**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

2023

# Essays on the Economics of Child Welfare and Social Experiences

Mills, Christopher

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01v692t952p

*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 on the Economics of Child Welfare and Social Experiences

Christopher Mills

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: Janet Currie

May 2023

© Copyright by Christopher Mills, 2023.

All Rights Reserved

# Abstract

This dissertation includes three chapters on the economics of child welfare and social experiences, through the lens of labor and public economics. A common theme in these papers is how targeted interventions in the forms of algorithmic information, peers, and social gatherings, respectively, shape long-run outcomes of adolescents and young adults. A second common thread is the use of proprietary linked administrative data and quasi-experimental methods for policy evaluation.

Chapter 1, co-authored with Marie-Pascale Grimon, studies the impact of algorithmic information on child protective services (CPS) social worker investigation decisions and child outcomes. We evaluate the effects of providing workers with randomized access to an algorithmic score that accurately predicts whether a child will be removed from their home due to maltreatment. Giving workers access to the score reduced child injury-related hospitalizations by 32 percent and narrowed Black-white racial disparities in CPS investigations. Text analysis of worker discussion notes suggests that algorithmic predictions allow workers to better focus their attention on supplemental features of the allegation.

Chapter 2, co-authored with Sarah Font, documents the role of peers on the outcomes of youth in congregate foster care. Using three decades of Wisconsin administrative data, we leverage exogenous variation in the relative imprisonment risk of foster care peers. A one standard deviation increase in peer imprisonment risk is associated with a modest 2.5 percent increase in a youth's likelihood of dropping out of high school. However, peer risk has no effect on a child's likelihood of entering prison by age 20, nor on several other indicators of deviant behavior.

Finally, Chapter 3 studies the impact of social gatherings on student giving and career trajectories. Using the rotational timing of the Urbana missions conference, a large collegiate religious gathering, and additional variation in weather-related flight cancellations, I find that conference attendance produces persistent long-run increases in giving and employment for the sponsor organization, and broader changes in initial industry of employment.

Gatherings are most influential for graduating-cohort attendees compared to younger students, suggesting that similar interventions may be particularly influential at key transition points.

# Acknowledgements

I am forever grateful for the unyielding and gracious support of my adviser and champion, Janet Currie. Janet has strengthened me professionally and personally in manifold ways throughout my time at Princeton. She has honed my abilities as a junior scholar and my capacity to serve other students through teaching and mentoring. Janet, thank you for your guidance and your friendship. You have been a wellspring of career and life advice and it has been a privilege learning under you.

I am also extremely thankful for committee members Ilyana Kuziemko and Leah Boustan. I am repeatedly wowed by Ilyana's knack for refining core research questions and exceptional ability to uplift and motivate. Leah helped me think carefully about how to communicate well and has been a model of excellence in how to give research feedback, develop teaching materials, and foster a thriving intellectual community in the Industrial Relations Section. She has also been a trusted conversation partner for ideas and musings. Leah and Ilyana, you have given me every opportunity to succeed and I am so grateful for you.

I have benefitted from the counsel of other current and former Princeton faculty: Orley Ashenfelter, Ellora Derenoncourt, Will Dobbie, Hank Farber, Kate Ho, Bo Honoré, Alan Krueger, David Lee, Alex Mas, Kelly Noonan, David Silver, Tom Vogl, Leeat Yariv, and Owen Zidar. I am grateful to Ellora for her guidance in framing my job market paper, Kelly for years of wisdom in teaching Princeton students, and Gianluca Violante for his help in navigating the job market. I am also indebted to external mentors and collaborators, namely Lonnie Berger, Maria Cancian, Anna Chorniy, Scott Cunningham, Natalia Emanuel, Sarah Font, Marie-Pascale Grimon, Dan Hungerman, Bob Kubinec, Paul Niehaus, Vesall Nourani, Emily Putnam-Hornstein, and Rhema Vaithianathan. You opened doors to study and improve the child welfare system, opened my mind to new ways of thinking and collaborating, and have repeatedly inspired me. Finally, I would not have advanced this far without the mentorship of my undergraduate adviser, Chris Barrett, who still manages to find new ways to surprise me with his kindness.

The Princeton Department of Economics, Center for Health and Wellbeing, and Industrial Relations Section have given me an academic home as I have completed my studies. Graduate Program Administrator Laura Hedden deserves an acknowledgements section all her own: Thank you, Laura. I am also grateful for administrators and staff including Valerie Ching, Gilbert Collins, Justine Conoline, Rachel Conner, Stephanie Cuccaro-Alamin, Indiana Gaffney, Ellen Graf, Larissa Lorimer, Dana Molina, Marcus Mullins, Thad Paul, Debra Pino Betancourt, John Prindle, Patti Tracey, custodial and dining staff, IRB staff, and generous funding from the Graduate School, Department of Economics, Center for Health and Wellbeing, and IR Section.

I have been incredibly fortunate for the economics friends who have accompanied me on this journey: all-star first-year study group Emily Battaglia, Faizaan Kisat, and Victoria Larsen, fellow skit-maker and GEC partner-in-crime Jonas Jin, job market confidantes Lukas Althoff, Jiwon Choi, and Christiane Szerman, and many other dear colleagues including David Arnold, Reyhan Ayas, Jessica Brown, Jacob Dorn, Allie Green, Stephanie Hao, Elisa Jácome, Andrew Langan, Nikita Melnikov, Boriana Miloucheva, Leon Musolff, Katherine Rittenhouse, Brandon Tan, Jenny Shen, Jeffrey Sun, Alex Wise, Kecy Wu, Luther Yap, Jason Zhang, and Yinuo Zhang. I also thank friends outside of economics who have uplifted me time and time again: Johnny, Chris, Peter, Eric, Darren, Daniel, Bolton, Felix, John, Matthew, Myeonhu, Susan, Naomi, Stephen, Grace, Liane, Maura, Tiffany, Jared, and Jess.

Economists know the importance of unexpected, sometimes-exogenous factors in facilitating professional opportunity. I am grateful to Lauren Elman and Saman Nazarian at Penn Medicine for keeping me on my feet, Fairfax County Public Schools and Cornell University for an excellent education, and Hope Church and the Princeton Christian community for their care. Foremost, I thank God, the giver of life, beauty, and hope, who has been gracious to me beyond measure.

Finally, I am grateful to my family. To my mother, Kathleen Mills, thank you for your camaraderie and near-constant optimism. To my father, Raymond Mills, thank you

for setting an example for me in foresight, self-sacrifice, and first inspiring me to study economics through atlases and books on poverty alleviation. You are amazing parents and I love you. Stephen and Tracey: I am more and more thankful we have each other as siblings each passing year.

For Mom and Dad.

