figure 2.3.

Figure B13: RK: Expert vs Non-Expert Performance Differences



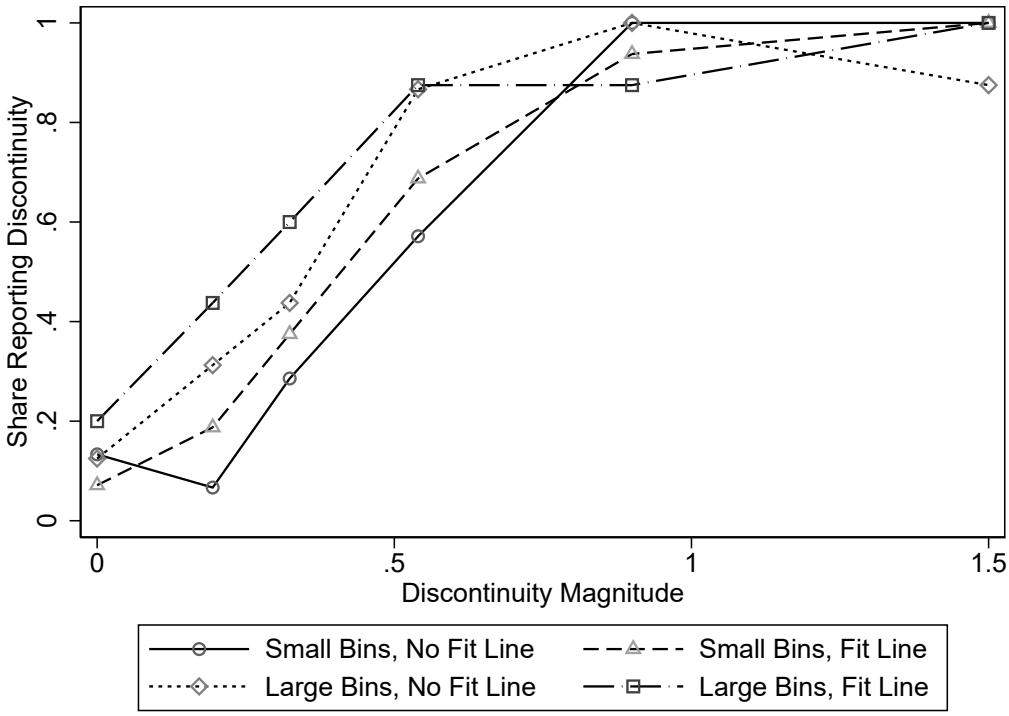
Notes: This figure plots the difference in the RK power functions between experts and non-experts from the right panel of Figure 2.3.

Figure B14: RD: Expert Preferences and Beliefs regarding Non-Expert Performance



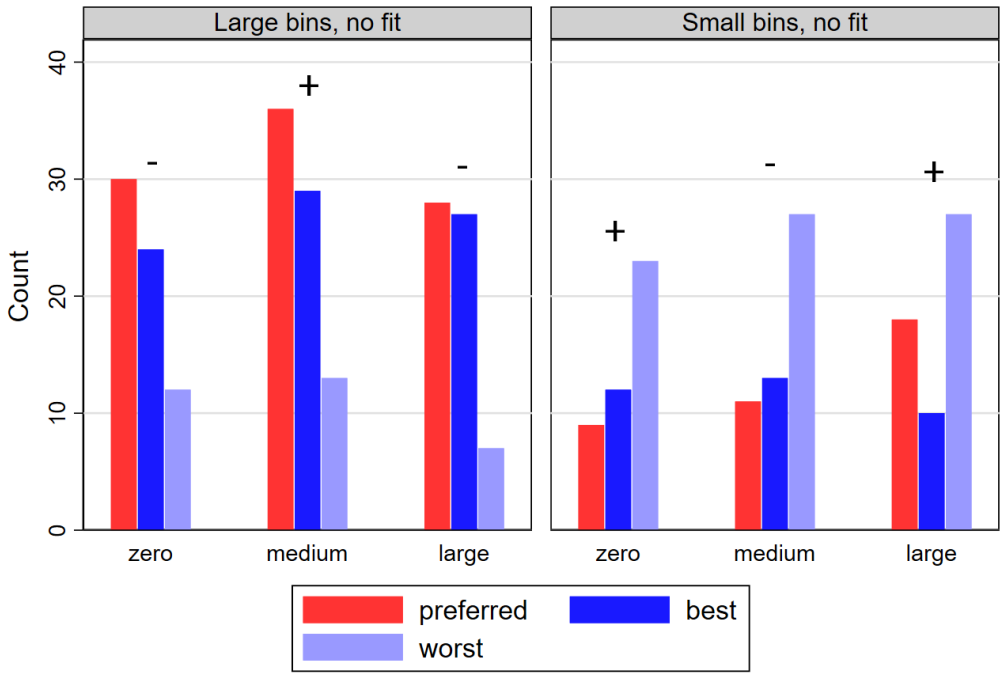
Notes: Each panel of this figure shows the number of experts who report the given treatment as being their preferred treatment at a given discontinuity level; believe it to be the best-performing treatment among our non-expert sample; or believe it to be the worst-performing treatment among our non-expert sample. For comparison, the treatments that performed best and worst in that sample are marked with a $+$ and $-$ sign respectively.

Figure B15: Phase RDD4: Power Functions for DGP Shown to Experts



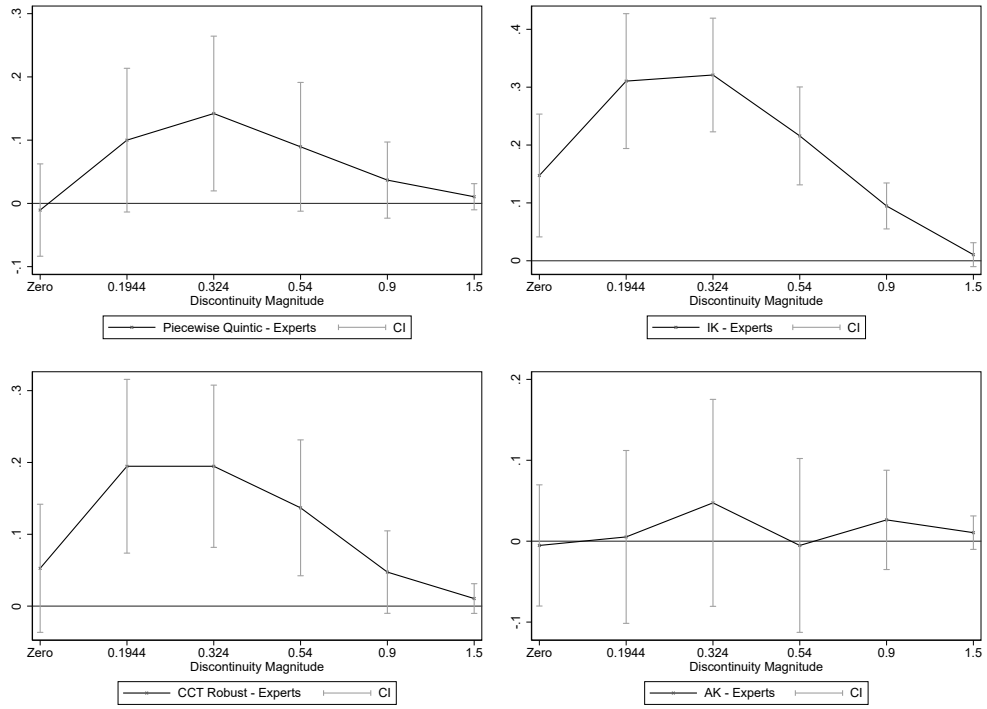
Notes: This figure shows the power functions for the DGP that was used to elicit expert preferences and beliefs across four considered treatments and three discontinuity levels in the second part of the expert survey. The corresponding expert preferences and beliefs are shown in Figure B14.

Figure B16: RK: Expert Preferences and Beliefs regarding Non-Expert Performance



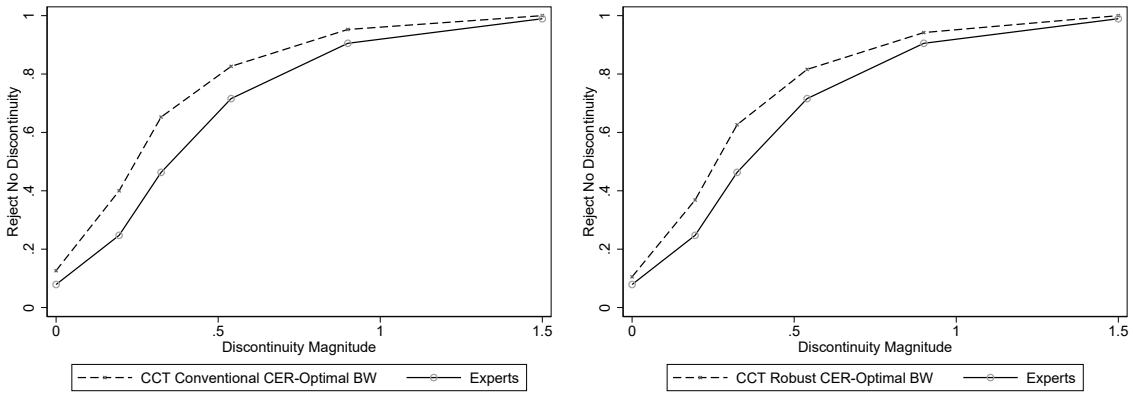
Notes: Each panel of this figure shows the number of experts who report the given treatment as being their preferred treatment at a given discontinuity level; believe it to be the best-performing treatment among our non-expert sample; or believe it to be the worst-performing treatment among our non-expert sample. For comparison, the treatments that performed best and worst in that sample are marked with a $+$ and $-$ sign respectively.

Figure B17: RD: Expert Visual vs Econometric Inference Power Function Differences



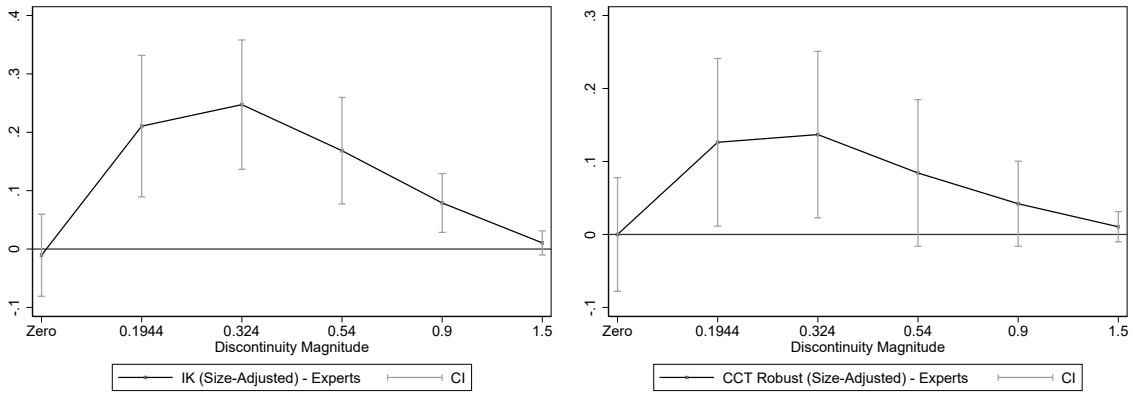
Notes: This figure plots the difference in the RD power functions between experts' visual inference and four econometric inference procedures from the left panel of Figure 2.4.

Figure B18: RD: Expert Visual vs CCT Procedure with Inference-Optimal Bandwidth Power Functions



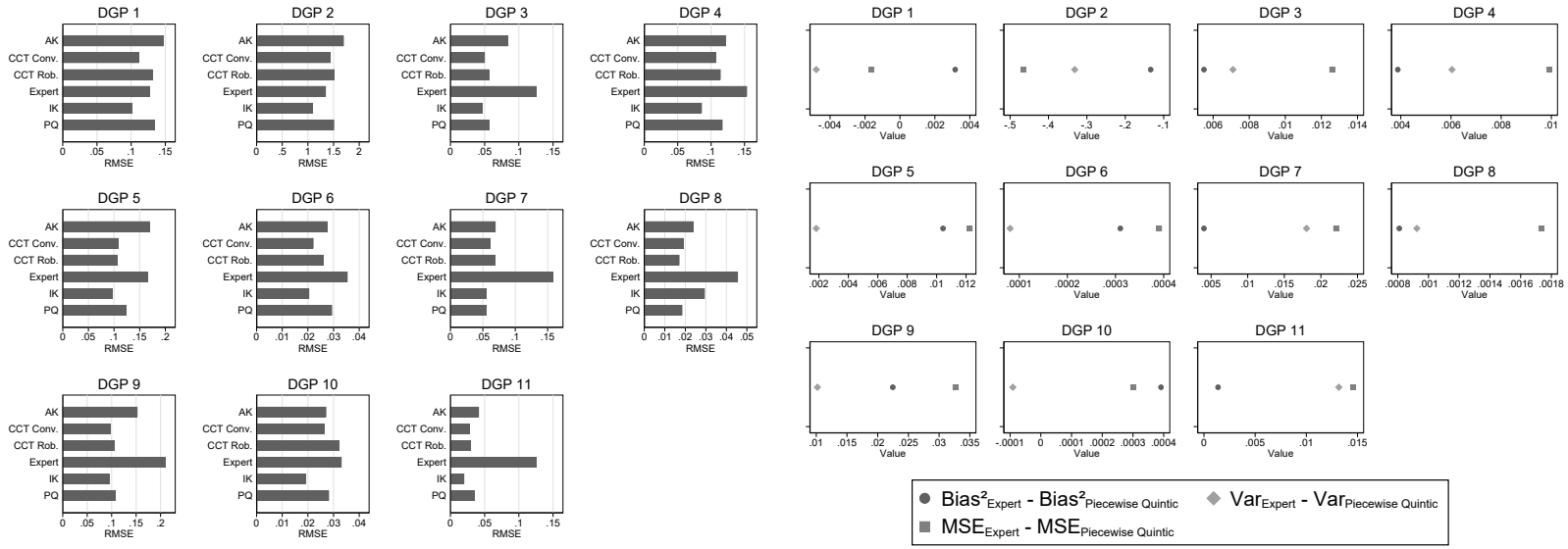
Notes: This figure plots the power functions of experts' visual inference and the CCT procedures with inference-optimal bandwidth per Calonico, Cattaneo, and Farrell (2020b)

Figure B19: RD: Expert Visual vs Size-Adjusted Econometric Inference Power Function Differences



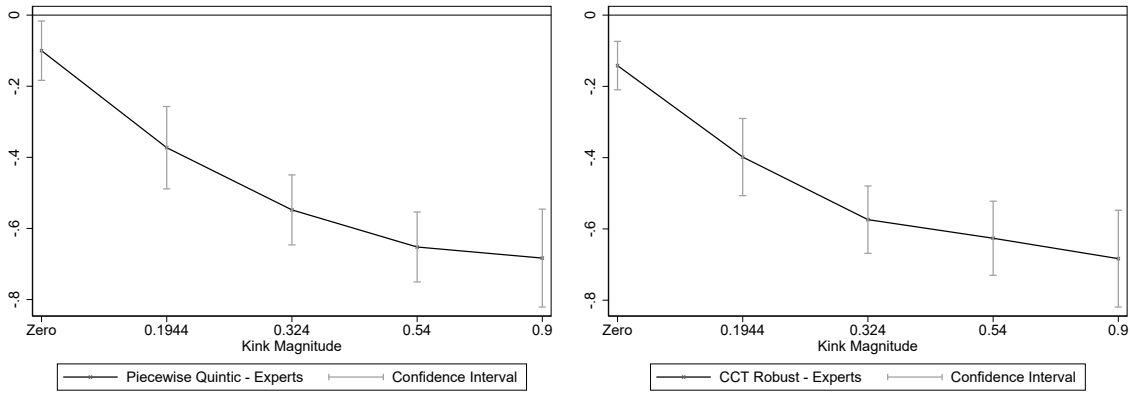
Notes: This figure plots the difference in the RD power functions between experts' visual inference and the size-adjusted IK and CCT procedures from the right panel of Figure 2.4.

Figure B20: RMSE of Point Estimates and MSE Decomposition by DGP



Notes: The left panel compares for each DGP the RMSE in estimating the discontinuity magnitude in visual and econometric procedures. The right panel decomposes the difference in RMSE between visual estimation and the correctly specified regression benchmark into the bias and variance components.

Figure B21: RK: Expert Visual vs Econometric Inferences Power Function Differences



Notes: This figure plots the difference in the RK power functions between experts' visual inference and the default CCT procedure of Figure 2.5.