# Contents

Abstract . . . . . iii

Acknowledgements . . . . . v

List of Tables . . . . . xiii

List of Figures . . . . . xvii

**1 The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial** **1**

1.1 Introduction . . . . . 2

1.2 Child Protection and Decision Tool Context . . . . . 8

1.2.1 Child Protective Services in the Evaluation County . . . . . 8

1.2.2 Decision Tool Background . . . . . 11

1.3 Trial . . . . . 14

1.3.1 Intervention Design . . . . . 14

1.3.2 Study Timeline . . . . . 16

1.3.3 Trial Sample . . . . . 17

1.4 Data and Descriptive Statistics . . . . . 18

1.4.1 Child Welfare Integrated Data System . . . . . 18

1.4.2 Risk Score Data . . . . . 19

1.4.3 Statewide Simulated Risk Scores . . . . . 19

1.4.4 Tool Randomization Database . . . . . 19

1.4.5 Discussion Text and Timestamp Files . . . . . 20

1.4.6 Hospital Records . . . 20
1.4.7 Descriptive Statistics . . . 21
1.5 Empirical Strategy . . . 22
1.5.1 Estimating Tool Impacts . . . 22
1.5.2 Confirming Random Assignment . . . 23
1.6 Key Findings . . . 24
1.6.1 Tool Validity . . . 24
1.6.2 Tool Use . . . 25
1.6.3 Impact of Tool on Child Health and Wellbeing . . . 26
1.6.4 Racial Disparities . . . 33
1.6.5 Impact of Algorithmic Information on Decisions . . . 35
1.6.6 Effect of Tool on Other Outcomes . . . 40
1.6.7 Team Discussions . . . 41
1.7 Conclusion . . . 43
1.8 Figures and Tables . . . 45

**Appendices** **64**

1.A Appendix Figures and Tables . . . 64
1.B Decision Protocol . . . 98

**2 Safe from Harm? Peers and Criminal Capital Formation in Congregate Foster Care** **99**

2.1 Introduction . . . 100
2.2 Background . . . 104
2.3 Data and Descriptive Statistics . . . 109
2.3.1 Risk Score Construction . . . 113
2.4 Empirical Strategy . . . 115
2.5 Results . . . 120

2.5.1 Exploring Possible Mechanisms of Peer Influence . . . . . 123
2.5.2 Additional Outcomes: High School Dropout and Teen Parenthood . . 126
2.5.3 Additional Discussion . . . . . 127
2.6 Conclusion . . . . . 129
2.7 Figures and Tables . . . . . 132

**Appendices** **141**

2.A Appendix Figures and Tables . . . . . 141

**3 Gatherings, Giving, and Life Trajectories: Evidence from Student Missions Conferences** **165**

3.1 Introduction . . . . . 166
3.2 The Urbana Missions Conference . . . . . 169
3.3 Data Sources . . . . . 171
3.3.1 Conference Registration Data . . . . . 171
3.3.2 Flight Cancellation Records . . . . . 172
3.3.3 Administrative Microdata . . . . . 172
3.3.4 Survey Data . . . . . 173
3.4 Short-Run Effects of Gatherings . . . . . 174
3.4.1 Short-Run Empirical Strategy . . . . . 174
3.4.2 Short-Run Sample and Descriptive Statistics . . . . . 176
3.4.3 Short-Run Results . . . . . 177
3.5 Cohort-Level Effects . . . . . 181
3.5.1 Cohort-Level Empirical Strategy . . . . . 181
3.5.2 Cohort-Level Sample and Descriptive Statistics . . . . . 183
3.5.3 Cohort-Level Results . . . . . 183
3.6 Long-Run Effects . . . . . 185
3.6.1 Long-Run Empirical Strategy . . . . . 185

3.6.2 Long-Run Sample and Descriptive Statistics . . . . . 187
3.6.3 Long-Run Results . . . . . 189
3.6.4 Dynamic Trends . . . . . 192
3.7 Mechanisms and Heterogeneity . . . . . 194
3.8 The Effect of Messaging on Representation . . . . . 197
3.9 Conclusion . . . . . 199
3.10 Figures and Tables . . . . . 201

**Appendices** **212**

3.A Appendix Figures and Tables . . . . . 212

**Bibliography** **227**

# List of Tables


1.1 Descriptive Statistics . . . . . 59

1.2 Tool Predictive Power Relative to Workers . . . . . 60

1.3 Summary Statistics for Main Hospitalization Outcomes . . . . . 61

1.4 Impact of Tool Availability on Hospitalization . . . . . 62

1.5 Impact of Tool Availability on Investigation . . . . . 63

1.A.1 Sample Restrictions for Trial Analysis Sample, November 2020 - March 2022 83

1.A.2 Test of Randomization of Intervention Status . . . . . 83

1.A.3 Frequency of Score Seen when Tool Available . . . . . 84

1.A.4 Impact of Tool on Child Health, Non-Active Referrals . . . . . 85

1.A.5 Impact of Tool on Child Health, Active Referrals . . . . . 85

1.A.6 Impact of Tool on Child Health, Short Run . . . . . 86

1.A.7 Tool and Child Health – Placebo Outcomes Prior to Randomization . . . 86

1.A.8 Supervisor Propensity Instrument Details . . . . . 87

1.A.9 Effect of Tool on Racial Disparities in Screen-Ins, CPS-Known Race . . . 88

1.A.10 Effect of Tool on Racial Disparities in Screen-Ins, CPS-Known Race and Hospital-Known Race . . . . . 88

1.A.11 Effect of Tool on Ethnic Disparities in Screen-Ins . . . . . 89

1.A.12 Level Effect of Tool on CPS-Related Outcomes, All Referrals . . . . . 90

1.A.13 Level Effect of Tool on CPS-Related Outcomes, Non-Active Referrals . . . 91

1.A.14 Level Effect of Tool on CPS-Related Outcomes, Active Referrals . . . . . 92

1.A.15 Effect of Tool on CPS-Related Outcomes, by Score . . . . . 93
1.A.16 Effect of Tool on CPS-Related Outcomes for Non-Active Referrals, by Score 94
1.A.17 Effect of Tool on CPS-Related Outcomes for Active Referrals, by Score . . 95
1.A.18 Effect of Tool Availability on Screen-In for Non-Active Referrals: Discontinuity Design . . . . . 96
1.A.19 Effect of Tool Availability on Screen-In: Child Fixed Effects . . . . . 97
1.A.20 Test of Within-Day Spillovers from Tool Availability . . . . . 97

2.7.1 Characteristics of Foster Care Youth by Provider Type . . . . . 133
2.7.2 Association between Peer Imprisonment Risk and Own Imprisonment Risk 134
2.7.3 Association between a Youth's Own Outcomes and Own Delinquency or Risk 135
2.7.4 Delinquent Peers and Youth Outcomes . . . . . 136
2.7.5 Delinquent Peers and Own Imprisonment Heterogeneity by Placement Type 137
2.7.6 Delinquent Peers and Own Imprisonment Heterogeneity by Child Welfare Experiences . . . . . 138
2.7.7 Delinquent Peers and Own Imprisonment Heterogeneity by Youth Background Characteristics . . . . . 139
2.7.8 Peer Imprisonment Risk and High School Dropout Heterogeneity by Youth Background Characteristics . . . . . 140
2.A.1 LASSO Coefficients of Standardized Variables Sorted by Penalty Importance 147
2.A.2 Delinquent Peers and Youth Outcomes: Heterogeneity by Own Delinquency 148
2.A.3 Delinquent Peers and Own Imprisonment, Robustness to Controls . . . . 149
2.A.4 Delinquent Peers and Youth Outcomes, Males Only . . . . . 150
2.A.5 Delinquent Peers and Youth Outcomes, First Congregate Care Placement Only . . . . . 151
2.A.6 Delinquent Peers and Youth Outcomes, Counting Peers of All Genders . . 152
2.A.7 Delinquent Peers and Youth Outcomes, Omitting Counties Near Milwaukee 152
2.A.8 Heterogeneity by Provider Median Placement Duration . . . . . 153