Figure B22: Lineup Protocols for DGPs Specified Before and After Adding Noise to the Running Variable

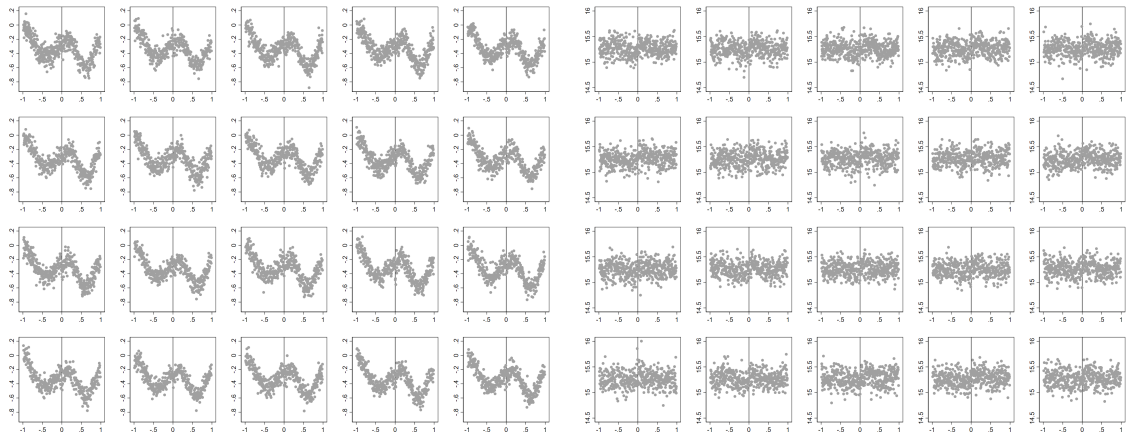


Figure B23: Comparison between CEFs from Original Microdata and Adding Noise to Running Variable

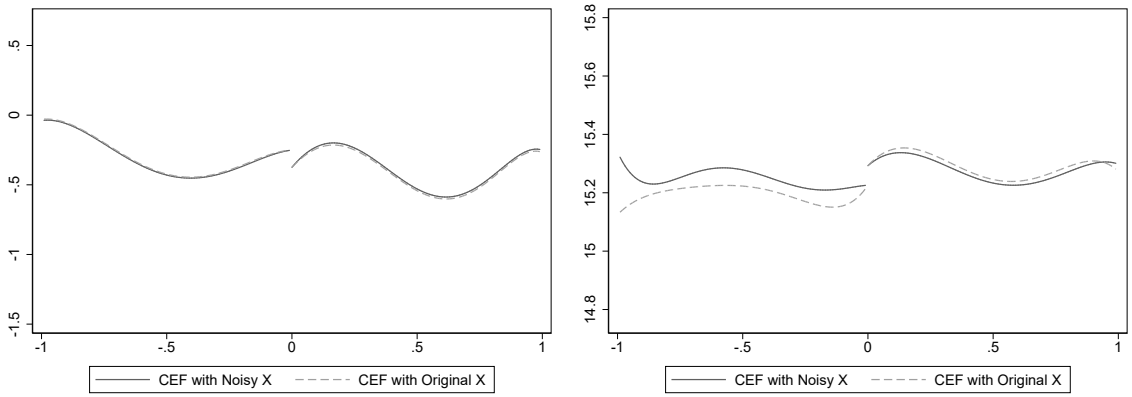


Figure B24: Video Tutorial Attention Check

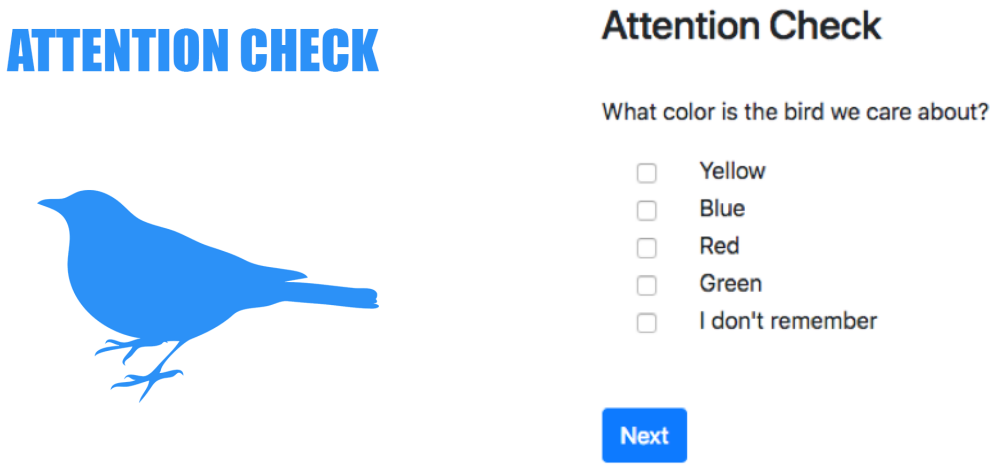
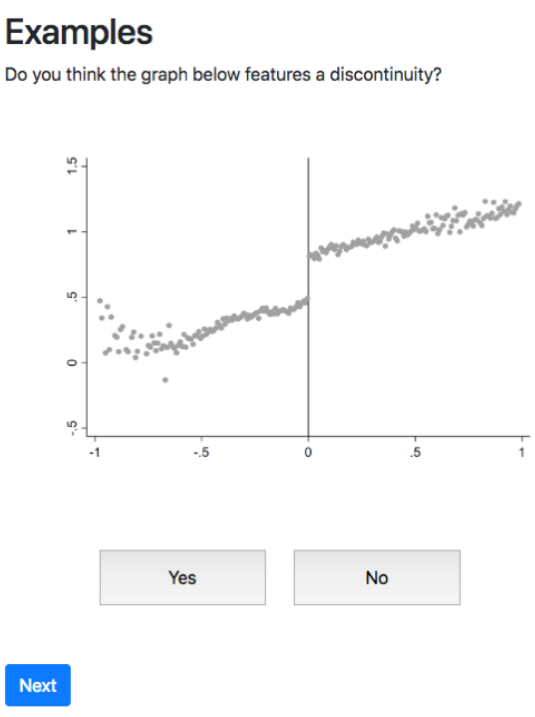


Figure B25: Example Tasks

(a) Example 1



(b) Example 1 - Feedback

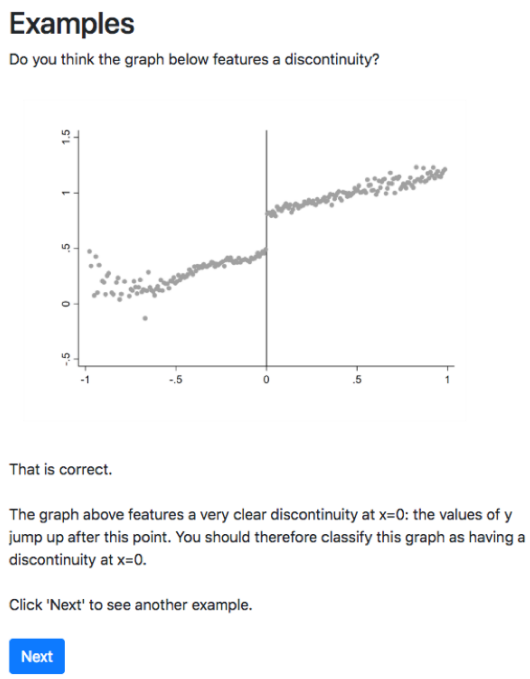


Figure B26: Example 4 - Feedback and Navigation Buttons

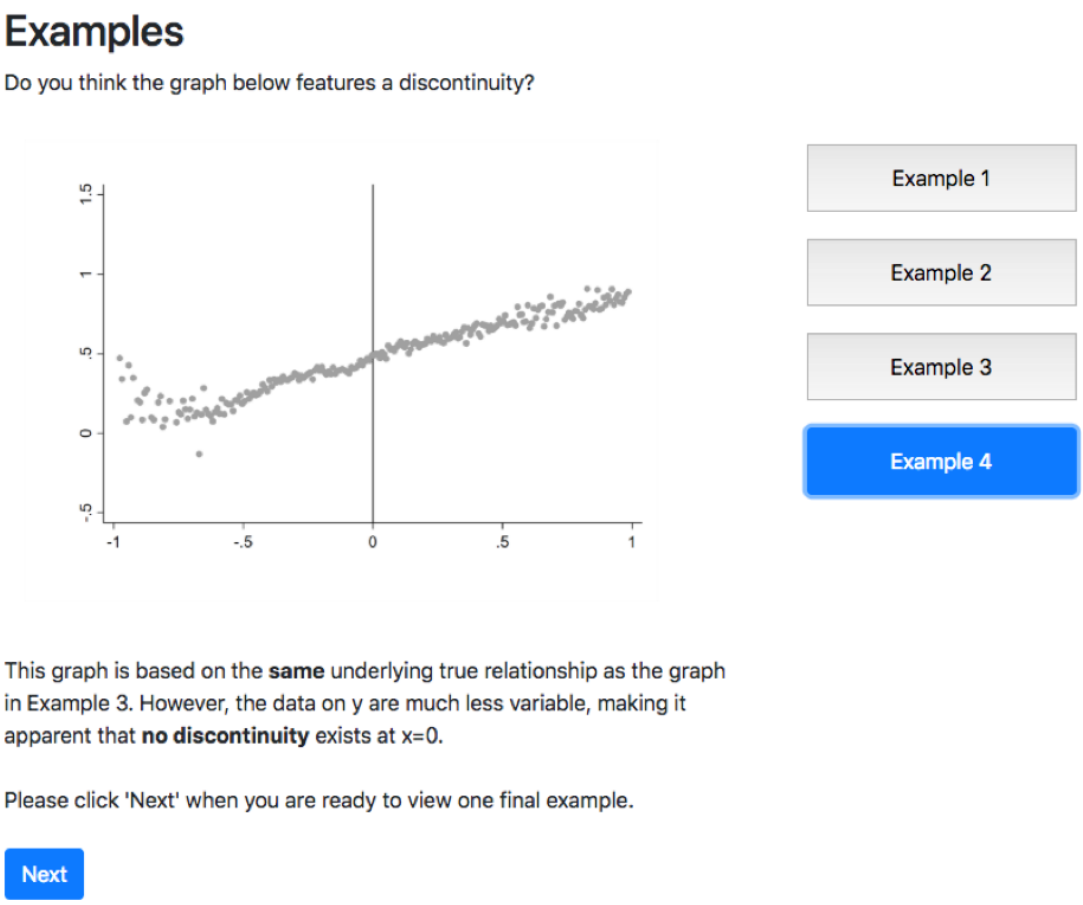


Figure B27: Classification Screen

(a) Incentives

Survey

In this survey, you will be asked to classify 11 graphs depending on whether you believe the underlying true relationship features a discontinuity at x=0 or not.

Please note that all, some or none of the 11 graphs you see in this survey may feature a discontinuity.

For each graph, you have two bonus options. You can select to be paid

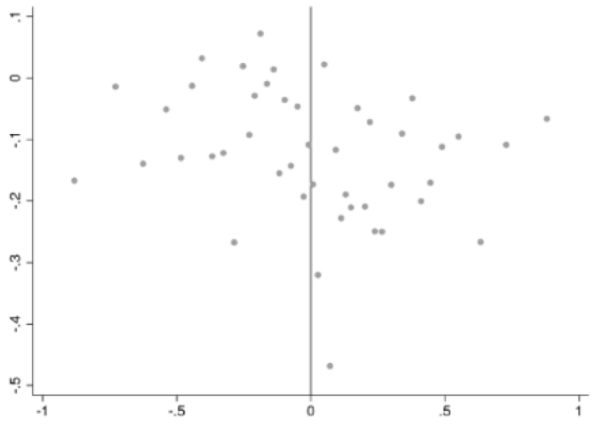
- $0.40 if you answer correctly and $0.00 otherwise, or
- $0.20 if you answer correctly and $0.20 otherwise.

Please click 'Next' to move on to the next page and begin the survey whenever you are ready.

Next

(b) Classification Task

Question 1 of 11



Do you think that there is a discontinuity in the true relationship at x=0?

Yes | No

Please make your bonus selection below:

Payment if Correct	Payment if Incorrect	Selection
$0.40	$0.00	☐
$0.20	$0.20	☐

Figure B28: Results Screen Example

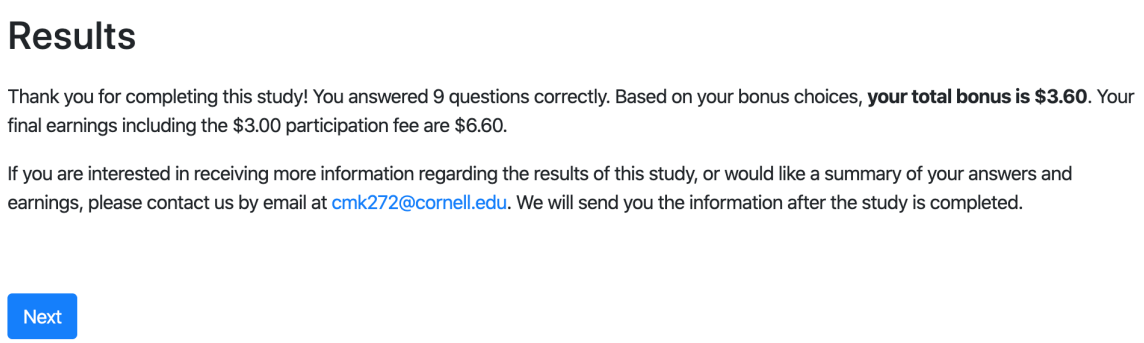


Figure B29: Sequence of Events - RDD Experiments

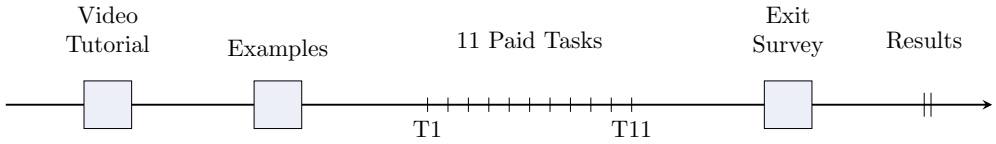
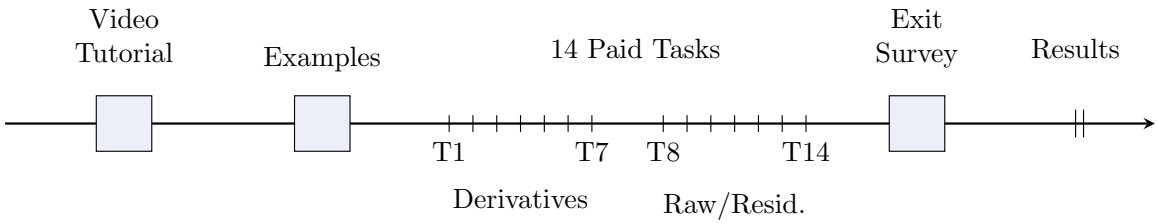


Figure B30: Sequence of Events - RKD Experiments

(a) Pilot RKD



(b) Phase RKD

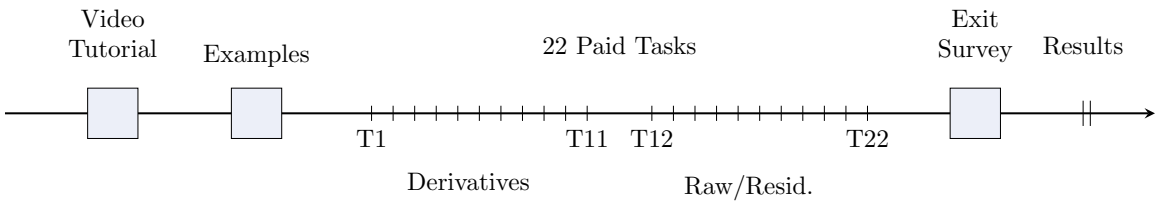


Figure B31: Sequence of Events - Expert Study

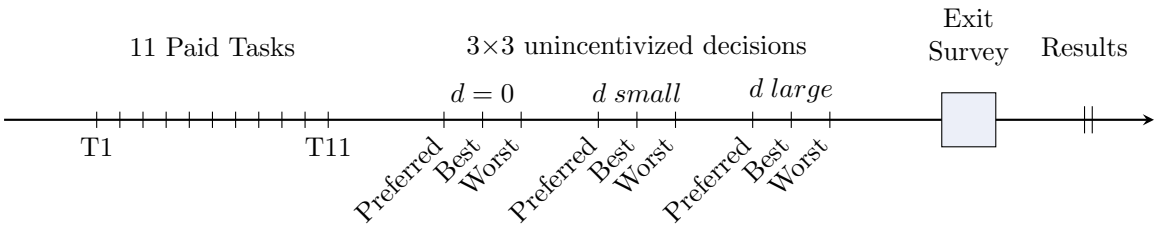


Figure B32: Expert Survey - Part 1 Introduction

Part 1

You are now ready to begin Part 1 of this survey.

In this part of the survey, we will present you with 11 graphs. You can navigate between graphs using a series of buttons at the bottom of the screen. For each graph, we will ask you to assess whether a discontinuity is present.

If you think that a graph exhibits a discontinuity, we will also ask you to determine the direction of the jump, and to visually estimate the size of the discontinuity with information on the y-axis.

Please note that you will not be able to move on to Part 2 until you have completed Part 1.