2.A.9 Peer Risk and Own Imprisonment Heterogeneity by Child Welfare Experiences . . . 154
2.A.10 Peer Risk and Own Imprisonment Heterogeneity by Background Characteristics . . . 155
2.A.11 Delinquent Peers and Youth Outcomes, Disaggregated by Gender . . . 156
2.A.12 Peer Demographics and Own Imprisonment . . . 157
2.A.13 Peer Demographics and Own Imprisonment, by Gender . . . 158
2.A.14 Alternative Models of Peer Risk and Own Imprisonment . . . 159
2.A.15 Nonlinear Model of Peer Risk and Own Imprisonment . . . 160
2.A.16 Association between Peer Imprisonment Risk and High School Dropout Heterogeneity by Placement Type . . . 160
2.A.17 Association between Peer Imprisonment Risk and High School Completion Heterogeneity by Placement Type . . . 161
2.A.18 Peer Imprisonment Risk and High School Dropout Heterogeneity by Child Welfare Experiences . . . 161
2.A.19 High School Dropout: Heterogeneity by Peer Characteristics . . . 162
2.A.20 Peers and Teen Parenthood . . . 163
2.A.21 Peers and Teen Parenthood Heterogeneity by Placement Type . . . 163
2.A.22 Peers and Teen Parenthood Heterogeneity by Risk Factors . . . 164

3.10.1 Descriptive Statistics by College and Year, 2003-2017 . . . 206
3.10.2 Conference Short-Run Effects Donors and Staff . . . 207
3.10.3 Conference Cohort-Level Effects . . . 208
3.10.4 Characteristics of Undergraduate Conference Registrants, 2000-2015 . . . 209
3.10.5 Estimates of Conference Attendance on Giving and Staff Employment . . . 210
3.10.6 Estimates of Conference Attendance on Giving and Staff Employment, by Grade . . . 211

3.A.1 Conference Short-Run Effects on Giving and Staff Employment Robustness to Simulated Attendance . . . . . 217
3.A.2 Conference Short-Run Effects on Giving and Staff Employment Restricted to Students who Filled Out Departing Student Form . . . . . 218
3.A.3 Conference Short-Run Effects Departing Student Form Responses . . . . . 219
3.A.4 Short-Run Effect of Conference on Reported Post-Graduation Job . . . . 220
3.A.5 Conference Short-Run Effects Select Campus Survey Responses . . . . . 220
3.A.6 Flight Cancellation Instrument Details . . . . . 221
3.A.7 Estimates of Conference Attendance Sample of Students who did not have First-Time Donation at the Conference . . . . . 223
3.A.8 Estimates of Conference Attendance on Giving and Staff Employment: Expanded Outcomes . . . . . 224
3.A.9 Future Event Participation . . . . . 225
3.A.10 Heterogeneity by Level of sponsor Organization Involvement . . . . . 225
3.A.11 Black Lives Matter Messaging Difference-in-Differences (DD) and Triple Difference (DDD) Estimates . . . . . 226

# List of Figures


1.1 Description of the Child Welfare System in the Evaluation County . . . . 45
1.2 Tool Display for Workers . . . . . 46
1.3 Trial Timeline and Data Sources . . . . . 46
1.4 Tool Predicted Risk and Investigator-Assessed Risk . . . . . 47
1.5 Tool Use when Tool Made Available . . . . . 48
1.6 Harm Index by Predicted Risk Level and Intervention Status . . . . . 48
1.7 Tool Availability and Targeting of Screen-Ins . . . . . 49
1.8 Impact of Tool Availability on Dynamic Harm Index over Time . . . . . 50
1.9 Supervisor Screen-In Propensity and Child Outcomes . . . . . 51
1.10 Tool Availability and Racial Disparities in Investigations . . . . . 52
1.11 Tool Availability and Screen-In Sensitivity to Predicted Risk . . . . . 53
1.12 Tool Impact Heterogeneity by Ongoing Case for Family . . . . . 54
1.13 Tool Availability and Re-Referrals . . . . . 55
1.14 Level Effect of Tool on CPS-Related Outcomes . . . . . 56
1.15 Impact of Tool on Screen-Out for "No Allegation of Abuse or Neglect" . . 57
1.16 Most Distinctive Discussion Phrases, Control and Intervention Groups . . 58
1.A.1 Team Decision Document . . . . . 64
1.A.2 Trial Description . . . . . 65
1.A.3 Randomization Database Flowchart . . . . . 65
1.A.4 Tool-Predicted Risk and Investigator-Assessed Risk . . . . . 66

1.A.5 Risk Score and Re-Referrals . . . . . 67
1.A.6 Risk Score and Preventable ER Visits over Time . . . . . 67
1.A.7 Harm Index by Predicted Risk Level and Intervention Status . . . . . 68
1.A.8 Injury Hospitalizations by Predicted Risk Level and Intervention Status . 69
1.A.9 Dynamic Harm Index Relative to First Appearance in Trial . . . . . 69
1.A.10 Association of Screen-In with Dynamic Harm Index . . . . . 70
1.A.11 Association of Supervisor Screen-In Propensity with Dynamic Harm Index 70
1.A.12 Counterfactual Estimation: Any Injury Conditional on Screen-Out . . . . 71
1.A.13 Impact of Tool on Investigations by Ethnicity . . . . . 72
1.A.14 Tool Availability and Screen-In Disparities . . . . . 72
1.A.15 Tool Availability and Screen-In Disparities, by Quartile . . . . . 73
1.A.16 Screen-In Heterogeneity by Race . . . . . 73
1.A.17 Distribution of Predicted Risk by Race . . . . . 74
1.A.18 Score Predictive Power by Race . . . . . 74
1.A.19 Tool Availability and Sensitivity to Risk – Alternative (Margins) Plot . . 75
1.A.20 Impact of Tool on Any Response to Family . . . . . 76
1.A.21 Tool Impact Across Supervisors . . . . . 77
1.A.22 Impact of Tool Availability on CPS-Related Outcomes, by Risk Score . . 78
1.A.23 Impact of Tool on Screen-Out for Other Reasons . . . . . 79
1.A.24 Impact of Tool, by Select Allegation Types . . . . . 79
1.A.25 Screen-In Sensitivity to Predicted Risk in Intervention and Control Groups over Time . . . . . 80
1.A.26 Screen-In Sensitivity to Predicted Risk over Time, Statewide . . . . . 80
1.A.27 Most Distinctive Discussion Phrases, Control Group . . . . . 81
1.A.28 Most Distinctive Discussion Phrases, Intervention Group . . . . . 82
1.B.1 Decision Protocol . . . . . 98

2.7.1 Distribution of Peer Delinquency and Relationship with Own Imprisonment 132

2.A.1 Child Welfare Context and Relation to Literature . . . . . 141
2.A.2 Continuum of Care . . . . . 142
2.A.3 Risk Score Accuracy, All Kids . . . . . 143
2.A.4 Assignment Process . . . . . 143
2.A.5 Peer Measures: Entire Placement vs. Beginning of Placement . . . . . 144
2.A.6 Distribution of Peer Imprisonment Risk and Relationship with Own Imprisonment . . . . . 145
2.A.7 Distribution of Youth Imprisonment Outcomes by Provider . . . . . 146