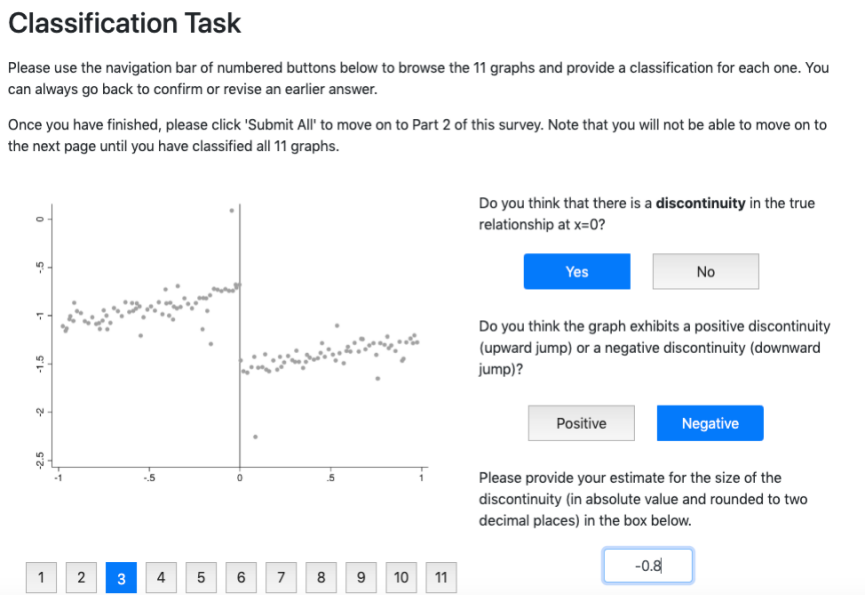
If you are one of the four randomly selected participants at the end of this study, you will earn a base fee of $450 plus $50 for every graph you classify correctly. To earn the bonus, you only need to correctly determine whether or not a discontinuity is present.

Please click 'Next' when you are ready to begin the classification task.

Next

Figure B33: Expert Survey - Part 1 Classification Task

(a) RD Case



(b) Expert Survey - Part 1 Task Screen (RK case)

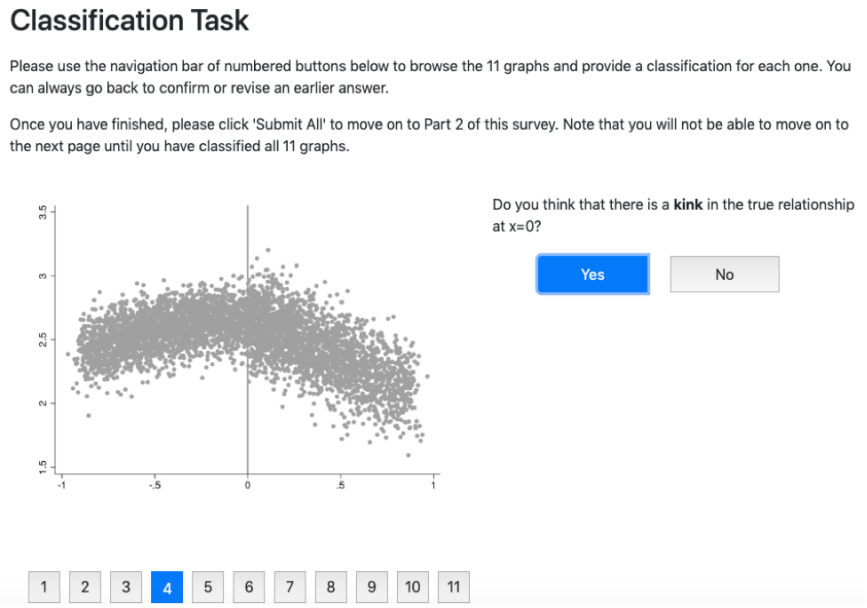
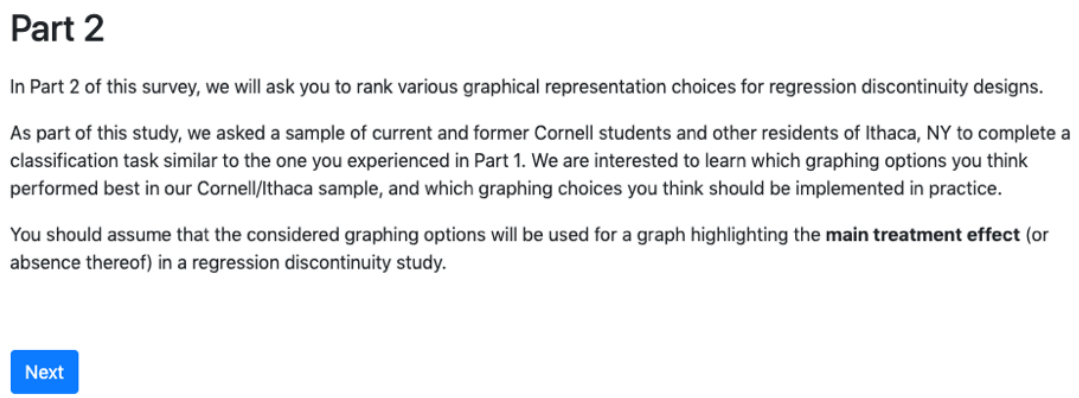


Figure B34: Expert Survey - Part 2 Instructions

(a) RD Case



(b) Expert Survey - Part 1 Task Screen (RK case)

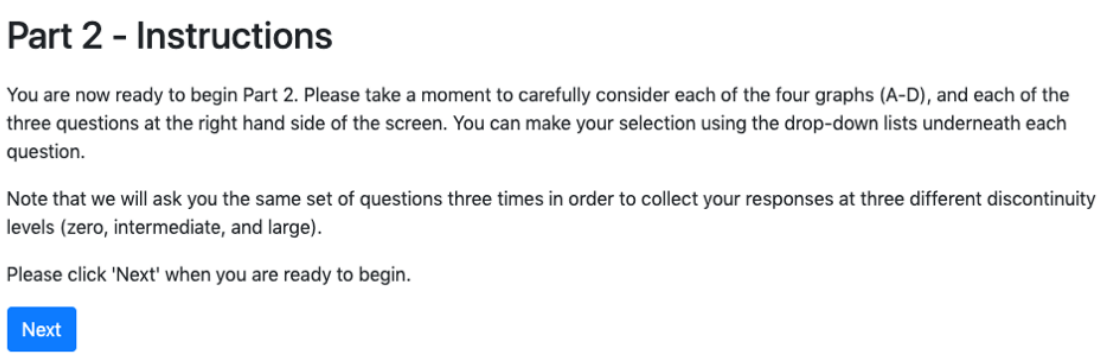


Figure B35: Expert Survey - Part 2 Decision Screen

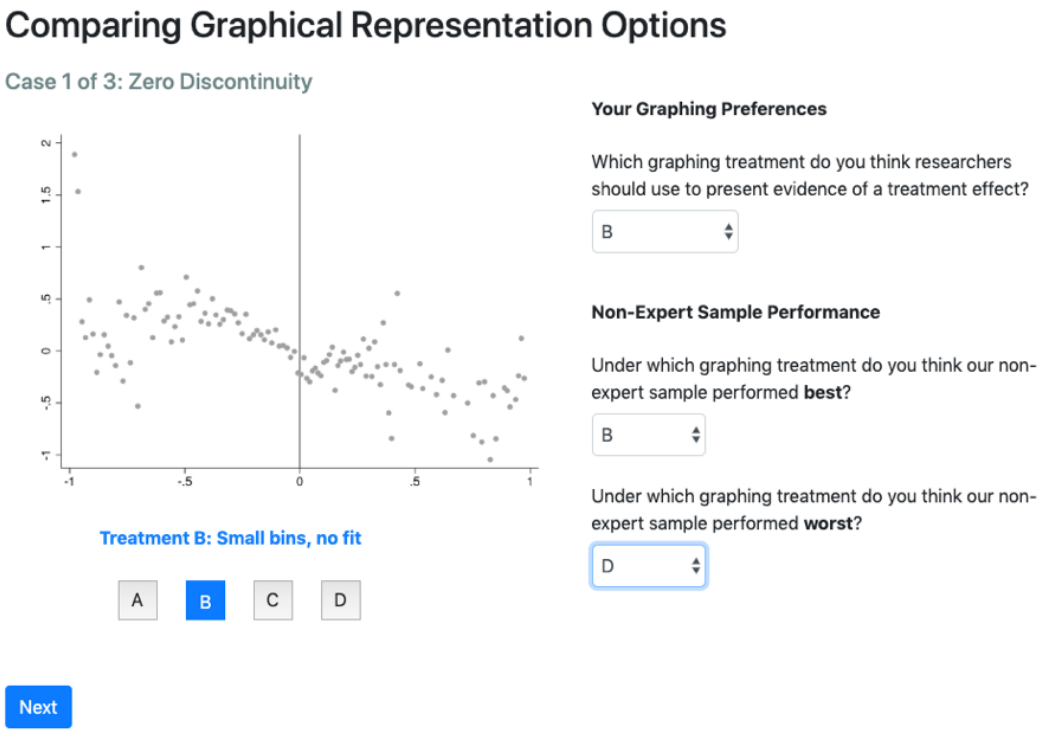
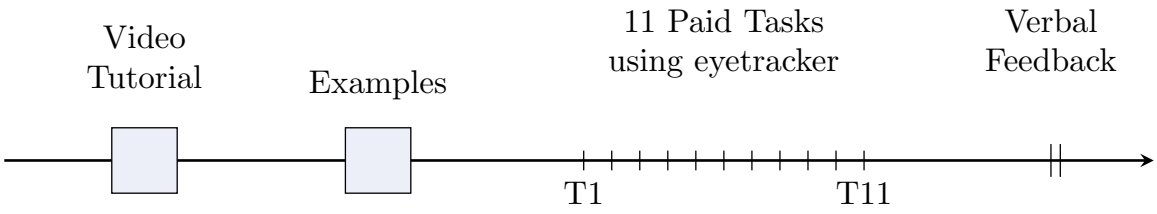


Figure B36: Sequence of Events - Eyetracking Study

Non-expert sample:



Expert sample:

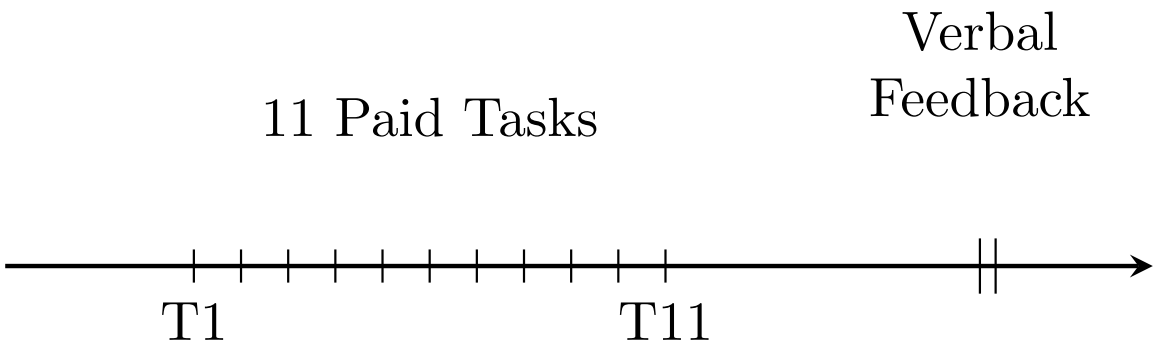


Figure B37: Sequence of Screens for Each Classification Task on Eyetracking Terminal

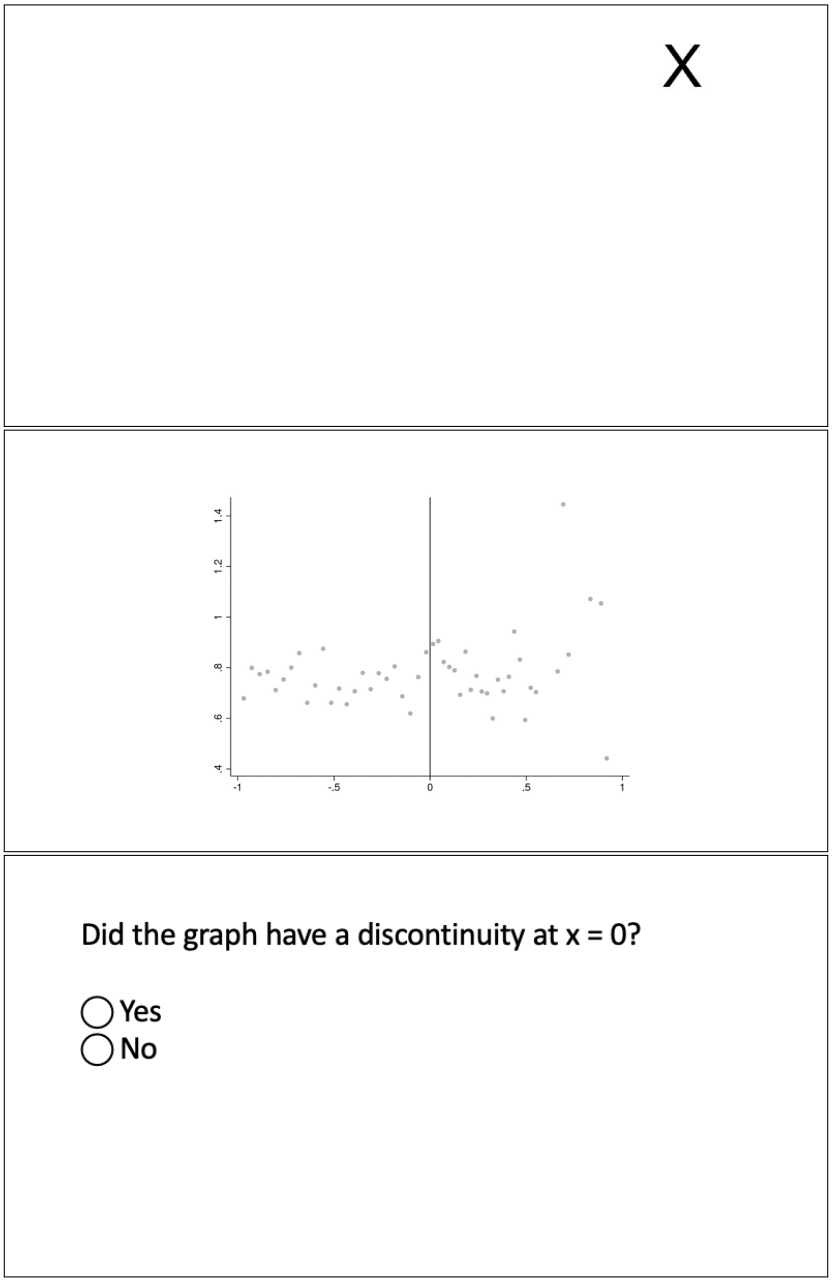
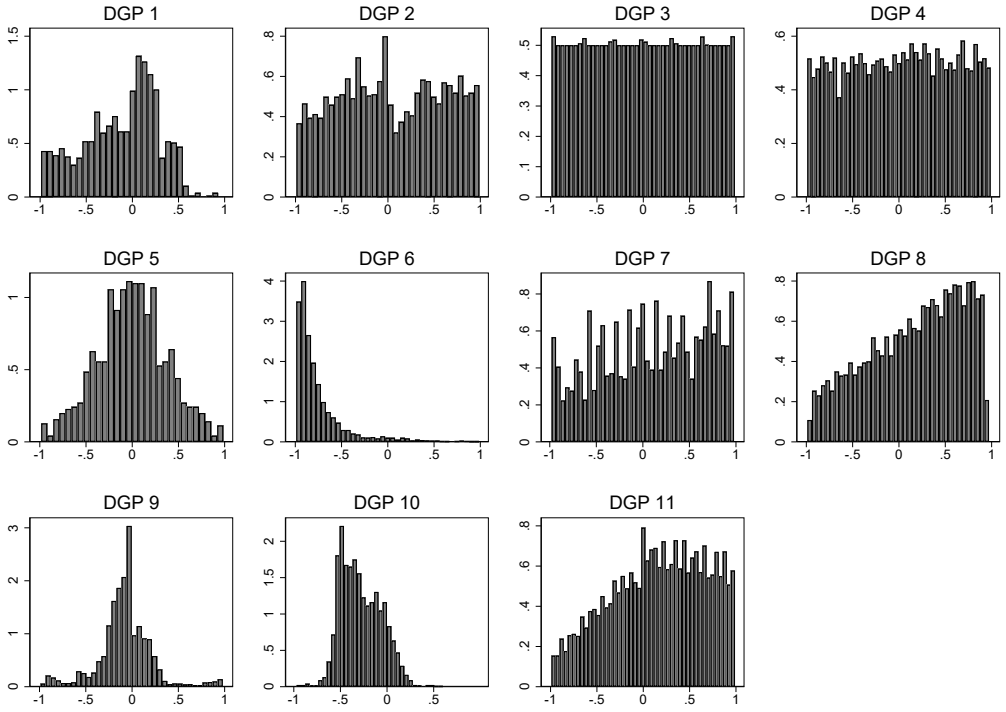
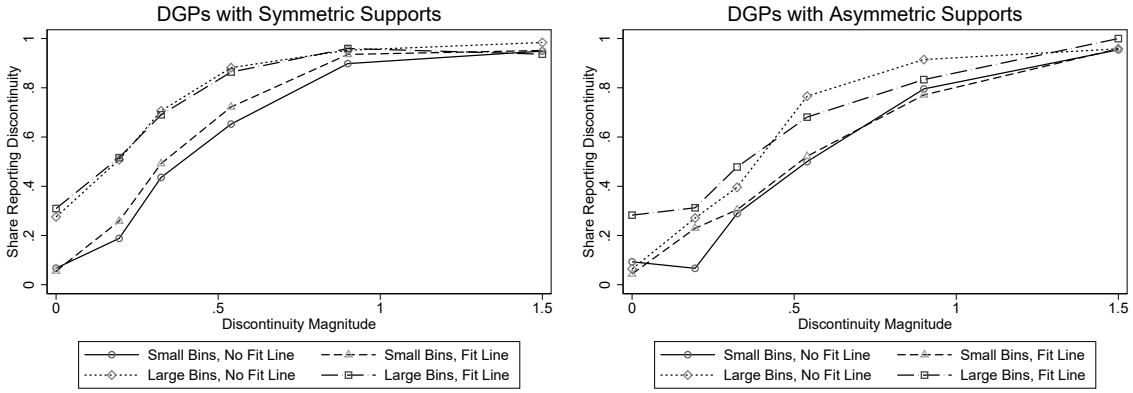