3.10.1 Descriptive Cyclical Patterns of Conference Attendance and Outcomes by Graduation Cohort . . . . . 201
3.10.2 Flight Cancellations, Conference Attendance, and Outcomes . . . . . 203
3.10.3 Descriptive Long-Run Giving Trajectories of Attending vs. Non-Attending Registrants . . . . . 204
3.10.4 Flight Cancellations and Outcomes: Heterogeneity by Grade . . . . . 205
3.A.1 Relationship between Flight Cancellations and Conference Attendance . . 212
3.A.2 Descriptive Long-Run Staff Employment Trajectories of Attending vs. Non-Attending Registrants . . . . . 213
3.A.3 Event Study Estimates of Attendance on Giving Over Time . . . . . 213
3.A.4 On Staff per Quarter, Among Conference Attendees . . . . . 214
3.A.5 Giving Among All Registrants, by Race and Conference Year . . . . . 215
3.A.6 Difference in Giving for Attendees vs. Non-Attendees, by Race and Conference Year . . . . . 215
3.A.7 Future Staff Employment Among All Registrants, by Race and Conference Year . . . . . 216
3.A.8 Difference in Future Staff Employment for Attendees vs. Non-Attendees by Race and Conference Year . . . . . 216

# Chapter 1

# The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial[1]

[1]We are grateful to Rhema Vaithianathan, Larissa Lorimer, and the Centre for Social Data Analytics (CSDA) for their indispensable support and feedback throughout the evaluation. We also thank Lukas Althoff, Leah Boustan, Janet Currie, Jiwon Choi, Diag Davenport, Ellora Derenoncourt, Joseph Doyle, Natalia Emanuel, Rachel Glennerster, Karin Hederos, Anne Karing, Ilyana Kuziemko, Thomas Le Barbanchon, David Lee, Matthew Lindquist, Erik Lindqvist, Alex Mas, Jessica Pac, Emily Putnam-Hornstein, Leeat Yariv, and conference and seminar participants at the NeurIPS Workshop on Human and Machine Decisions, the Princeton Center for Health and Wellbeing, the Princeton Industrial Relations Section, and SOFI for their valuable comments and suggestions. The Princeton Center for Health and Wellbeing provided generous funding for this project. We thank members of the partner agency for their detailed comments and feedback. This randomized controlled trial is supported as part of a wider research program on the use of risk modeling in human services led by Rhema Vaithianathan (Centre for Social Data Analytics, Auckland University of Technology, and The University of Queensland) and Emily Putnam-Hornstein (UNC Chapel Hill). Evaluation is one of several safeguards in the implementation of these tools. Both authors received IRB approval for the analysis from their respective institutions. This experiment was preregistered on the AEA RCT Registry under ID AEARCTR-0006311. The authors have no additional disclosures.

## 1.1 Introduction

A pivotal challenge in government and service provision more broadly is how to best allocate scarce interventions to recipients. Across a range of policy domains, from home visiting programs to criminal sentencing, improving the targeting of (non-)interventions could yield far-reaching social benefits. Machine learning-based tools hold the potential to improve the allocation of interventions to recipients (Kleinberg et al. 2018), particularly in contexts where a wealth of administrative data are available and where worker turnover is high. In many high-stakes contexts where algorithmic tools could be used, however, human experts must make the final decision. The combined effect of algorithms as used by experts in practice is relatively unknown in most policy domains.

This paper studies the impact of human-algorithm interaction on efficiency and equity in a challenging policy context: Child Protective Services (CPS). Each year, CPS agencies must decide how to respond to four million calls of alleged maltreatment involving eight million children (Children's Bureau 2021).[2] These time-constrained decisions – often made in a matter of minutes – present an allocation problem. On the one hand, failing to investigate a child suffering from untreated abuse or neglect can result in continued victimization and, in some instances, death.[3] On the other hand, an unnecessary investigation is a misuse of agency resources and can be traumatic and destabilizing for families, particularly low-income, Black, and Native American families who are disproportionately exposed to law enforcement contact.[4] Child welfare experts acknowledge there is likely some misallocation of CPS in-

[2]Child maltreatment is defined as child abuse (e.g., physical or sexual abuse) or neglect (e.g., caregiver incapacitation or hazardous living environment). 75 percent of substantiated cases of maltreatment are for neglect. One in three children in the United States are investigated for alleged abuse or neglect before age 18 (Kim et al. 2017), and at least one in eight will be a confirmed victim (Wildeman et al. 2014).

[3]Approximately 1,800 U.S. child deaths each year are attributed to maltreatment, making it one of the largest causes of mortality among young children (U.S. Department of Health and Human Services 2021). This estimate is likely a lower bound on the actual number of maltreatment-related deaths.

[4]"[F]amilies find child welfare investigations profoundly disruptive, humiliating and even traumatic. Caseworkers making unannounced visits strip-search children looking for bruises and peer into refrigerators and around homes looking for signs of bad parenting. One A.C.S. worker ... compared the experience to being stopped and frisked for 60 days." https://www.nytimes.com/2022/11/22/nyregion/nyc-acs-racism-abuse-neglect.html

terventions: fewer than 20 percent of investigations result in substantiated maltreatment, and more than half of first-time confirmed maltreatment victims were previously reported but not investigated.[5] Maltreatment is a significant social problem affecting at least one in eight children in the United States (Wildeman et al. 2014) and associated with a significant increase in criminal activity and reductions in future economic wellbeing (Currie and Widom 2010; Currie and Tekin 2012). Accordingly, child welfare practitioners have taken an interest in tools that can help improve worker decisions.

We design and evaluate a randomized controlled trial to study the effects of an algorithmic tool that predicts a child's likelihood of being removed from their home due to abuse or neglect, in partnership with a child protection agency. The trial randomizes the availability of a recent reputable machine learning-based decision support tool that uses family administrative data history to predict a child's likelihood of being removed within two years, and provides a risk score ranging from 1 to 20. We first show that the algorithmic tool has theoretical value for decision-making. Although the algorithm relies on information that is already accessible to workers, we find that it predicts children's future CPS involvement and injuries better than traditional worker risk assessments alone, even though the tool was not trained on these outcomes, and even in contexts where workers have additional information after visiting the home. Although workers were not required to consult the tool (and associated risk score), workers choose to do so in at least two-thirds of instances when the tool is available.

We estimate beneficial impacts of providing algorithmic information on child wellbeing. Using a relatively novel linkage of statewide hospitalization records to child welfare administrative data, we find that giving workers access to the tool reduces children's future hospitalization for conditions potentially related to maltreatment. Specifically, we estimate a 32 percent reduction in injury-related admissions and a 0.08 standard deviation reduction

[5]Authors' analysis using California statewide child welfare referrals from 2014-2019. 19 percent of investigations resulted in substantiated maltreatment. 50 percent of first-time confirmed victims had previously had a screened out (uninvestigated) referral.

in a combined hospitalization index that includes confirmed incidents of abuse and neglect. Further, we find a modest reduction in future allegations of abuse or neglect, suggesting that the tool may improve decision quality. We present evidence that access to the algorithmic tool enhanced the targeting of screen-ins (investigations), with screened-out children in the intervention group having less severe hospitalization histories compared to screened-out children in the control group for whom the tool was not available. We supplement experimental analysis with natural variation in the assignment of supervisors to children to provide new suggestive evidence that screen-in reduces injuries and that the tool improved outcomes for low screen-in-propensity workers.

Access to the algorithmic tool reduced racial disparities. Racial disproportionality is a significant concern in law enforcement and social service contexts, and policymakers and academics have expressed concern that algorithmic tools that rely on potentially-biased historical data could widen inequities (Obermeyer et al. 2019). In the trial county, Black children make up fewer than two percent of the county's child population yet comprise four percent of the county's CPS-involved children and eight percent of children in foster care, comparable with national disparities. Consistent with a behavioral framework of automatic and deliberative thinking (e.g., Kahneman 2011, Agan et al. 2022), workers in the control group investigate a disproportionately high share of Black children, but when given the tool they investigate low-risk Black children at rates comparable to those for other low-risk children.

Next, we consider mechanisms behind the effects of tool availability on worker decision-making and discover an unexpected result. In the absence of a tool, workers more often investigate children predicted as high-risk, but once the tool is available workers surprisingly become *less* likely to investigate predicted high-risk children and more likely to investigate low-risk children compared to the control group. We find no evidence of offsetting changes in the intensive margin of interventions, alternative service and law enforcement visits, spillovers to the control group, nor evidence of increases in time to make a decision. We also find

substantial heterogeneity in effects depending on whether the family already has an open case: For referrals (calls) of families with an ongoing intervention, tool use increases team sensitivity to predicted risk. Supplemental quasi-experimental designs, including a regression discontinuity (RD)-based approach and a child fixed effects strategy, yield similar findings. This counterintuitive set of findings highlights the potential unintended impacts of human-algorithm interaction and the importance of real-world empirical evaluation of algorithmic tools. One practical takeaway is that, even with a highly predictive risk assessment that aligns with professional objectives, efficiency cannot be measured purely by risk assessment compliance.

With this puzzle in mind, we process proprietary text data written during team discussions and provide suggestive evidence that algorithmic information caused workers to use supplemental information more effectively. When the tool is available, workers more often mention particular allegation details ("no mark or bruise") that are not algorithm inputs and record other details about statutory abuse or neglect definitions. Furthermore, among high-risk children, providing algorithmic information shifted investigations toward children with greater past hospitalization histories, despite this history only being observed by the econometrician and not the workers themselves. These findings suggest that the tool caused workers to focus on particular details from the present allegation that supplement information from the algorithmic risk score, leading them to make better decisions than in the absence of the tool.

This study is among the first to implement a randomized controlled trial (RCT) to evaluate how professionals use algorithmic information in a social services or criminal justice context. In contrast to lab-like experiments (Green and Chen 2019 using Amazon Mechanical Turk), our intervention takes place in a high-stakes field setting where decision-makers face professional incentives. In contrast with other RCTs that focus on a single decision-maker (e.g., Imai et al. 2020 evaluate one judge's use of a public safety assessment), our findings generalize across three independently-minded decision teams that vary by day in worker

composition. This study's sample includes more than a dozen supervisors and hundreds of social worker team members. Finally, though not exclusive to our study, the decision aid tool used in this evaluation has substantial independent support from the scientific community (Goldhaber-Fiebert and Prince 2019), and several agencies across the country are considering implementing it.

Emerging findings on the theoretical and actual impacts of algorithmic decision aids have been mixed. On the one hand, seminal work by Kleinberg et al. (2018) has established the hypothetical potential of machine learning tools to improve bail decisions by reducing both crime and jailing rates.[6] Although consumers exhibit aversion to using decision tools in certain contexts (e.g., health insurance decisions: Abaluck and Gruber 2016, Bundorf, Polyakova and Tai-Seale 2019), Gruber et al. (2020) show in contrast that providing algorithmic tools to professional agents advising senior Medicare enrollees improves enrollment outcomes. Their study and others such as De-Arteaga, Fogliato and Chouldechova (2020), who show that CPS workers successfully override algorithmic glitches, demonstrate that algorithms and human experts can complement one another's shortcomings, yet can also provide broad improvements over low-performing workers (Hoffman, Kahn and Li 2018 re: job testing for potential hires). On the other hand, decision-makers can override recommendations in ways that introduce new inefficiencies. Stevenson and Doleac (2019) find that the introduction of a risk score in felony sentencing did not reduce recidivism rates in part because judges were likely to override risk score recommendations for young, high-risk defendants. Judges have also been shown to override risk scores in ways that may systematically favor or disadvantage particular groups, for example being less lenient toward Black defendants in bail decisions (Albright 2019). As algorithmic decision aids continue to be adopted in high-stakes policy domains, it is imperative to understand their real-time impacts in these

[6]We too find evidence for the theoretical potential of machine learning decision tools: the tool used in this study is predictive of higher rates of high-priority hospitalization over time, which is usually unobserved by CPS social workers, yet at the moment of CPS contact hospitalization is high for both predicted low and high-risk kids, implying that algorithms can provide additional predictive value for latent risk prior to the moment of crisis (Figure 1.A.6).

settings.

This paper provides four additional contributions to advance the study of algorithm impacts and the changing nature of high-skilled work (Frank et al. 2019; Cao et al. 2021). First, our trial is designed to credibly speak to the overall distributional effects of algorithmic information. Important prior research of early-stage decision tools has relied primarily on quasi-experimental, discontinuity-based designs that are more suited to estimating the effects of tool use on a context-specific cutoff margin (Stevenson and Doleac 2019; Albright 2019). Randomization in our setting allows us to estimate the full distributional effects of the tool. Notably, we find considerably different effects of providing the tool across the distribution of child risk. Our empirical results also show substantial heterogeneity in the effects of the tool by the decision-maker's own familiarity with the family in question, highlighting personnel case-specific knowledge as an important feature to consider in the rollout of decision aids (Currie and MacLeod 2020; Cuddy and Currie 2020). Second, our study provides novel medium and long-run estimates of the impact of tool availability on family outcomes. This paper is most similar to important ongoing work by Fitzpatrick, Sadowski and Wildeman (2022), who study the implementation of a similar child welfare prediction tool using a randomized controlled trial but, to the best of our knowledge, are limited in detecting statistically significant long-run effects due to a different choice of randomization unit. Our trial randomizes tool access at the household (mother) level, allowing us to credibly estimate the dynamic effects of tool availability on children over time compared to a counterfactual set of children for whom the tool was never available. Third, incorporating text and timestamp data allows for increased understanding of how experts complement algorithmic information in practice. Lastly, variation in supervisor behavior allows us to study the heterogeneous impacts of new technologies on user productivity (Choi et al. 2022).

Finally, this study offers further insights about the impacts of public services on child health. While prior studies have evaluated the effect of downstream CPS interventions such as case openings and foster care placement on crime, academic achievement (Doyle 2007*b*;

Doyle 2008; Gross and Baron 2022; Baron and Gross 2022; Bald et al. 2022), and parents' service utilization (Grimon 2022), this paper provides some of the first evidence of the impact of CPS policy on children's health and wellbeing (Doyle 2007*a*; Chorniy and Mills 2022).

The remainder of the paper is organized as follows. Section 2 describes the empirical context of Child Protective Services and the decision tool. Section 3 describes the trial design. Section 4 discusses the sources of data used in the paper and presents descriptive statistics. Section 5 introduces the estimation framework. Section 6 presents and discusses key results. Section 7 concludes.

## 1.2 Child Protection and Decision Tool Context

The goal of the child welfare system, also referred to as Child Protective Services (CPS), is to strengthen families and keep children safe from abuse and neglect.[7] Its objective is thus closely related to a prediction problem: which families should child welfare workers visit? This section provides a brief description of the child welfare system, accompanied by a description of the machine learning tool developed to aid decision makers.

### 1.2.1 Child Protective Services in the Evaluation County

Figure 1.1 provides a visual depiction of the child welfare system in the evaluation county. In the United States, CPS first learns about a child who is at risk of maltreatment through reports ("referrals") to a child protection hotline. These referrals are typically phone calls from mandated reporters such as teachers, law enforcement, and medical professionals who are required by law to report any instances of suspected abuse or neglect, as well as additional calls from friends, family members, neighbors, and anonymous reporters. In practice, some allegations are merely expressions of concern, and some are malicious false reports. CPS must decide which allegations warrant an investigation.