Figure B38: RDD DGP Histograms



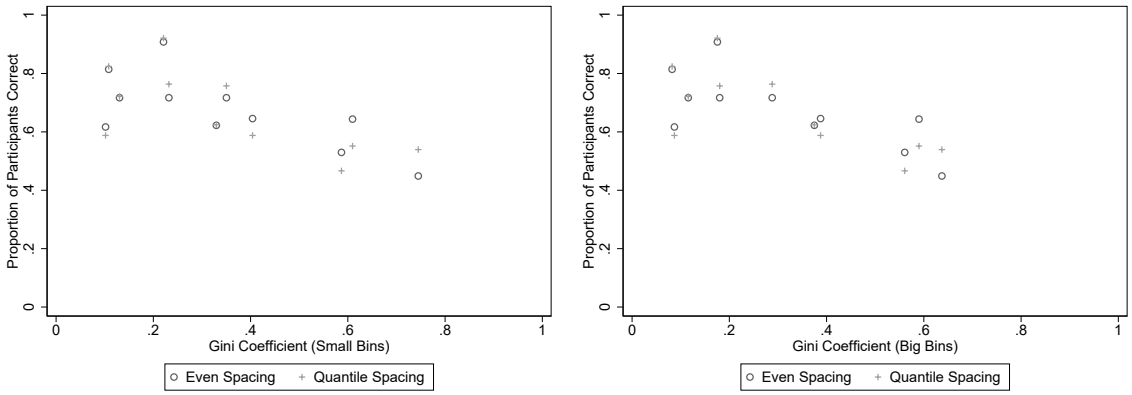
Notes: This figure presents histograms of the running variable for each RDD DGP.

Figure B39: Comparison between Power Functions for Phase RDD4 by Symmetry of DGP Supports



Notes: This figure plots the power functions from Phase RDD4 broken down into DGPs with symmetric supports prior to normalization and those with asymmetric supports prior to normalization.

Figure B40: DGP Gini Coefficients and Non-Expert Performance



Notes: This figure compares non-expert accuracy in classifying RDD graphs by DGP based on the uniformity of the distribution of the running variable as measured by the Gini coefficient. The left panel contains results for graphs with small bins, and the right panel contains results for graphs with big bins.

Figure B41: Discontinuity Easier to Detect When Both Slopes at Cutoff are in the Same Direction as the Discontinuity

(a) Positive Slopes and Positive Discontinuity

(b) Positive Slopes and Negative Discontinuity

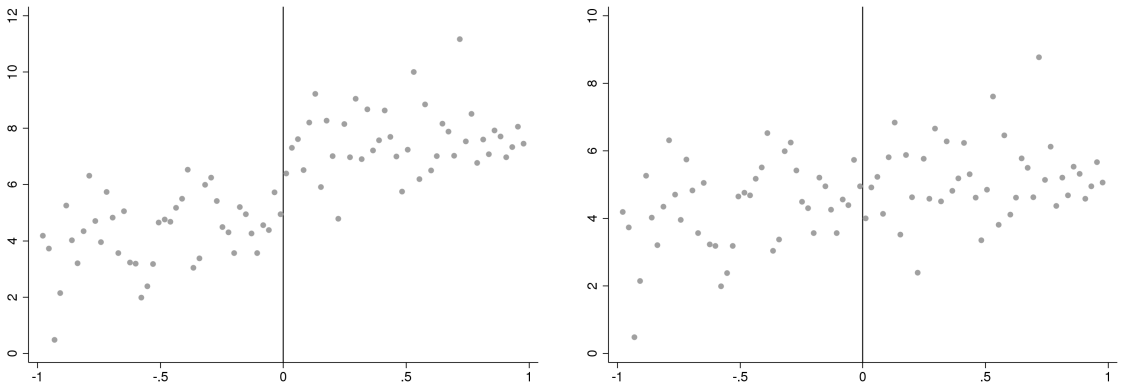
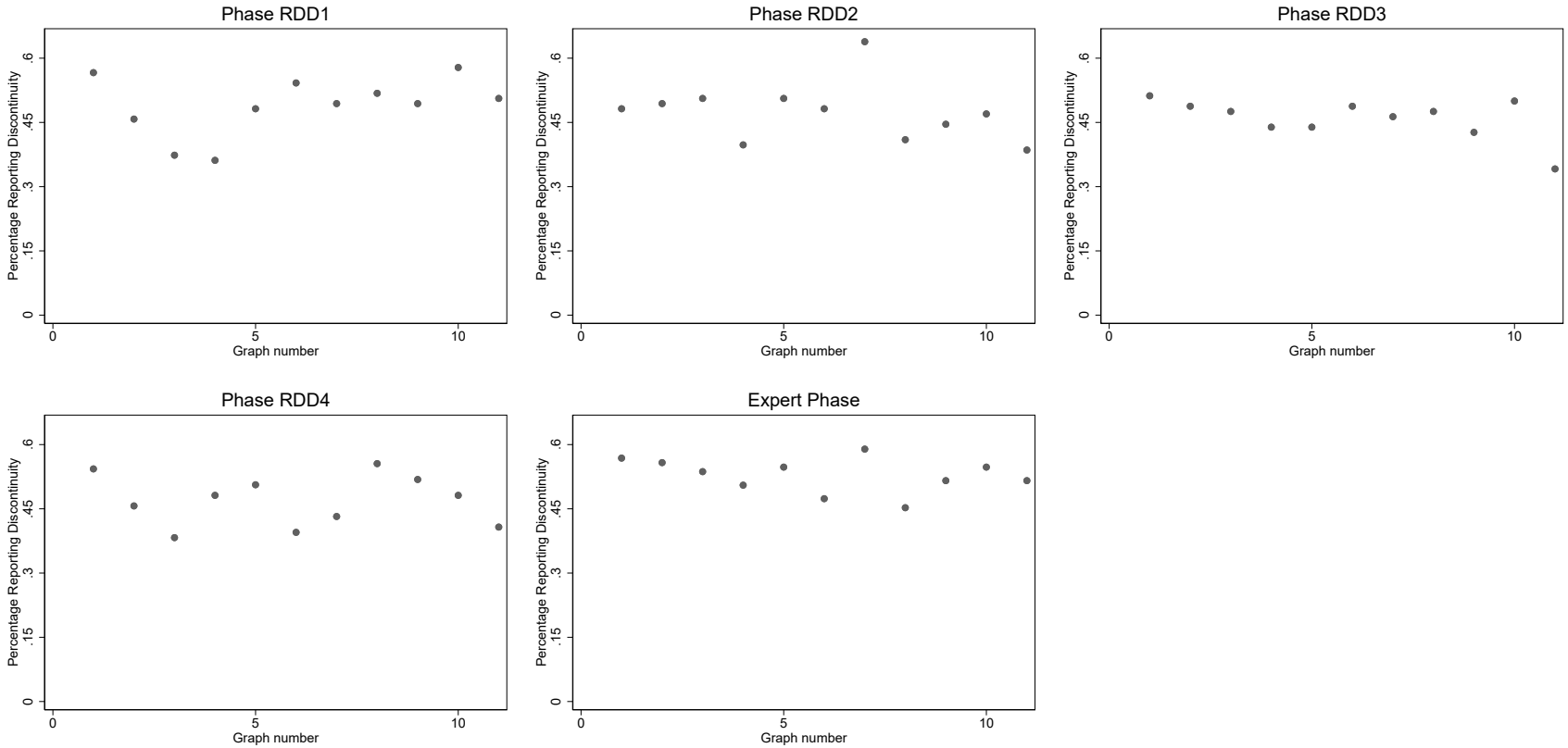
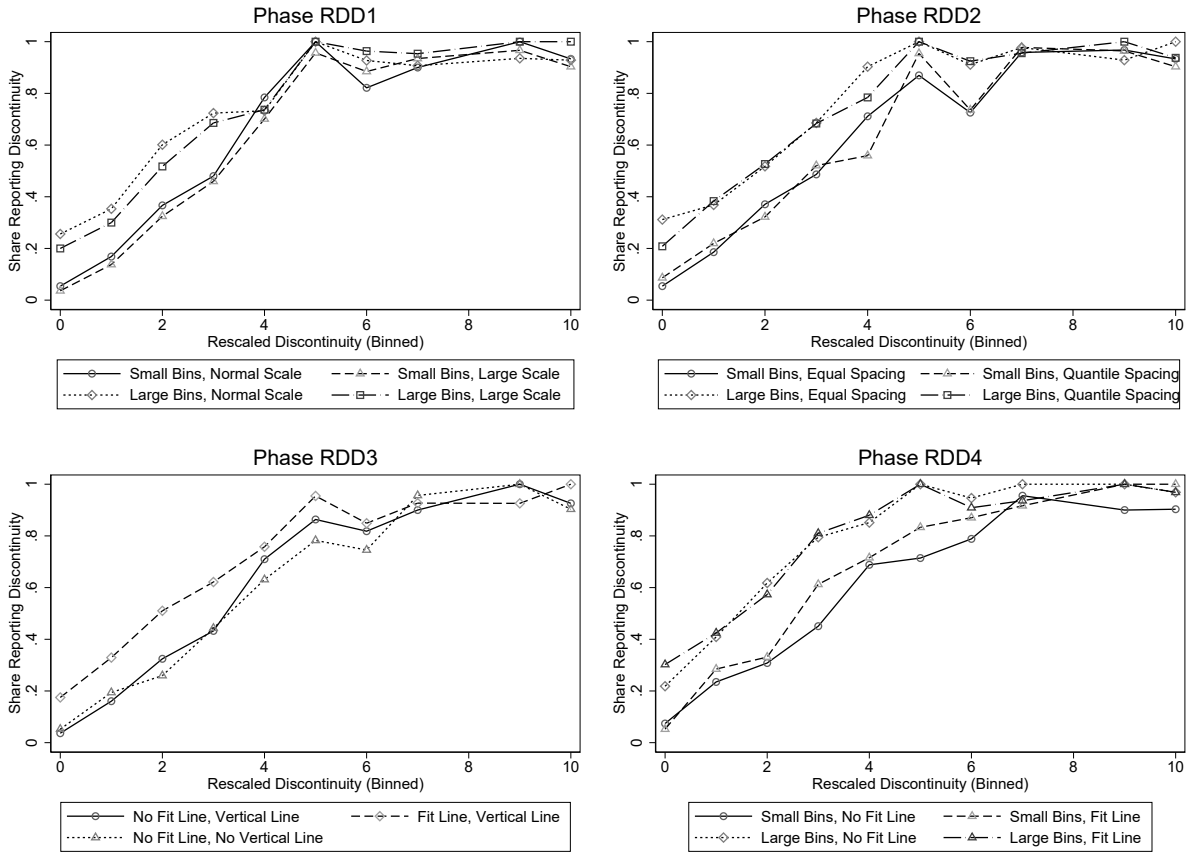


Figure B42: RDD Discontinuity Classifications over 11 Graphs



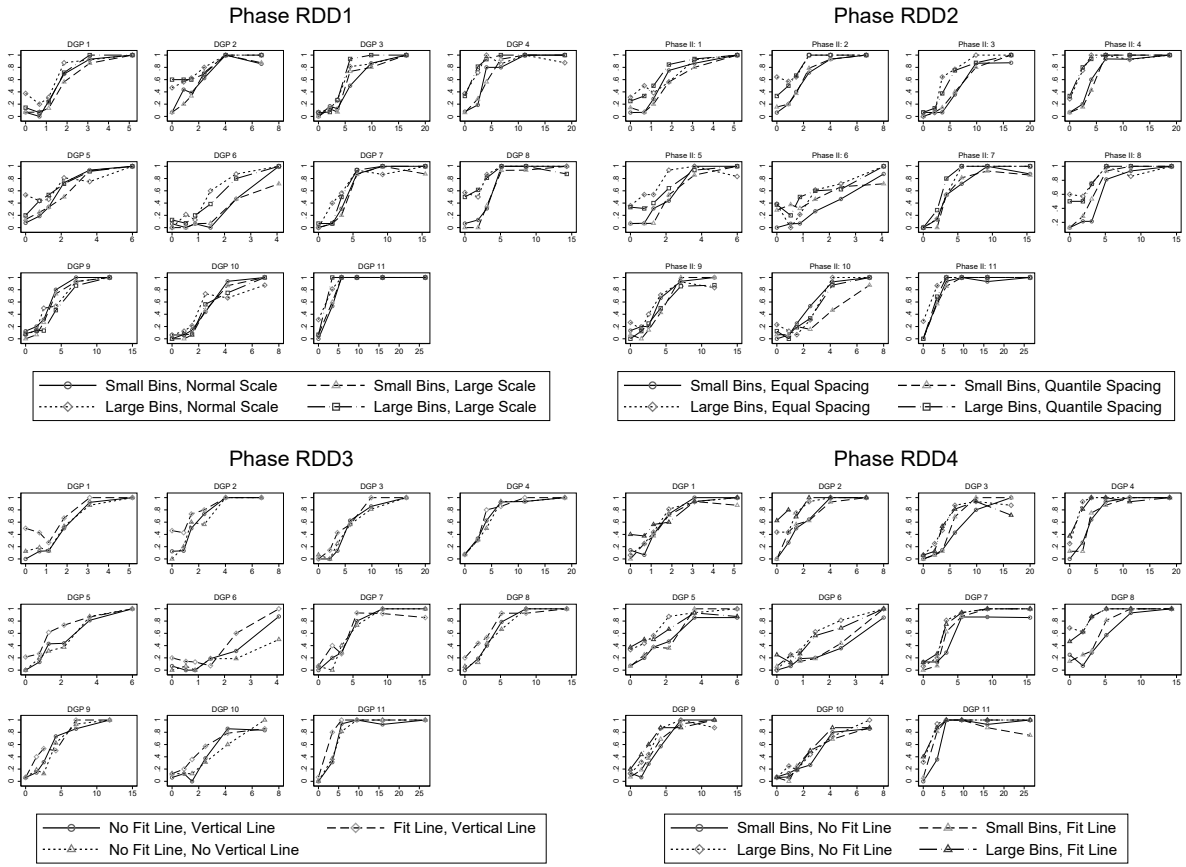
Notes: This figure plots participants' likelihood of reporting a discontinuity over the course of the 11 graphs seen.

Figure B43: RDD Phases: Power Functions against Asymptotic t-Statistics



Notes: Plotted are empirical power functions from the four non-expert RDD experiments. The power functions are defined in Section 2.2. The x-axis reflects binned values of the asymptotic t-statistic discussed in Appendix B.5. The y-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure B44: RDD Phases: Power Functions by DGP against Asymptotic t-Statistics



Notes: Plotted are empirical power functions from the four non-expert RDD experiments broken down by DGP per phase. The power functions are defined in Section 2.2. The x-axis represents binned values of the asymptotic t-statistic discussed in Appendix B.5. The y-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure B45: Visual Representation of Eyetracking Data

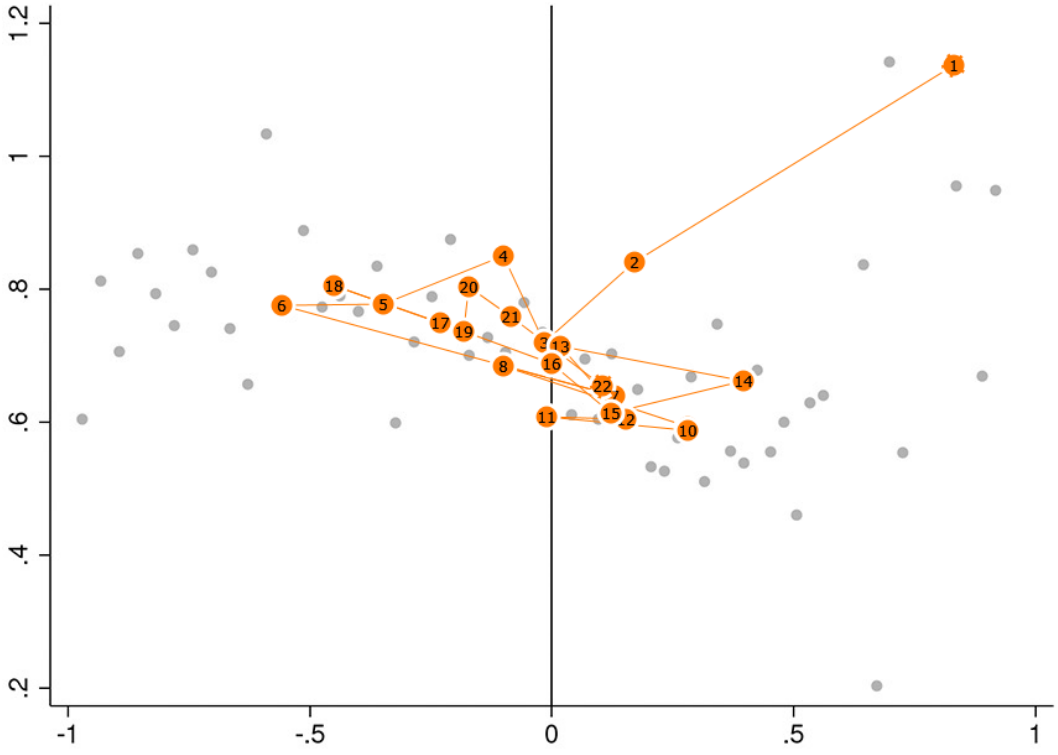
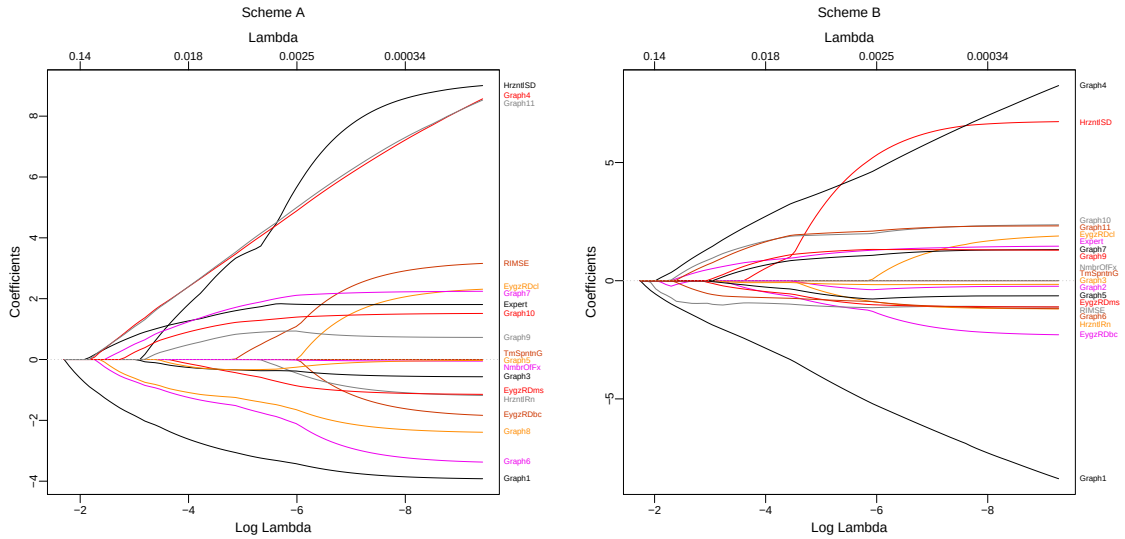
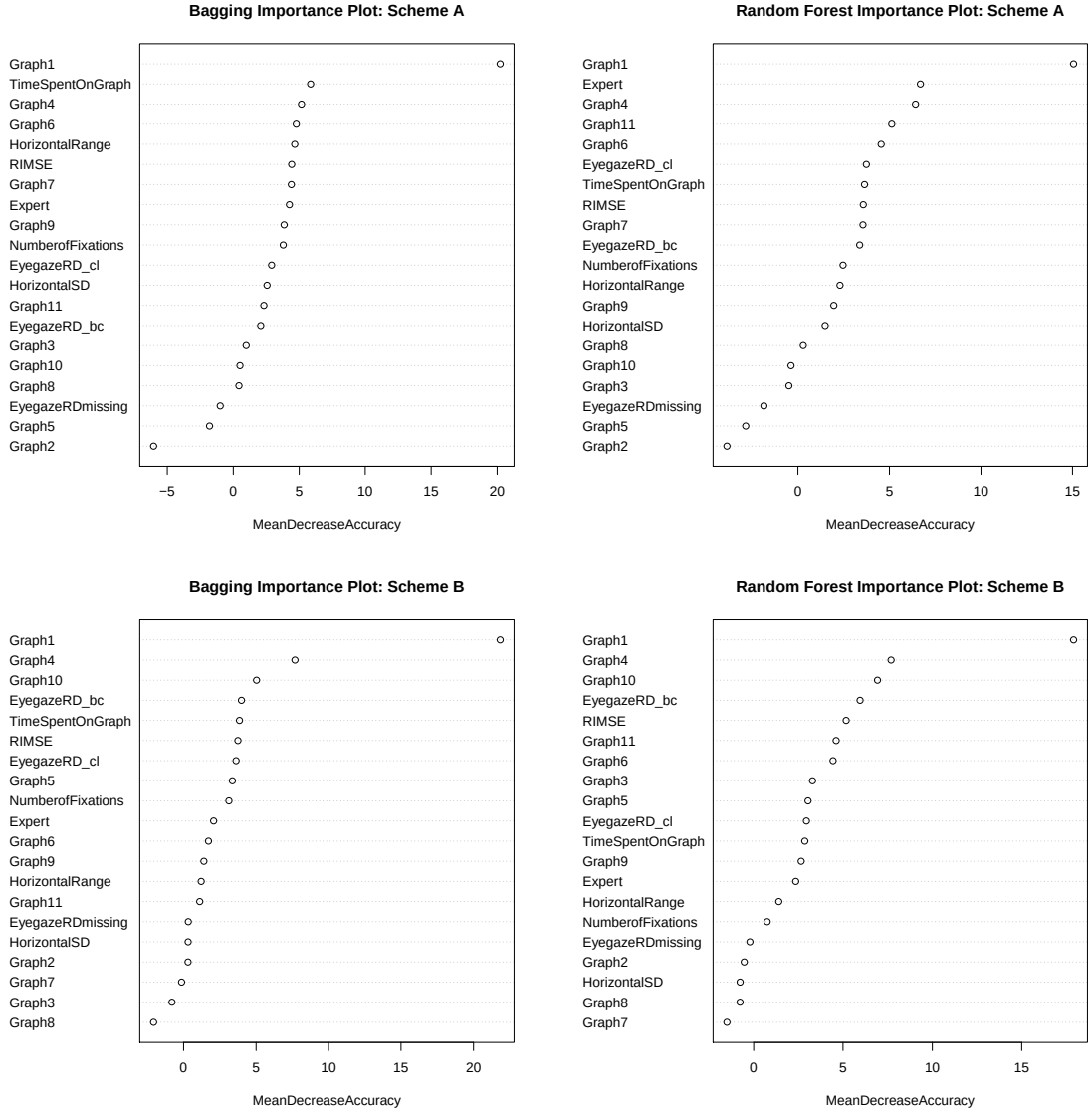


Figure B46: Penalized Logit Coefficient Path Plots



Notes: For the abbreviated variables: "TmSpntnG" reflects the time spent per graph in seconds; "NmbrOfFx" is the number of fixations per graph; "HrzntlRn" is the horizontal range of eyegazes; "HrzntlSD" is the horizontal standard deviation of eyegazes; "EygzRDcl" is the absolute deviation of the conventional local linear RD estimates from the true discontinuity (scaled by σ); "EygzRDbc" is the absolute deviation of bias-corrected local linear RD estimates from the true discontinuity; and "EygzRDms" reflects whether the discontinuity estimate is unavailable.

Figure B47: Bagging and Random Forest Importance Plots



Notes: For the abbreviated variables: "EyegazeRD_cl" reflects the absolute deviation of conventional local linear RD estimates from the true discontinuity (scaled by σ); "EyegazeRD_bc" is the absolute deviation of bias-corrected local linear RD estimates from the true discontinuity; and "EyegazeRDmissing" reflects whether the discontinuity estimate is unavailable.

Appendix C

Appendix for Chapter 3

C.1 Additional Estimates

Table C1: Calloway and Sant'Anna Difference-in-Differences Estimates, Domestic Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	0.002	-0.004	0.072	0.087	0.000	0.003	0.027	-0.021
	(0.003)	(0.002)	(0.096)	(0.123)	(0.003)	(0.002)	(0.067)	(0.068)
t=-1	0.001	0.000	0.007	0.028	0.004	0.000	0.020	-0.014
	(0.001)	(0.001)	(0.031)	(0.044)	(0.001)	(0.002)	(0.044)	(0.024)
t=0	0.002	0.004	-0.042	0.017	0.001	0.004	-0.033	-0.027
	(0.002)	(0.002)	(0.042)	(0.048)	(0.001)	(0.001)	(0.061)	(0.046)
t=1	0.006	0.003	-0.037	0.064	-0.002	0.001	-0.104	0.037
	(0.003)	(0.003)	(0.037)	(0.098)	(0.002)	(0.002)	(0.048)	(0.037)
t=2	0.005	0.000	-0.028	-0.037	-0.003	-0.001	-0.051	0.009
	(0.003)	(0.002)	(0.037)	(0.038)	(0.003)	(0.002)	(0.050)	(0.019)
t=3	0.005	-0.001	0.000	-0.011	0.001	0.005	-0.040	0.004
	(0.003)	(0.002)	(0.043)	(0.032)	(0.003)	(0.003)	(0.045)	(0.033)
Clusters	53	53	53	52	338	338	337	337

Notes: Table reports matched differences-in-differences estimates based on the group-time average treatment effect and event study aggregation scheme from Callaway and Sant'Anna (2020). Period $t = 0$ refers to the year in which the local minimum wage was passed. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers).

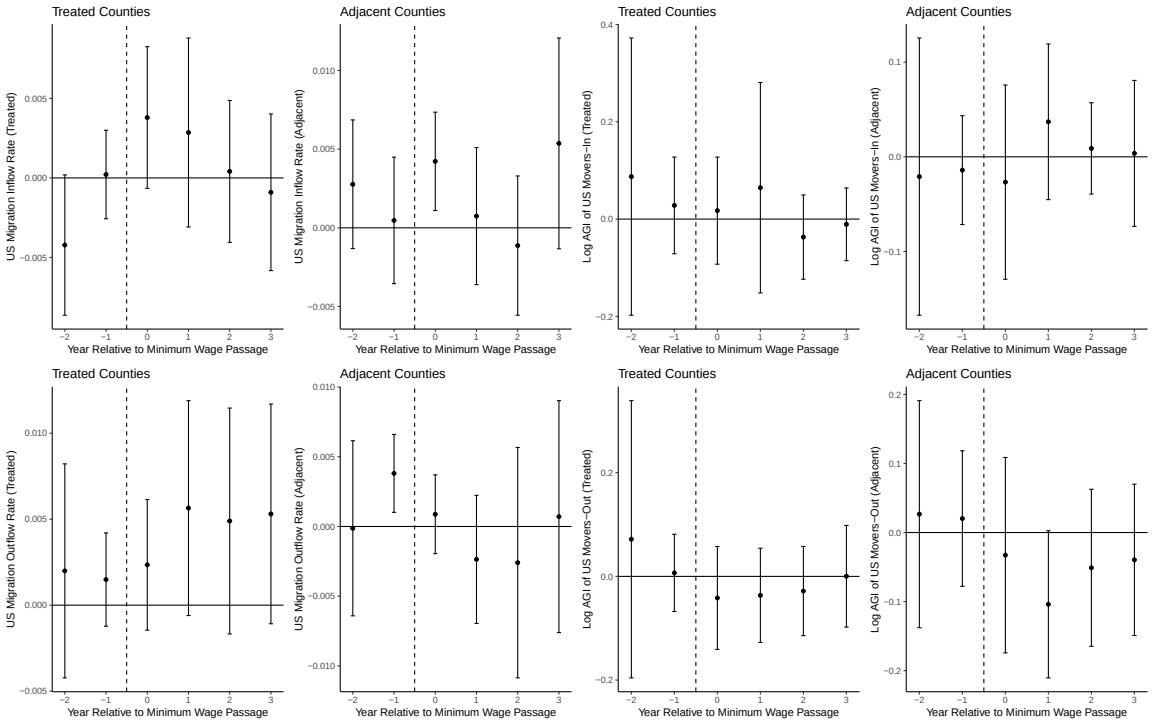


Figure C1: CS DiD Estimates: Domestic

Notes: Figure presents CS difference-in-differences estimates.

C.2 TWFE Estimates

Table C2: Matched Difference-in-Differences Estimates, Total Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.002	−0.000	−0.012	−0.045	−0.003	−0.002	0.023	0.013
	(0.001)	(0.001)	(0.036)	(0.045)	(0.002)	(0.002)	(0.031)	(0.024)
t=0	0.002	0.004	−0.049	0.003	0.002	0.003	0.012	−0.023
	(0.002)	(0.002)	(0.046)	(0.041)	(0.001)	(0.001)	(0.056)	(0.044)
t=1	0.006	0.003	−0.048	0.056	−0.001	−0.001	−0.064	0.041
	(0.003)	(0.002)	(0.045)	(0.096)	(0.002)	(0.002)	(0.034)	(0.043)
t=2	0.005	0.001	−0.033	−0.056	−0.002	−0.003	−0.009	0.008
	(0.003)	(0.002)	(0.039)	(0.036)	(0.003)	(0.001)	(0.030)	(0.023)
t=3	0.005	−0.001	−0.003	−0.025	0.002	0.004	0.002	0.003
	(0.003)	(0.002)	(0.049)	(0.027)	(0.003)	(0.002)	(0.032)	(0.034)
R^2	0.943	0.942	0.900	0.918	0.941	0.919	0.721	0.903
Clusters	53	53	53	53	296	296	296	296

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C3: Matched Difference-in-Differences Estimates, Domestic Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.001	−0.000	−0.008	−0.033	−0.003	−0.002	0.021	0.013
	(0.001)	(0.001)	(0.035)	(0.044)	(0.002)	(0.002)	(0.031)	(0.024)
t=0	0.002	0.004	−0.041	0.019	0.002	0.003	0.010	−0.026
	(0.002)	(0.002)	(0.044)	(0.048)	(0.001)	(0.001)	(0.057)	(0.043)
t=1	0.006	0.003	−0.036	0.064	−0.001	−0.001	−0.063	0.038
	(0.003)	(0.002)	(0.044)	(0.101)	(0.002)	(0.002)	(0.034)	(0.043)
t=2	0.005	0.001	−0.029	−0.041	−0.001	−0.003	−0.009	0.009
	(0.003)	(0.002)	(0.039)	(0.040)	(0.003)	(0.001)	(0.030)	(0.023)
t=3	0.005	−0.001	−0.001	−0.011	0.002	0.004	0.002	0.004
	(0.003)	(0.002)	(0.050)	(0.030)	(0.003)	(0.002)	(0.032)	(0.034)
R^2	0.943	0.940	0.909	0.917	0.938	0.914	0.719	0.902
Clusters	53	53	53	53	296	296	296	296

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C4: Matched Difference-in-Differences Estimates, Foreign Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.000	−0.000	−0.093	−0.312	−0.000	−0.000	0.115	−0.108
	(0.000)	(0.000)	(0.173)	(0.339)	(0.000)	(0.000)	(0.128)	(0.076)
t=0	0.000	0.000	−0.343	−0.721	0.000	0.000	0.051	−0.035
	(0.000)	(0.000)	(0.208)	(0.487)	(0.000)	(0.000)	(0.152)	(0.195)
t=1	−0.000	0.000	−0.424	−0.265	0.001	−0.000	−0.061	−0.074
	(0.000)	(0.000)	(0.291)	(0.134)	(0.001)	(0.000)	(0.195)	(0.158)
t=2	−0.000	0.000	−0.346	−0.580	−0.001	−0.000	−0.229	−0.358
	(0.000)	(0.000)	(0.137)	(0.325)	(0.000)	(0.000)	(0.101)	(0.128)
t=3	−0.000	0.000	0.210	−0.110	−0.000	−0.000	0.063	−0.023
	(0.000)	(0.000)	(0.218)	(0.482)	(0.000)	(0.000)	(0.128)	(0.132)
R^2	0.812	0.910	0.638	0.714	0.745	0.883	0.565	0.640
Clusters	36	35	36	35	143	135	143	135