[7] In this paper, we use the terms "child welfare", "Child Protective Services", and "child protection" interchangeably.

A hotline worker records the allegations of maltreatment and asks additional questions about the alleged incident, the alleged perpetrator(s), the child's family composition, and other demographic information. Hotline workers record information about home address and guardian identity, often known to the reporter, and attempt to match individuals to person-level records already in the agency's administrative data system. The hotline worker is responsible for evaluating if an allegation is truthful and warrants the creation of a referral record. In some situations, the hotline worker may decide a child requires an immediate investigation within hours, for example due to an emergency or risk of the family leaving county jurisdiction, regardless of the time of the call. For all other referrals, the referral is sent to a decision team that meets the following weekday morning.

The decision tool was developed for worker discussion teams, which handle remaining allegations of abuse or neglect.[8] Teams consist of a group of social workers, each led by at least one supervisor. There are typically three teams that meet each weekday morning to each review referrals that arrived since their last meeting. For example, referrals called in on Thursday afternoon would be reviewed by a team on Friday morning. Referrals called in after the Friday meeting and during the weekend would be reviewed on Monday morning. Each referral is only sent to one team and typically a family that is being re-reported is sent back to their original discussion team.

Worker teams discuss the referral collaboratively by writing relevant information on a document called a framework (Figure 1.A.1). The framework document includes information about family structure, child welfare history, alleged perpetrators' criminal records in the state courts system, potential strengths and weaknesses pertaining to a supportive home environment, and allegation details. A member of the team – often a designated supervisor – will record information during an in-person or virtual team discussion for all the members to see. At the conclusion of recording information, the team holds a discussion and writes

[8]Discussion team referrals are typically less urgent than emergency reports that are investigated directly from the hotline, although certain allegations may still indicate a severe risk of harm and warrant a swift response.

down additional steps for following up on the referral. Ultimately, a team will choose either to send an investigator to the family ("screen in") or not ("screen out"). 30 percent of referrals are screened in for investigation in our analysis sample, and 70 percent are screened out. For some screen-outs, a family may receive a visitor who offers optional voluntary services. Otherwise, the screened-out family receives no contact; they often remain unaware that they were even reported.

When a team decides to screen in a family, they also decide if they are going to screen in the family for assessment on one of two tracks: a low-risk track or a high-risk track. The low-risk track is part of a differential response framework common in child welfare, where the level of intervention with a family is reduced and certain findings do not need to be recorded. Families screened in on low-risk tracks can be switched to a high-risk track within the first 45 days, but the reverse is not true. Teams also determine how fast to send someone to visit the family: immediately, a 3-day response, or a 5-day response.

Finally, if a family is screened in for investigation, they are assigned to an investigator. An investigator records additional information about the allegation from visiting the home and from interviews with alleged victims, perpetrators, and reporters. As is the case in many child welfare agencies, the investigator is required to fill out risk and safety assessments within 60 days, which document strengths, weaknesses, and current and future risk of harm. The investigator decides whether to substantiate any of the allegations and whether to recommend any voluntary services, mandatory services, or remove the child from the home and place them in foster care.

The machine learning tool was designed to help teams decide which families to screen in and screen out, and determine the most appropriate follow-up response. Team objectives are difficult to quantify but align approximately with reducing future instances of severe abuse or neglect. These instances, when observed, often result in a child being removed from the home, which the tool was designed to predict. In future sections we describe how the tool is also predictive of other outcomes of interest to the agency. Although child welfare workers

are not typically evaluated based on performance metrics, agencies have an incentive to avoid negative news coverage from extreme events such as child death, subject to a constraint on the number of investigations workers can carry out and which reports are legally justifiable to screen in under statutory definitions of abuse or neglect.

### 1.2.2 Decision Tool Background

The decision aid tool used in this study was developed by Rhema Vaithianathan and colleagues at the Centre for Social Data Analytics (CSDA). Like two previous iterations of this tool in Colorado and Pennsylvania, the tool predicts a child's likelihood of being removed from their home (i.e., placed in foster care) by Child Protective Services within two years of the referral using a LASSO regularized logistic regression. Although more flexible and unstructured methodologies, notably random forests and XGBoost, were considered for the algorithm, the LASSO model was chosen for runtime purposes. The model was trained on a dataset of past child welfare referrals in the intervention county and other statewide counties using historical data from 2010-2014 on approximately 500,000 child-referrals. The model uses several hundred features relating to demographics (excluding race/ethnicity and disability), past child welfare involvement of household members, as well as limited case management data with social services. The initial choice of which variables to include for the model was made in agreement with local implementing partners. Although more variables were included, most of the predictive power of the tool comes from child welfare history of household members. The area under the receiver operating characteristic curve (AUC) of this model is 76.2 percent, which indicates good discrimination.

Given the behavioral challenges of interpreting probabilities, predictions from the model were binned into 20 equally sized groups, similar to the implementing partner's deployment of tools in Allegheny County (PA) and Douglas County (CO). The lowest-risk five percent of children thus receive a score of 1 and the highest-risk five percent receive a score of 20 for the full sample on which the model was trained (not just the county of this study). The

highest risk level (20) is associated with a high probability of future foster care placement, a relatively rare outcome that is considered a proxy for risk of harm and need for intervention.

Figure 1.2 shows an image of the algorithmic tool used during the trial. The tool and its associated risk score was designed as an aid to decision-making. Social workers were highly encouraged to consult the tool, but the tool was not meant to replace human decision-making. Workers were even specifically trained at the start of the trial with examples to make sure they did not make decisions solely using the tool. The tool included no information about why the family was called in on the current referral. However, this additional allegation information is relevant for a decision. For example, suppose a high-SES family with no past child welfare history was referred with a clear instance of sexual abuse. This family would have one of the lowest predicted risk scores, but it would be considered professional best practice to screen in the referral. An opposite situation could also occur, where a family has substantial child welfare history and receives a high-risk score, but the present referral allegation is minor and does not meet the state's definition of what is legally justifiable for a child welfare visit. The machine learning risk score, therefore, was only intended to be one part of a worker's decision determination.

The score was made exclusively accessible during the team meetings where workers decide whether or not to screen in a family. The reason for restricting access to the score at this stage was to minimize the effect that the tool might have on workers' assessments of the child's safety when they visited the home. Teams of workers were asked to write down the score in the notes from their discussion, however, so it cannot be ruled out that an investigating worker visiting a family for an assessment knew the score if he or she consulted the discussion notes or happened to be a member of the meeting that discussed the referral.

An important design choice is not only the content of information presented to decision-makers, but the *timing* of when that information is presented (c.f., Lieder et al. 2018 for a review and interpretation of anchoring). When deciding whether to present algorithmic information at the beginning or middle of the decision process, we balanced the potency of

the intervention with worker preferences for tool use. Following a pre-trial pilot in September and October 2020, workers decided that they preferred to consult the score in the middle of the of the decision process: after going through the details of the family and prior to their discussion. We found no significant differences in tool use across pre-trial treatment arms, and therefore implemented the workers' preferred workflow. For each referral, a member of the team reads out loud all call details while another member of the team fills out the framework used for decision-making. In particular, teams construct a genogram of the household (all family and non-family relationships), list all sources of danger or harm being faced by the child, risk factors, child vulnerability, past history, cultural considerations, any safety measures already taken, protective factors, and gray areas. After this first stage is completed, one of the workers in the team meeting – typically the supervisor – consults the scores for the children on the referral. Each child's score is listed on the tool interface, and the maximum score of all children in the household is made most salient. Workers then begin a discussion about the most appropriate response until the team reaches a unanimous decision.