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C5: Matched Difference-in-Differences Estimates, Same State Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.001	−0.000	0.034	−0.005	−0.001	−0.002	0.045	−0.019
	(0.001)	(0.001)	(0.026)	(0.071)	(0.001)	(0.001)	(0.027)	(0.034)
t=0	0.001	0.001	−0.042	0.040	0.001	0.001	0.001	−0.022
	(0.002)	(0.001)	(0.024)	(0.027)	(0.001)	(0.001)	(0.024)	(0.040)
t=1	0.004	0.002	−0.030	0.095	0.000	−0.001	−0.020	0.042
	(0.002)	(0.002)	(0.030)	(0.161)	(0.001)	(0.001)	(0.028)	(0.053)
t=2	0.001	−0.000	0.012	−0.074	−0.001	−0.002	0.014	0.008
	(0.002)	(0.001)	(0.040)	(0.050)	(0.002)	(0.001)	(0.020)	(0.029)
t=3	−0.000	−0.003	0.016	−0.005	−0.001	0.001	0.014	−0.004
	(0.002)	(0.001)	(0.047)	(0.050)	(0.001)	(0.001)	(0.021)	(0.035)
R^2	0.950	0.934	0.880	0.854	0.945	0.901	0.833	0.889
Clusters	53	53	53	53	296	296	296	296

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C6: Matched Difference-in-Differences Estimates, Different State Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.001	0.000	0.001	−0.024	−0.002	−0.000	0.002	0.043
	(0.001)	(0.001)	(0.068)	(0.034)	(0.001)	(0.001)	(0.051)	(0.039)
t=0	0.001	0.002	0.039	−0.021	0.001	0.002	0.039	−0.005
	(0.001)	(0.002)	(0.080)	(0.097)	(0.001)	(0.001)	(0.136)	(0.037)
t=1	0.002	−0.000	−0.007	−0.015	−0.001	−0.001	−0.080	0.018
	(0.002)	(0.001)	(0.039)	(0.044)	(0.001)	(0.001)	(0.052)	(0.035)
t=2	0.004	0.001	−0.035	−0.017	0.001	−0.002	−0.034	−0.018
	(0.002)	(0.002)	(0.067)	(0.058)	(0.001)	(0.001)	(0.052)	(0.038)
t=3	0.005	0.002	0.002	−0.004	0.002	0.002	0.017	0.011
	(0.002)	(0.002)	(0.055)	(0.055)	(0.002)	(0.001)	(0.054)	(0.040)
R^2	0.926	0.934	0.893	0.836	0.957	0.952	0.806	0.809
Clusters	53	53	53	53	295	295	295	295

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C7: Matched Difference-in-Differences Estimates, Neighboring Counties Migration

	Treated Counties				Adjacent Counties			
	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In	Outflow Rate	Inflow Rate	Log AGI Out	Log AGI In
t=-2	−0.002	−0.001	0.047	−0.012	−0.001	−0.003	0.040	−0.026
	(0.002)	(0.001)	(0.042)	(0.069)	(0.001)	(0.001)	(0.036)	(0.048)
t=0	0.001	0.002	−0.061	−0.000	0.001	0.001	0.015	−0.046
	(0.001)	(0.001)	(0.018)	(0.033)	(0.001)	(0.001)	(0.026)	(0.044)
t=1	0.002	−0.000	−0.027	0.014	−0.000	−0.001	−0.009	−0.053
	(0.001)	(0.002)	(0.027)	(0.068)	(0.001)	(0.001)	(0.042)	(0.045)
t=2	0.001	−0.001	−0.013	−0.018	−0.001	−0.001	−0.006	−0.015
	(0.001)	(0.001)	(0.043)	(0.024)	(0.001)	(0.001)	(0.024)	(0.038)
t=3	0.002	0.001	0.029	0.030	0.001	0.003	0.020	−0.009
	(0.002)	(0.001)	(0.047)	(0.046)	(0.001)	(0.001)	(0.019)	(0.036)
R^2	0.943	0.959	0.914	0.815	0.919	0.865	0.849	0.865
Clusters	53	53	53	53	292	292	292	292

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Migration events are based on IRS tax filings and approximate the universe of tax-filing households. Denominators for migration rates are based on the IRS' annual filings totals for each county (non-moving filers plus moving-out filers). Standard errors are clustered at the county level.

Table C8: Additional Matched Difference-in-Differences Estimates

	Treated Counties			Adjacent Counties		
	Log Population	Non-Mover Log AGI	Log Home Value	Log Population	Non-Mover Log AGI	Log Home Value
t=-2	−0.028	0.050	−0.027	−0.028	0.050	−0.027
	(0.015)	(0.023)	(0.035)	(0.015)	(0.023)	(0.035)
t=0	0.016	0.006	0.031	0.016	0.006	0.031
	(0.005)	(0.020)	(0.020)	(0.005)	(0.020)	(0.020)
t=1	0.030	−0.021	0.052	0.030	−0.021	0.052
	(0.010)	(0.034)	(0.039)	(0.010)	(0.034)	(0.039)
t=2	0.033	0.009	0.062	0.033	0.009	0.062
	(0.014)	(0.036)	(0.062)	(0.014)	(0.036)	(0.062)
t=3	0.033	0.025	0.033	0.033	0.025	0.033
	(0.020)	(0.039)	(0.061)	(0.020)	(0.039)	(0.061)
R^2	1.000	0.980	0.993	1.000	0.980	0.993
N Clusters	53	53	52	53	53	52

Notes: Table reports matched difference-in-differences estimates following the model in Equation 3.1. Event time is based on the passage of the local minimum wage, with the final pre-period being omitted and periods outside of the balanced support binned and unreported. Standard errors are clustered at the county level. Home values come from the Zillow Home Value Index.

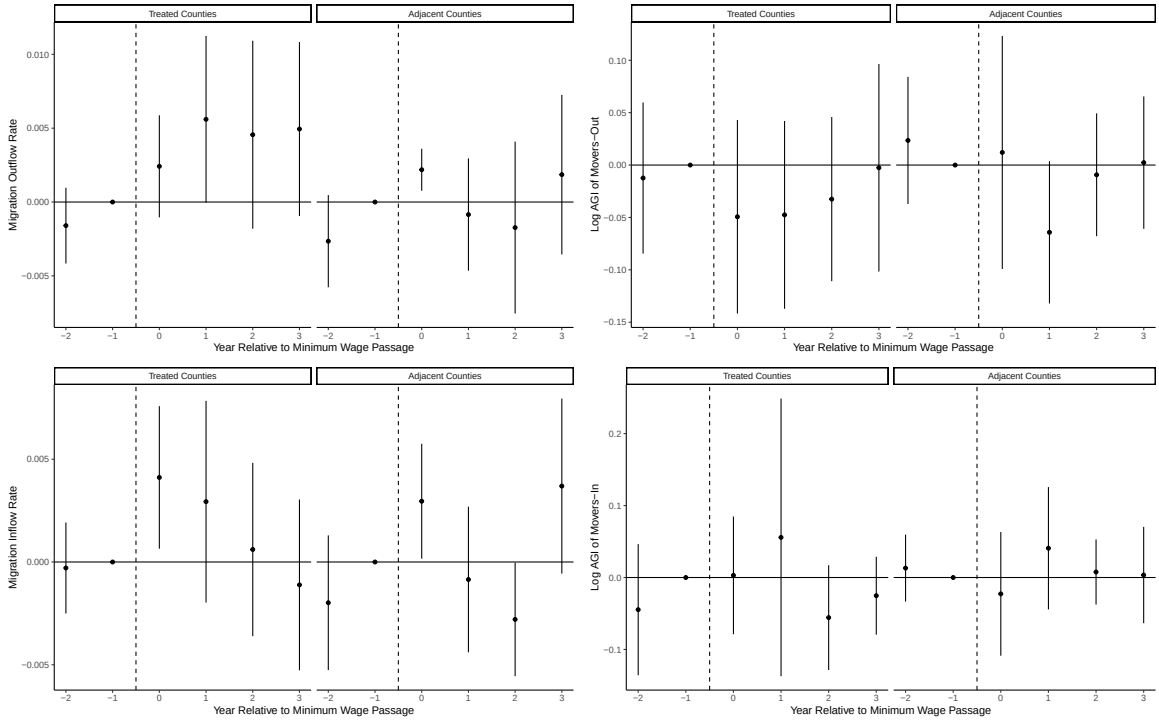


Figure C2: Matched DiD Estimates: Totals

Notes: Figure presents matched difference-in-differences estimates.

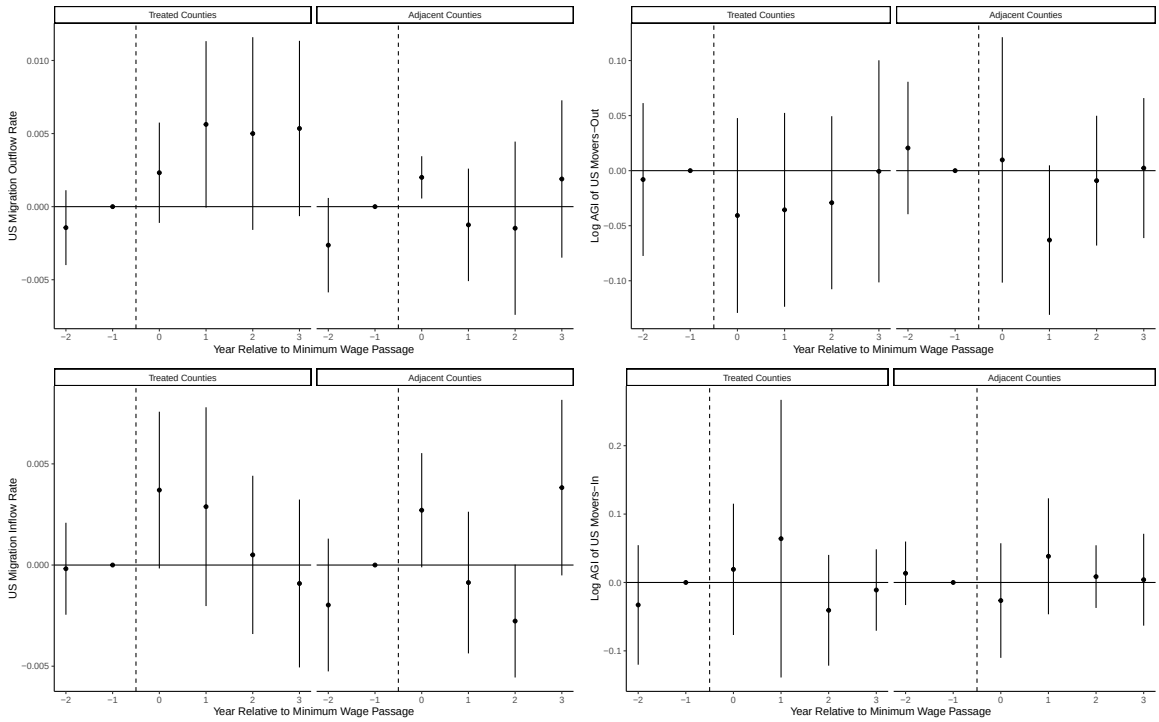


Figure C3: Matched DiD Estimates: Domestic

Notes: Figure presents matched difference-in-differences estimates.

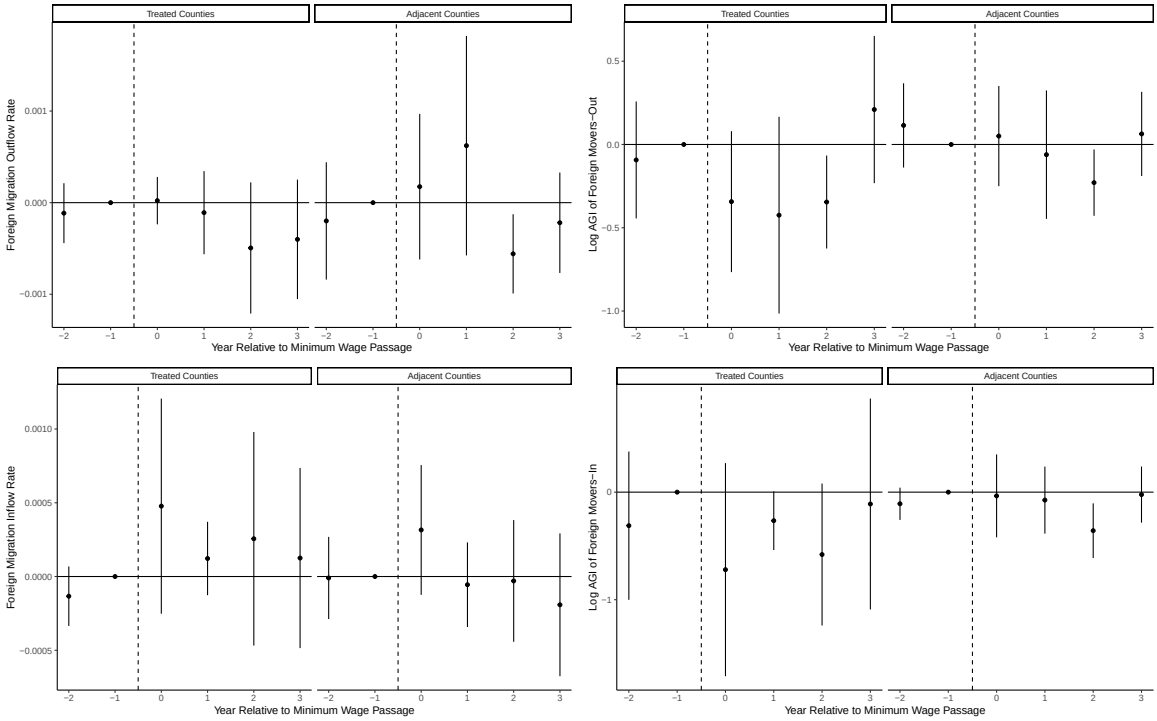


Figure C4: Matched DiD Estimates: Foreign

Notes: Figure presents matched difference-in-differences estimates.

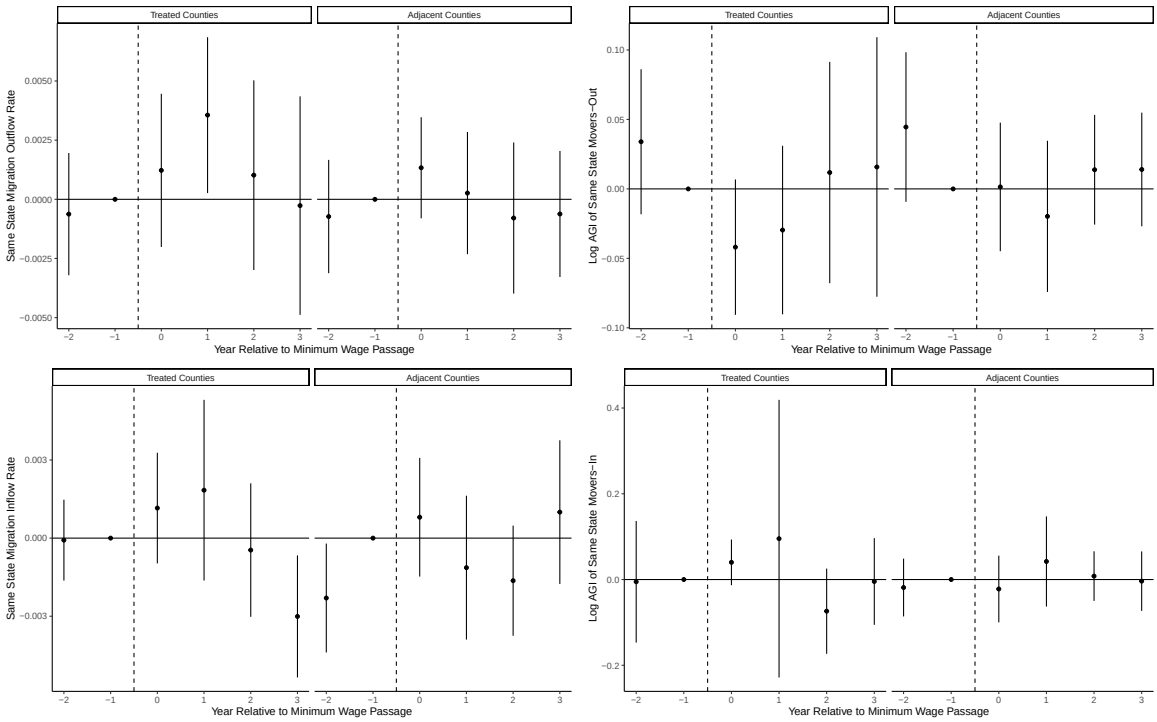


Figure C5: Matched DiD Estimates: Same State

Notes: Figure presents matched difference-in-differences estimates.