Notably, the predicted likelihood of a child being removed increases exponentially at the top of the score distribution. The placement probabilities on the bottom-left corner of Figure 1.2 were available to the worker looking up the score, to remind them that a high risk score is disproportionately likely to be at risk of foster care. For example, below a score of 15, the difference in likelihood of foster care placement is relatively small, but above 15 each additional score indicates a substantially increased likelihood of future removal, ranging from less than 0.5 percent for a child with a score of 1, to 14 percent for a child scored 15, to 48 percent for a child with a score of 20. One concern is potential worker misperceptions about probability. Importantly, the 1-20 score bin was made more salient to teams than the associated probability: teams were asked to write down the risk score on a scale from 1 to 20, thus conceptualizing risk in an ordinal rather than cardinal way, which is a format consistent with other types of risk assessment measures already used by the agency.

## 1.3 Trial

### 1.3.1 Intervention Design

We designed a randomized controlled trial (RCT) to estimate the causal effect of the decision aid tool on CPS decisions and child outcomes. The first time a household is seen during the trial, we randomized whether or not the algorithmic risk score is observable to workers. The decision of whether to investigate a household is common to all children in the home, so intervention was randomized at the household level. For each referral, workers would look up the referral in the machine learning tool interface. When a family was in the intervention arm of the trial, the scores of all children in the household were shown, with salience placed on the maximum child score, also called the referral risk score (Figure 1.2). When the family was in the control arm of the trial, the interface did not show any of the scores and workers would continue their discussion of the referral as usual (Figure 1.A.2).

One novel contribution of our study design is that randomization is instantiated at the household level, such that once in the trial families retain their original intervention or control status. This strategy provides a best approximation for estimating the effect of interest: outcomes in a world where the tool is available compared to a counterfactual world without the tool. The design provides advantages relative to other designs that suffer from estimation challenges. For example, team-day randomization would result in the same family being re-randomized over time, and a difference-in-differences design would be plagued by rapid changes in agency and seasonal case composition across regions, such as occurred during the COVID-19 pandemic.

Ensuring that households kept their original intervention status was difficult in a setting where households are often non-traditional in structure. Multiple families living in the same household and unstable parental relationships are more common among child-welfare-involved families. Our solution randomized biological mothers to either intervention or control, and then associated their children to the biological mother's intervention status.

In case of intervention status conflict in the household over time, our design required families to be re-randomized. As will be discussed in Section 1.3.3, however, this was not without implementation errors. The advantage of maintaining household randomization intervention status is to ensure that we could estimate the long-run effects of tool availability by ensuring that the control group would not have a score shown during the trial period. We find that families indeed do get called in multiple times, with the average family in our sample being referred just more than twice from November 2020 through July 2022 (Table 1.1).

Unfortunately, the household randomization procedure did not work as smoothly as expected. Our implementing partners accidentally overwrote the intervention status for a small subset of referrals from November 2020-March 2021. In March 2021, the implementing partners attempted to correct the error, but in doing so inadvertently changed the procedure that was designed to re-randomize treatment when some children on a referral were in control and others were in the intervention group. From March 2021 onward all children in conflicting-status referrals were assigned to the intervention group. We therefore have a small ($N = 75$) subset of children in our sample who switch from control to intervention during the trial.[9] The March 2021 correction introduced an additional problem: children without a listed biological mother were randomly assigned an intervention status at the individual level instead of at the sibling-group level. Therefore, given the concurrent change that referrals with conflicting-status children were assigned to the intervention group, beginning March 2021 larger sibling groups were more likely to be treated. For example, a group of three siblings without a mother was assigned to control with a probability of: $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ instead of the intended $\frac{1}{2}$. Thankfully, most referrals have a mother listed and were not subject to this error. Second, for children without a mother, treatment assignment remained random conditional on the number of listed children. To address the non-randomization concern, therefore, we include dummies for each sibling size above one for children listed without a mother and first-randomized on or after March 2021.

[9] Future versions of this paper will include robustness checks for this subset of children.

### 1.3.2 Study Timeline

Figure 1.3 provides a visual illustration of the timeline for the study. The study can be divided into approximately four stages, which we refer to as (i) Pre-Trial, (ii) Phase 1, (iii) Trial, and (iv) Post-Trial.

The pre-trial period of the study consists of observations prior to August 2020. During this period, no teams had access to the tool, and conversations were held between implementing partners and the county about how to best implement to the tool.

From August 2020 through October 2020, two teams were given access to the tool for implementation testing. Each team was assigned to a treatment arm randomly each day: tool consulted at the beginning of a framework, or tool during the middle of a framework prior to the discussion. This portion of the study we call "Phase 1." We do not find significant differences between the two treatment arms for this part of the study, which was designed to test anchoring and to tailor the tool process to team preferences. At the conclusion of Phase 1 in October 2020, teams elected to use the tool during the middle of the framework (prior to discussion) during the main trial. One of the teams was purposely held out of Phase 1 of the trial so as to help estimate spillovers to the control group and general equilibrium effects of tool use on team decisions. In practice, implementation glitches resolved during Phase 1 provided an additional source of variation where the tool was randomly unavailable for many days of Phase 1 for all teams.

The main trial began on November 1, 2020. From this date forward, each referral was randomized into intervention (score available) or control (score not available). Workers were instructed to check for a score during each discussion and write down the score or write "No [name of tool]." If a referral had previously been randomized during the trial, the original intervention status was kept for the household the next time a family was called in. Households that had been part of Phase 1 were re-randomized during their first appearance in the main trial. In practice, the total number of referrals of households that previously

appeared in Phase 1 was relatively small ($N = 140$).[10] Household randomization continued through the end of the trial, March 29, 2022.

In late March 2022, the trial ended and the post-trial period began. At this point full access to the tool was given to all teams, such that the score was available for every referral. Relevant for the interpretation of results, children in the control group now had a score observable to workers if they were re-referred following the trial, April 2022 and onward. Certain long-run estimates of the impact of tool use may therefore be attenuated toward zero.

### 1.3.3 Trial Sample

The trial ran from November 1, 2020 to March 29, 2022. Over this time period, the county received 8,309 referrals involving 15,023 child-referral observations (Table 1.A.1). Our analysis unit of observation, however, is focused on a child's initial randomization in the trial. 6,852 unique children were randomized on 4,213 first-time-in-trial referrals that were successfully scored.[11]

Approximately 24 percent of randomized children are omitted from the analysis sample for various reasons. First, we exclude $N = 251$ children who were first scored during the Phase 1 pilot prior to the rollout of the full trial, as teams may have remembered their score. Children with a sibling already randomized during the trial were also excluded from the analysis, as a sibling's prior assignment could have already impacted the newly randomized child prior to their own assignment ($N = 603$). We restrict our sample to standard calls

[10] Our primary results omit households that appeared during Phase 1.