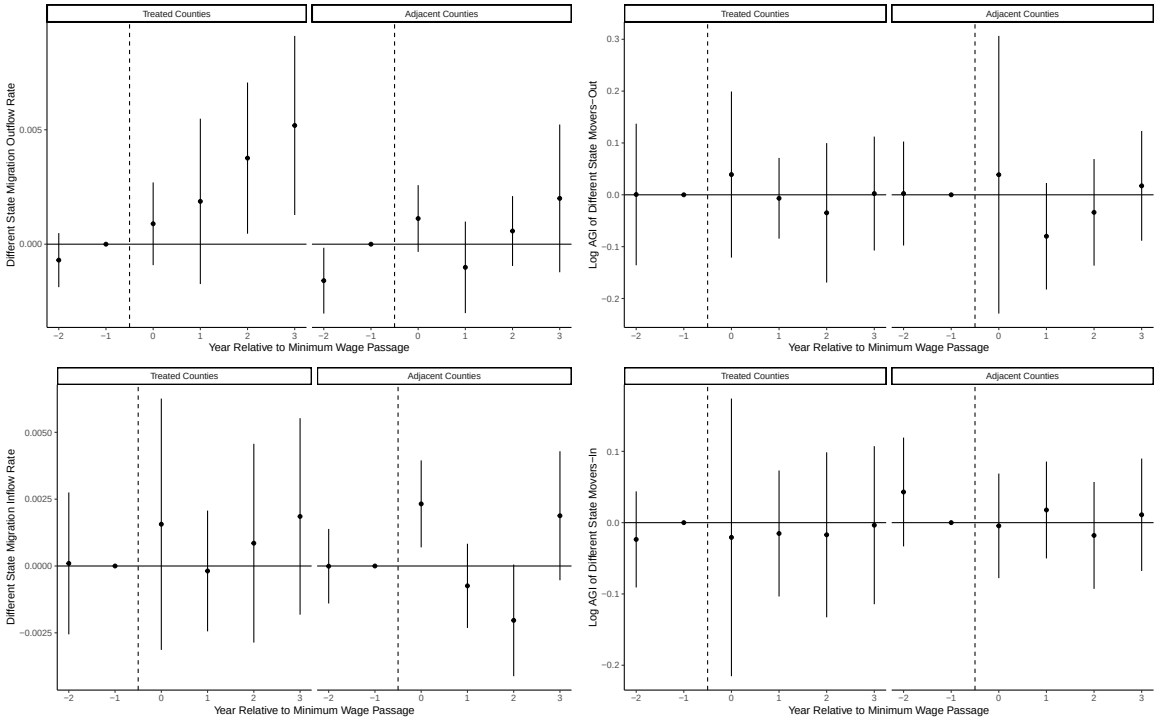


Figure C6: Matched DiD Estimates: Different State

Notes: Figure presents matched difference-in-differences estimates.

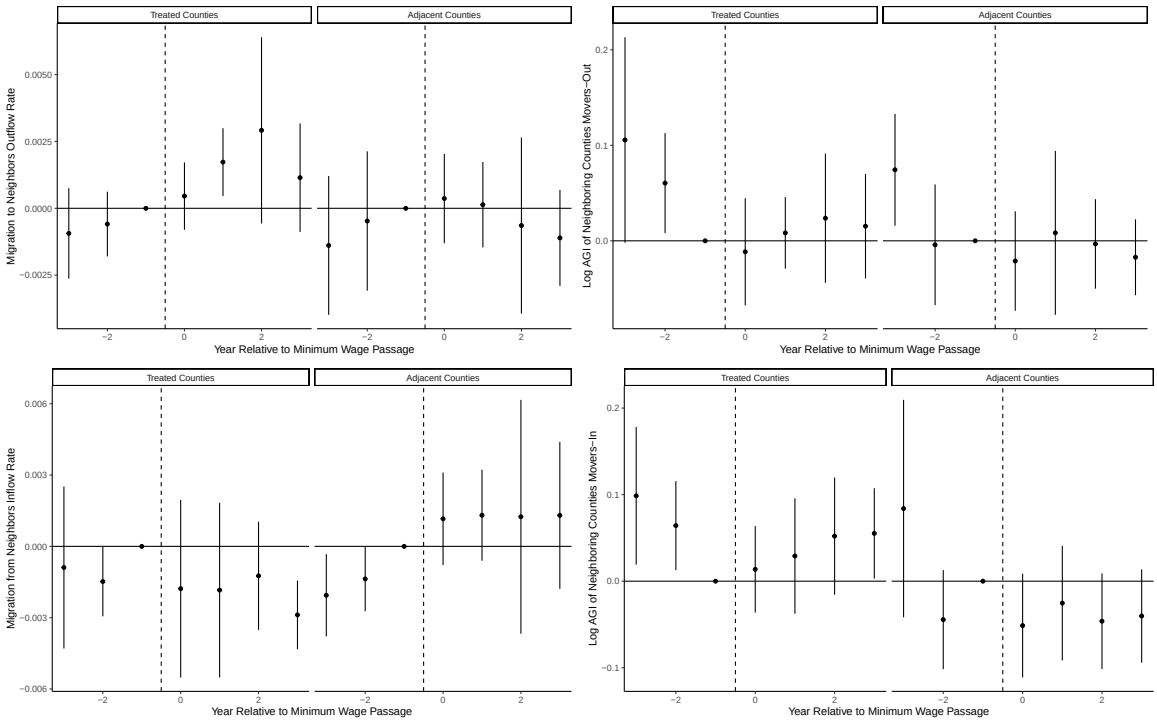


Figure C7: Matched DiD Estimates: Neighboring Counties

Notes: Figure presents matched difference-in-differences estimates.

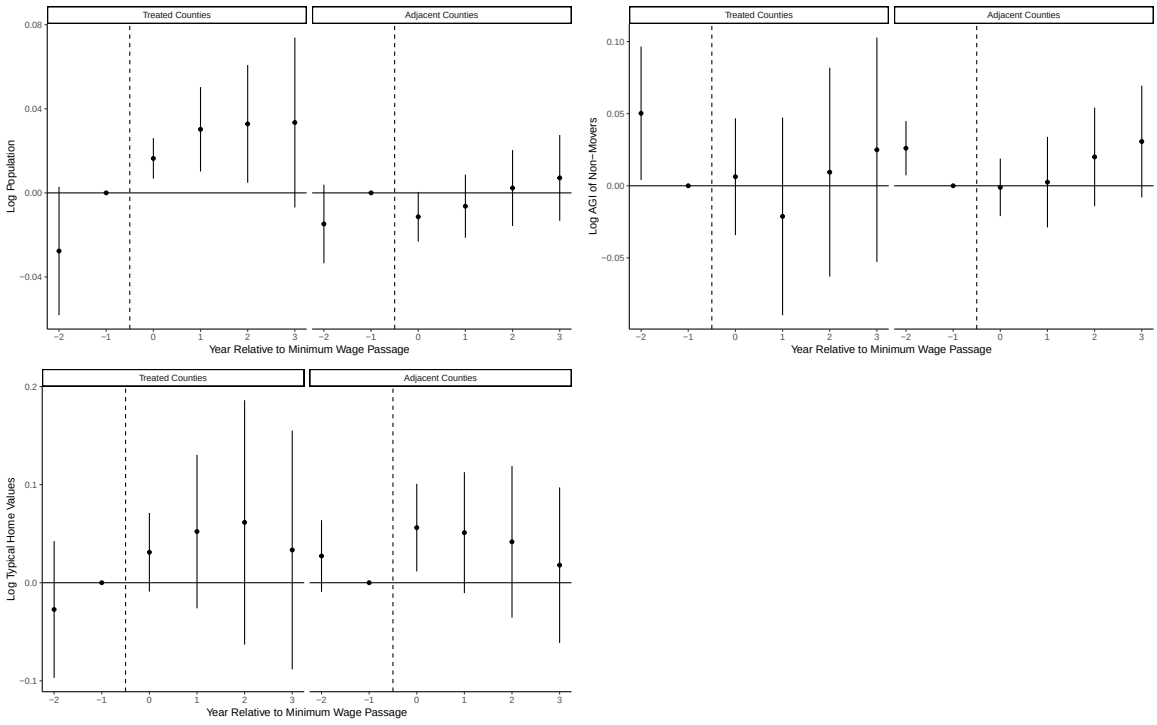


Figure C8: Matched DiD Estimates: Log Population and Non-Mover Income

Notes: Figure presents matched difference-in-differences estimates.

Bibliography

Agresti, Alan (1992). "A Survey of Exact Inference for Contingency Tables". *Statistical Science* 7.1, pp. 131–153. ISSN: 08834237.

Allegretto, Sylvia et al. (Mar. 2013). *Credible Research Designs for Minimum Wage Studies*. Tech. rep. 7638. IZA. URL: http://ftp.iza.org/dp7638.pdf.

Allegretto, Sylvia et al. (2018). "The New Wave of Local Minimum Wage Policies: Evidence from Six Cities". *CWED Policy Report*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3251977.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro (2020). *Transparency in Structural Research*. National Bureau of Economic Research Working Paper 26631.

Andrews, Isaiah and Jesse M Shapiro (2020). *A Model of Scientific Communication*. National Bureau of Economic Research Working Paper 26824.

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". *Quarterly Journal of Economics* 144.2, pp. 533–575.

Angrist, Joshua D. and Jörn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics". *Journal of economic perspectives* 24.2, pp. 3–30.

Angrist, Joshua D. and Miikka Rokkanen (2015). "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff". *Journal of the American Statistical Association* 110.512, pp. 1331–1344.

Antonovics, Kate and Brian G Knight (2009). "A New Look at Racial Profiling: Evidence from the Boston Police Department". *The Review of Economics and Statistics* 91.1, pp. 163–177.

Anwar, Shamena and Hanming Fang (2006). "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence". *American Economic Review* 96.1, pp. 127–151.

Armstrong, Timothy and Michal Kolesár (2018). "Optimal Inference in a Class of Regression Models". *Econometrica* 86(2), pp. 655–683.

Armstrong, Timothy B and Michal Kolesár (2020). "Simple and Honest Confidence Intervals in Nonparametric Regression". *Quantitative Economics* 11.1, pp. 1–39.

Arnold, David, Will Dobbie, and Crystal S. Yang (2018). "Racial Bias in Bail Decisions". *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.

Athey, Susan and Guido W. Imbens (2021). "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption". *Journal of Econometrics*.

Ba, Bocar et al. (2020). *The Impact of Racial and Ethnic Diversity in Policing*. Tech. rep.

URL: https://scholar.princeton.edu/sites/default/files/jmummolo/files/bkmr_racialethnicdiversitypolicing.pdf (visited on 04/03/2020).

Barrett, Nathan et al. (2019). "Disparities and Discrimination in Student Discipline by Race and Family Income". *Journal of Human Resources*.

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

Bielen, Samantha, Wim Marneffe, and Naci H. Mocan (2019). *Racial Bias and In-group Bias In Judicial Decisions: Evidence from Virtual Reality Courtrooms*. Working Paper 25355. National Bureau of Economic Research. URL: http://www.nber.org/papers/w25355.

Boffy-Ramirez, Ernest (2013). "Minimum Wages, Earnings, and Migration". *IZA Journal of Migration* 2.1, pp. 1–24.

Bohren, J. Aislinn et al. (2020). *Inaccurate Statistical Discrimination: An Identification Problem*. Tech. rep. National Bureau of Economic Research.

Bol, Nadine et al. (2016). "How Are Online Health Messages Processed? Using Eye Tracking to Predict Recall of Information in Younger and Older Adults". *Journal of health communication* 21.4, pp. 387–396.

Bond, Raymond et al. (2014). "Assessing Computerized Eye Tracking Technology for Gaining Insight into Expert Interpretation of the 12-Lead Electrocardiogram: an Objective Quantitative Approach". *Journal of electrocardiology* 47.6, pp. 895–906.

Borusyak, Kirill and Xavier Jaravel (2017). *Revisiting Event Study Designs*. Tech. rep. 2826228. SSRN.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh (2019). "Inference under covariate-adaptive randomization with multiple treatments". *Quantitative Economics* 10.4, pp. 1747–1785. DOI: 10.3982/QE1150. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1150. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/QE1150.

Buja, Andreas et al. (2009). "Statistical Inference for Exploratory Data Analysis and Model Diagnostics". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906, pp. 4361–4383.

Bureau of Justice Statistics (2016). *Law Enforcement Management and Administrative Statistics (LEMAS)*. DOI: 10.3886/ICPSR37323.v1.

Byrne, Sahara, Motasem Kalaji, and Jeff Niederdeppe (2018). "Does Visual Attention to Graphic Warning Labels on Cigarette Packs Predict Key Outcomes Among Youth and Low-income Smokers?" *Tobacco Regulatory Science* 4.6, pp. 18–37.

Cadena, Brian C. (2014). "Recent Immigrants as Labor Market Arbitrageurs: Evidence from the Minimum Wage". *Journal of Urban Economics* 80, pp. 1–12.

Callaway, Brantly and Pedro H.C. Sant'Anna (2020). "Difference-in-Differences with Multiple Time Periods". *Journal of Econometrics*. ISSN: 0304-4076.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell (2020a). "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs". *The Econometrics Journal* 23.2, pp. 192–210.

— (2020b). "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs". *The Econometrics Journal* 23.2, pp. 192–210.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik (2014). "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs". *Econometrica* 82.6,

pp. 2295–2326.
Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik (2015). "Optimal Data-Driven Regression Discontinuity Plots". *Journal of the American Statistical Association* 110.512, pp. 1753–1769.
Calonico, Sebastian et al. (2019). "Regression Discontinuity Designs Using Covariates". *Review of Economics and Statistics* 101.3, pp. 442–451.
Camerer, Colin F. and Eric J. Johnson (1997). "The Process-Performance Paradox in Expert Judgment: How can Experts Know so Much and Predict so Badly". *Research on judgment and decision making: Currents, connections, and controversies* 342.
Cameron, A. Colin and Douglas L. Miller (2015). "A practitioner's guide to cluster-robust inference". *Journal of Human Resources* 50.2, pp. 317–372.
Card, David, David S. Lee, and Zhuan Pei (2009). *Quasi-Experimental Identification and Estimation in the Regression Kink Design*. Princeton University Industrial Relations Section Working Paper 553.
Card, David et al. (2015b). *Inference on Causal Effects in a Generalized Regression Kink Design*. Upjohn Institute Working Paper 15-218.
— (2015a). "Inference on Causal Effects in a Generalized Regression Kink Design". *Econometrica* 83.6, pp. 2453–2483.
Card, David et al. (2015c). "The Effect of Unemployment Benefits on the Duration of Unemployment Insurance Receipt: New Evidence from a Regression Kink Design in Missouri, 2003-2013". *American Economic Review: Papers & Proceedings* 105.5, pp. 126–130.
Card, David et al. (2017). "Regression Kink Design: Theory and Practice". In: *Regression Discontinuity Designs: Theory and Applications*. Ed. by Matias D. Cattaneo and Juan Carlos Escanciano. Vol. 38. Advances in Econometrics. Emrald. Chap. 5, pp. 341–382.
Castillo-Freeman, Alida and Richard B Freeman (1992). "When the Minimum Wage Really Bites: The Effect of the US-Level Minimum on Puerto Rico". In: *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*. University of Chicago Press, pp. 177–212.
Cattaneo, Matias D, Max H Farrell, and Yingjie Feng (2020). "Large Sample Properties of Partitioning-Based Series Estimators". *Annals of Statistics* 48.3, pp. 1718–1741.
Cattaneo, Matias D. et al. (2019). "On Binscatter". *arXiv preprint arXiv:1902.09608*.
Cengiz, Doruk et al. (May 2019). "The Effect of Minimum Wages on Low-Wage Jobs". *The Quarterly Journal of Economics* 134.3, pp. 1405–1454.
Chapman, Jonathan et al. (2018). *Loss attitudes in the US population: Evidence from dynamically optimized sequential experimentation (DOSE)*. National Bureau of Economic Research Working Paper 25072.
Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments". *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.
Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates". *American Economic Review* 104.9, pp. 2593–2632.
Chetty, Raj et al. (Nov. 2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star". *The Quarterly Journal of Economics* 126.4,

pp. 1593–1660. ISSN: 0033-5533. DOI: 10.1093/qje/qjr041. eprint: https://academic.oup.com/qje/article-pdf/126/4/1593/17089543/qjr041.pdf.

Chetty, Raj et al. (Sept. 2014). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States". *The Quarterly Journal of Economics* 129.4, pp. 1553–1623. ISSN: 0033-5533. DOI: 10.1093/qje/qju022. eprint: https://academic.oup.com/qje/article-pdf/129/4/1553/30631636/qju022.pdf.

Cleveland, William S., Persi Diaconis, and Robert McGill (1982). "Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased". *Science* 216.4550, pp. 1138–1141.

Cleveland, William S. and Robert McGill (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". *Journal of the American Statistical Association* 79.387, pp. 531–554.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". *Educational and Psychological Measurement* 20.1, pp. 37–46. eprint: https://doi.org/10.1177/001316446002000104.

Coviello, Decio and Nicola Persico (2015). "An Economic Analysis of Black-White Disparities in the New York Police Department's Stop-and-Frisk Program". *The Journal of Legal Studies* 44.2, pp. 315–360.

Davenport, Ernest C. and Nader A. El-Sanhurry (1991). "Phi/Phimax: Review and Synthesis". *Educational and Psychological Measurement* 51.4, pp. 821–828. eprint: https://doi.org/10.1177/0013164491051004003.

De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9, pp. 2964–96.

DellaVigna, Stefano and Devin Pope (2018). "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126.6, pp. 2410–2456.

Devi, Tanaya and Roland G. Fryer (2020). *Policing the Police: The Impact of "Pattern-or-Practice" Investigations on Crime*. Working Paper 27324. National Bureau of Economic Research. URL: http://www.nber.org/papers/w27324.

Donoho, David L (1994). "Statistical Estimation and Optimal Recovery". *The Annals of Statistics*, pp. 238–270.

Dube, Arindrajit, T William Lester, and Michael Reich (2010). "Minimum wage effects across state borders: Estimates using contiguous counties". *The Review of Economics and Statistics* 92.4, pp. 945–964.

Dube, Arindrajit and Attila Lindner (2021). "City Limits: What Do Local-Area Minimum Wages Do?" *Journal of Economic Perspectives* 35.1, pp. 27–50.

Dube, Arindrajit, Suresh Naidu, and Michael Reich (2007). "The Economic Effects of a Citywide Minimum Wage". *Industrial and Labor Relations Review* 60.4, pp. 522–543.

Durett, Rick (2010). *Probability: Theory and Examples*. 4th. Cambridge University Press.

Durlauf, Steven N. and James J. Heckman (2020). "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment". *Journal of Political Economy* 128.10, pp. 3998–4002.

Dynarski, Susan (2014). "What We Mean when We Say Student Debt is Bad". *New York Times*. URL: https://www.nytimes.com/2014/08/09/upshot/what-we-mean-when-we-say-student-debt-is-bad.html.

Eells, Walter Crosby (1926). "The Relative Merits of Circles and Bars for Representing Component Parts". *Journal of the American Statistical Association* 21.154, pp. 119–132.

Elliott, Graham, Ulrich K. Mäller, and Mark W. Watson (2015). "Nearly Optimal t-Test when a Nuisance Parameter is Present under the Null Hypothesis". *Econometrica* 83.2, pp. 771–811.

Fisher, Ronald Aylmer (1925). *Statistical Methods for Research Workers*. Oliver and Boyd (Edinburgh).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Fryer, Roland G. (2019). "An Empirical Analysis of Racial Differences in Police Use of Force". *Journal of Political Economy* 127.3, pp. 1210–1261.

— (2020). "An Empirical Analysis of Racial Differences in Police Use of Force: A Response". *Journal of Political Economy* 128.10, pp. 4003–4008.

Gail, M. and R. Simon (1985). "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets". *Biometrics* 41.2, pp. 361–372.

Garcia Moreno-Esteva, Enrique et al. (2017). "Identifying Key Visual-Cognitive Processes in Students' Interpretation of Graph Representations Using Eye-Tracking Data and Math or Machine Learning Based Data Analysis". In: *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1-5, 2017)*. Dublin City University.

Gelman, Andrew (2008). "Scaling Regression Inputs by Dividing by Two Standard Deviations". *Statistics in Medicine* 27.15, pp. 2865–2873.

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, pp. 813–823.

Giulietti, Corrado (2014). "Is the Minimum Wage a Pull Factor for Immigrants?" *ILR Review* 67.3_suppl, pp. 649–674.

Goncalves, Felipe and Steven Mello (Mar. 2017). *A Few Bad Apples? Racial Bias in Policing*. Working Papers 608. Princeton University, Department of Economics, Industrial Relations Section. URL: https://ideas.repec.org/p/pri/indrel/608.html.

Goodman-Bacon, Andrew (2020). *Difference-in-Differences with Variation in Treatment Timing*. Tech. rep. URL: http://goodman-bacon.com/pdfs/ddtiming.pdf.

Grogger, Jeffrey and Greg Ridgeway (2006). "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness". *Journal of the American Statistical Association* 101.475, pp. 878–887.

Habiger, Joshua D. and Edsel A. Peña (2011). "Randomised P-values and Nonparametric Procedures in Multiple Testing". *Journal of Nonparametric Statistics* 23.3, pp. 583–604.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (1999). *Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design*. National Bureau of Economic Research Working Paper 7131.

Heaton, Paul (2010). "Understanding the Effects of Antiprofiling Policies". *The Journal of Law and Economics* 53.1, pp. 29–64.

Hegarty, Mary, Matt S Canham, and Sara I Fabrikant (2010). "Thinking About the Weather:

How Display Salience and Knowledge Affect Performance in a Graphic Inference Task." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36.1, p. 37.

Hoeffding, Wassily (1956). "On the Distribution of the Number of Successes in Independent Trials". *Annals of Mathematical Statistics* 27.3, pp. 713–721.

Hoekstra, Mark and CarlyWill Sloan (2020). *Does Race Matter for Police Use of Force? Evidence from 911 Calls*. Working Paper 26774. National Bureau of Economic Research. URL: http://www.nber.org/papers/w26774.

Hofmann, Heike et al. (2012). "Graphical Tests for Power Comparison of Competing Designs". *IEEE Transactions on Visualization and Computer Graphics* 18.12, pp. 2441–2448.

Horrace, William C. and Shawn M. Rohlin (2016). "How Dark Is Dark? Bright Lights, Big City, Racial Profiling". *The Review of Economics and Statistics* 98.2, pp. 226–232.

Iacus, Stefano M, Gary King, and Giuseppe Porro (2012). "Causal Inference without Balance Checking: Coarsened Exact Matching". *Political analysis*, pp. 1–24.

Imai, Kosuke, In Song Kim, and Erik Wang (2020). *Matching Methods for Causal Inference with Time-Series Cross-Sectional Data*. Tech. rep. URL: https://imai.fas.harvard.edu/research/files/tscs.pdf.

Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". *The Review of economic studies* 79.3, pp. 933–959.

Imbens, Guido and Stefan Wager (2019). "Optimized Regression Discontinuity Designs". *Review of Economics and Statistics* 101.2, pp. 264–278.

Imbens, Guido W. and Thomas Lemieux (2008). "Regression Discontinuity Designs: A Guide to Practice". *Journal of Econometrics* 142, pp. 615–635.

Jardim, Ekaterina et al. (2017). *Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle*. Working Paper 23532. National Bureau of Economic Research. URL: http://www.nber.org/papers/w23532.

— (2018). *Minimum Wage Increases and Individual Employment Trajectories*. Working Paper 25182. National Bureau of Economic Research. URL: http://www.nber.org/papers/w25182.

Just, Marcel A. and Patricia A. Carpenter (1980). "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological review* 87.4, p. 329.

Kahneman, Daniel and Amos Tversky (Mar. 1979). "Prospect Theory: An Analysis of Decision under Risk". *Econometrica* 47.2, pp. 263–291.

Kane, Thomas J. and Douglas O. Staiger (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation". National Bureau of Economic Research working paper 14607.

Kline, Patrick and Christopher Walters (2019). *Audits as Evidence: Experiments, Ensembles, and Enforcement*. Tech. rep. URL: https://arxiv.org/abs/1907.06622.

Knowles, John, Nicola Persico, and Petra Todd (2001). "Racial Bias in Motor Vehicle Searches: Theory and Evidence". *Journal of Political Economy* 109.1, pp. 203–229.

Knox, Dean, Will Lowe, and Jonathan Mummolo (2020). "Administrative Records Mask Racially Biased Policing". *American Political Science Review* 114.3, pp. 619–637.

Korting, Christina et al. (2019a). "A Study on Graphical Representation". *AEA RCT Reg-*

istry (https://doi.org/10.1257/rct.4331-1.0).
Korting, Christina et al. (2019b). "A Study on Graphical Representation". *Retrieved from osf.io/jeax5*.
Kriz, Sarah and Mary Hegarty (2007). "Top-Down and Bottom-Up Influences on Learning from Animations". *International Journal of Human-Computer Studies* 65.11, pp. 911–930.
Kuehn, Daniel (2016). "Spillover Bias in Cross-Border Minimum Wage Studies: Evidence from a Gravity Model". *Journal of Labor Research* 37.4, pp. 441–459.
Lee, David S. (2008). "Randomized experiments from non-random selection in US House elections". *Journal of Econometrics* 142.2, pp. 675–697.
Lee, David S. and Thomas Lemieux (2010). "Regression Discontinuity Designs in Economics". *Journal of Economic Literature* 48.2, pp. 281–355.
Lee, David S. et al. (2019). *Are Sufficient Statistics Necessary? Nonparametric Measurement of Deadweight Loss from Unemployment Insurance*. National Bureau of Economic Research Working Paper 25574.
Leek, Jeffrey T. and Roger D. Peng (2015). "Statistics: P Values are Just the Tip of the Iceberg". *Nature News* 520.7549, p. 612.
Li, Housen et al. (2020). "The Essential Histogram". *Biometrika* 107.2, pp. 347–364.
Luh, Elizabeth (2020). *Not So Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports*. Tech. rep. URL: https://elizluh.github.io/files/highwaybias.pdf (visited on 06/06/2020).
Majumder, Mahbubul, Heike Hofmann, and Dianne Cook (2013). "Validation of Visual Statistical Inference, Applied to Linear Models". *Journal of the American Statistical Association* 108.503, pp. 942–956.
McCarthy, Craig (2019). *Help us continue to improve N.J's first statewide database of police use of force*. URL: https://www.nj.com/news/2018/12/help-us-continue-to-improve-njs-first-statewide-database-of-police-use-of-force.html (visited on 06/01/2019).
McCarthy, Craig and Blake Nelson (2019). *Every way a police officer is legally allowed to harm another person, ranked*. URL: https://www.nj.com/news/2018/11/every_way_a_police_officer_is_legally_allowed_to_harm_another_person_from_a_to_z.html (visited on 06/01/2019).
McCarthy, Craig and Stephen Stirling (2019). *How we built the most comprehensive statewide database of police force in the United States*. URL: https://www.nj.com/news/2018/11/how_we_built_the_most_comprehensive_statewide_database_of_police_force_in_the_us.html (visited on 06/01/2019).
McCrary, Justin (2008). "Manipulation of the running variable in the regression discontinuity design: A density test". *Journal of Econometrics* 142.2, pp. 698–714.
McKinnish, Terra (2017). "Cross-State Differences in the Minimum Wage and Out-of-State Commuting by Low-Wage Workers". *Regional Science and Urban Economics* 64, pp. 137–147.
Meer, Jonathan and Jeremy West (2016). "Effects of the Minimum Wage on Employment Dynamics". *Journal of Human Resources* 51.2, pp. 500–522.
Merrell, Melissa (2007). "The Impact of Unauthorized Immigrants on the Budgets of State and Local Governments". In: Congress of the United States, Congressional Budget Of-

fice.

Molloy, Raven, Christopher L Smith, and Abigail Wozniak (2011). "Internal Migration in the United States". *Journal of Economic perspectives* 25.3, pp. 173–96.

Monras, Joan (2019). "Minimum Wages and Spatial Equilibrium: Theory and Evidence". *Journal of Labor Economics* 37.3, pp. 853–904. DOI: 10.1086/702650. eprint: https://doi.org/10.1086/702650.

Morris, Carl N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications". *Journal of the American Statistical Association* 78.381, pp. 47–55.

Mosteller, Frederick and Ronald Aylmer Fisher (1948). "Questions and Answers". *The American Statistician* 2.5, pp. 30–31. ISSN: 00031305.

Nelson, Blake (2019). *Frequently asked questions about The Force Report*. URL: https://www.nj.com/news/2018/11/frequently_asked_questions_about_the_force_report.html (visited on 06/01/2019).

Neumark, David, JM Ian Salas, and William Wascher (2014). "Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater?" *Ilr Review* 67.3_suppl, pp. 608–648.

Nicholson-Crotty, Sean, Jill Nicholson-Crotty, and Sergio Fernandez (2017). "Will More Black Cops Matter? Officer Race and Police-Involved Homicides of Black Citizens". *Public Administration Review* 77.2, pp. 206–216.

Nickell, Stephen (1981). "Biases in Dynamic Models with Fixed Effects". *Econometrica: Journal of the Econometric Society* 49(6), pp. 1417–1426.

Nix, Justin et al. (2017). "A Bird's Eye View of Civilians Killed by Police in 2015: Further Evidence of Implicit Bias". *Criminology & Public Policy* 16.1, pp. 309–340.

NJ Advance Media (2019). *The Force Report*. URL: http://force.nj.com (visited on 06/01/2019).

Orrenius, Pia M. and Madeline Zavodny (2008). "The Effect of Minimum Wages on Immigrants' Employment and Earnings". *ILR Review* 61.4, pp. 544–563.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt (2019). "Poorly measured confounders are more useful on the left than on the right". *Journal of Business & Economic Statistics* 37.2, pp. 205–216.

Pei, Zhuan and Yi Shen (2017). "The Devil is in the Tails: Regression Discontinuity Design with Measurement Error in the Assignment Variable". *Advances in Econometrics* 38, pp. 455–502.

Pei, Zhuan et al. (2018). *Local Polynomial Order in Regression Discontinuity Designs*. Princeton University Industrial Relations Section Working Paper 622.

Percus, Ora E. and Jerome K. Percus (1985). "Probability Bounds on the Sum of Independent Nonidentiacally Distributed Binomial Random Variables". *SIAM Journal of Applied Mathematics* 45.4, pp. 621–640.

Pérez, Jorge Pérez (2020). *City Minimum Wages and Spatial Equilibrium effects*. Tech. rep. URL: https://jorgeperezperez.com/files/Jorge_Perez_JMP.pdf (visited on 02/15/2021).

Persico, Nicola and Petra Todd (2006). "Generalising the Hit Rates Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita". *The Economic Journal* 116.515, F351–F367.

Potter, Nicholas (2006). "Measuring the Employment Impacts of the Living Wage Ordi-

nance in Santa Fe, New Mexico".
Raphael, Steven and Sandra V. Rozo (2019). "Racial Disparities in the Acquisition of Juvenile Arrest Records". *Journal of Labor Economics* 37.S1, S125–S159.
Rayner, Keith (1998). "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological bulletin* 124.3, p. 372.
Rehavi, M. Marit and Sonja B. Starr (2014). "Racial Disparity in Federal Criminal Sentences". *Journal of Political Economy* 122.6, pp. 1320–1354.
Reich, Michael, Sylvia Allegretto, and Anna Godoey (2017). "Seattle's Minimum Wage Experience 2015-16". *CWED Policy Report*. URL: https://www.irle.berkeley.edu/files/2017/Seattles-Minimum-Wage-Experiences-2015-16.pdf.
Rosen, Rebecca J. (2015). "Slight Changes in Yelp Ratings can Mean Huge Losses for Small Businesses". *The Atlantic*. URL: https://www.theatlantic.com/technology/archive/2012/09/slight-changes-in-yelp-ratings-can-mean-huge-losses-for-small-businesses/261943/.
Rosenholtz, Ruth, Yuanzhen Li, and Lisa Nakano (2007). "Measuring Visual Clutter". *Journal of vision* 7.2, pp. 17–17.
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.
Sanders, Michael, Freddie Mitchell, and Aisling Ni Chonaire (2015). *Just Common Sense? How Well Do Experts and Lay-People Do at Predicting the Findings of Behavioural Science Experiments*. Harvard University Kennedy School Working Paper.
Santiago, Carlos E (1993). "The Migratory Impact of Minimum Wage Legislation: Puerto Rico, 1970–1987". *International Migration Review* 27.4, pp. 772–795.
Schmitt, John and David Rosnick (2011). "The Wage and Employment Impact of Minimum-Wage Laws in Three Cities". *Center for Economic and Policy Research*.
Shen, Yi and Tony S. Wirjanto (2019). "Stationarity as a Path Property". *Probability and Mathematical Statistics* 39(2), pp. 403–422.
Sides, John (2015). "How to Get Young People to Vote? Register Them Before they Turn 18." *The Washington Post*.
Silverman, Bernard W. (1986). *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC Press.
Stock, James H. and Mark W. Watson (2011). *Introduction to Econometrics, 3rd edition*. Pearson.
Sun, Liyang and Sarah Abraham (2020). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". *Journal of Econometrics*.
Terrill, William and Michael D. Reisig (2003). "Neighborhood Context and Police Use of Force". *Journal of Research in Crime and Delinquency* 40.3, pp. 291–321.
Watson, Geoffrey S. (1964). "Smooth Regression Analysis". *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.
Weisburst, Emily K. (2019). "Police Use of Force as an Extension of Arrests: Examining Disparities across Civilian and Officer Race". *AEA Papers and Proceedings* 109, pp. 152–56.
West, Jeremy (2018). *Racial Bias in Police Investigations*. Tech. rep. URL: https://people.ucsc.edu/~jwest1/articles/West_RacialBiasPolice.pdf (visited on 06/07/2019).

Wilkins, Vicky M. and Brian N. Williams (2008). "Black or Blue: Racial Profiling and Representative Bureaucracy". *Public Administration Review* 68.4, pp. 654–664.

Wilson, Daniel J. (2019). "The harmonic mean p-value for combining dependent tests". *Proceedings of the National Academy of Sciences* 116(4), pp. 1195–1200. ISSN: 0027-8424. eprint: https://www.pnas.org/content/116/4/1195.full.pdf.

Wolfe, Jeremy M and Todd S Horowitz (2004). "What Attributes Guide the Deployment of Visual Attention and How Do They Do it?" *Nature reviews neuroscience* 5.6, pp. 495–501.