[11] Nine percent of calls could not be scored due to being high-emergency or calls that came in during the evening. In order to generate a score for each child, the child's referral needs to be transferred into the state central data system. This happens automatically at the end of each work day: the process begins around 5pm locally but typically takes 12 to 14 hours to complete. Emergency referrals that are *immediately* screened in will thus not have had time to be copied over and have a score generated. Furthermore, referrals that arrive approximately after 5pm and before 8am the following morning will be reviewed by a team before information is copied into the main data system, and therefore will not have a score generated. On certain days, the main data system is down or the transfer process took longer than usual and so the team started reviewing referrals before a score could be generated. Hence, scores will not be generated for every referral. Even with these exceptions, 91 percent of unique children referred during the trial were scored.

assigned to the normal team meeting procedure, and therefore omit $N = 217$ non-standard calls that were not seen by a discussion team, such as calls listing concerns about a foster family. Given the implementation complications listed in the previous section, we make two further sample restrictions to guarantee random assignment. First, we omit the small subset of referrals where there were multiple calls placed for the family before the first team meeting (e.g., a family referred on Friday evening, re-referred on Saturday, and both referrals are seen by a team on Monday). Some of these children were re-randomized with multiple statuses before first being seen and we therefore drop the entire group ($N = 157$). Second, we exclude all children whose original status was overwritten as a result of the implementation issue from Nov 2020-Mar 2021. We thus omit an additional $N = 378$ children. Our remaining analysis sample consists of $5,225$ unique children from $3,180$ referrals.

## 1.4 Data and Descriptive Statistics

The trial county has an extensive administrative data system, which we integrate with a tool randomization database and statewide hospitalization records.

### 1.4.1 Child Welfare Integrated Data System

The county child welfare integrated data system includes individually linked identifiers that follow each person, child and adult, on each referral over time. The data system includes datasets with detailed information about the referral (e.g., allegation types), response to a referral, investigator assessments of safety and risk upon visiting the home, case openings, foster care placements, authorized services, court proceedings, relationships of persons to one another on a referral, and individual demographic information when known. We integrate an additional data source from call logs that allow us to determine which referrals apply to children residing in the county of interest.

### 1.4.2 Risk Score Data

Our implementation partners at the Centre for Social Data Analytics (CSDA) generated a risk score for each incoming referral.[12] This process linked persons on the referral to past child welfare history, and public benefits files, and limited demographics to generate a score predicting a child's likelihood of foster care placement within two years using a model based on statewide administrative data. These real-time scores are available for most children in the study even when randomization was not successful. Notably, a child's score could change over time on subsequent referrals.

### 1.4.3 Statewide Simulated Risk Scores

For robustness checks comparing the intervention county with neighboring counties, our implementing partners created an ex-post backscored (simulated) risk score for all children on all referrals from the state prior to the trial and during the trial.

### 1.4.4 Tool Randomization Database

The tool randomization database maintained the intervention status of each person in the trial, to determine whether the tool (score) would be available for a given referral. The authors designed the randomization database to maintain consistent intervention status by household. In practice, this proved to be a complex process given changes in household composition over time: Figure 1.A.3 illustrates the database design. The implementing partners then constructed and maintained the database. Implementation errors notwithstanding (Section 1.3.1), the database provided information on each referral's randomization status and the people on the referral – usually one mother but sometimes a larger set of individuals – who contributed to the final randomization status of either intervention or control.

[12] All data processing and score generation occurred on state servers. This was a deliberate choice by the implementing partner to protect data privacy. The score was unable to be generated for the nine percent of referrals that came in overnight or where there was a system outage.

### 1.4.5 Discussion Text and Timestamp Files

Team discussion data are used to understand how the tool changed the content of team discussions. These data were exported from digital documents written for a given referral during each team decision-making process, and redacted by the implementing partner to remove names and other personal identifiers. In particular, these data include fields for follow-up next steps (the decision about how to follow up with the home and a justification for that decision) and gray areas (areas of potential family strengths and weaknesses the team was weighing during discussion and was still unsure about). Additionally, the documents include a timestamp of last edit, which is usually at the end of a discussion. These timestamps allow us to estimate the length of time to reach a decision for a given referral.

### 1.4.6 Hospital Records

Statewide hospital inpatient and emergency room (ER) records are maintained by the trial state's hospital association. We partnered with the hospital association to integrate linked hospitalization records for all children in the sample as a way to measure ground truth outcomes with respect to medical utilization and maltreatment. Records were matched probabilistically using person-level identifiers (full name, date of birth, gender) and include fields for: dates of admission and discharge, point of origin, facility type, admission priority, charged amount, diagnosis codes, and procedure codes.[13] Notably, children were matched to all hospitalizations in the state – not just hospitalizations in the trial county – to address concerns about differential migration. 86 percent of matched hospitalizations occurred within the trial county, and 93 percent occurred within the trial county or a neighboring county.

[13]In practice, match probabilities were 95 percent or greater for nearly all children.

### 1.4.7 Descriptive Statistics

Table 1.1 presents descriptive statistics from the analysis sample, with additional columns for referral type subsamples. The combined analysis sample includes 5,225 children from 3,156 family clusters listed on 3,180 first-in-trial referrals.[14] Of these, 4,514 children (86 percent of sample) first appear on a non-active referral, where the family was not receiving attention from CPS, and 711 children (14 percent) first appear on an active referral, for example where a CPS case was already ongoing for the family.

The sample is relatively balanced in terms of gender, with 52 percent girls. At least 80.5 percent of children are white (including Hispanic ethnicity), at least 4.6 percent are Black (including Hispanic ethnicity), and 14.4 percent of observations have an unknown race demographic category across all data sources, where race is more commonly known for children who have already been investigated. Similarly, Hispanic or non-Hispanic ethnicity status is missing for 36.1 percent of observations and at least 16.7 percent of children are Hispanic. As is true of many CPS agencies, Black children are overrepresented in the sample relative to the county population. Finally, children in the sample are referred on average 2.3 times during the trial, including the initial randomized referral. This average is greater for children on active referrals (3.3 per child) than for children on non-active referrals (2.2 per child), as children with active referrals tend to be at higher risk of maltreatment and are more frequently engaged with CPS.

The average referral in the sample has approximately 1.7 (median of 2) children listed; some families consist of a single newborn whereas others consist of several older children. Of these, on average 1.3 children are alleged victims and 75 percent of children are referred by a mandated reporter, for example a school teacher, physician, or police officer. 2.2 percent of children that are the subject of a team discussion in our sample are on a referral with an

[14] A family cluster is defined conservatively as any network of individuals who share a connected path with one another, as measured by common appearances on a referral. For example, if child A and child B appear together on the same referral, they are considered part of the same family cluster. If child B later appeared on a referral with child C, then A, B, C and are all considered part of the same family cluster, as there is a connected path from referrals with A to referrals with C through B.

alleged child in danger, which tends to be more common for non-active referrals. Referrals are categorized at the hotline with allegations in up to 24 categories. Alleged injurious environment, which is a form of neglect, is the most common allegation, appearing for 62.1 percent of children. Two of the other largest allegation categories are physical abuse and parental substance abuse. 20.3 percent of children appear on referrals with alleged physical abuse and 14.7 percent with alleged parental substance abuse, respectively. Other common allegation types include sexual abuse, emotional abuse, and medical neglect.

## 1.5 Empirical Strategy

### 1.5.1 Estimating Tool Impacts

The empirical strategy in this paper estimates (1) the level effect of tool availability and (2) the effect of tool availability across the risk distribution. For child $i$ first scored on referral $r(i)$, we estimate:

$$Y_i = \beta_0 + \beta_1 Tool_{r(i)} + \gamma_{s(i)} + \epsilon_i \tag{1.1}$$

and

$$Y_i = \beta_0 + \beta_1 Tool_{r(i)} + \beta_2 Score_{r(i)} + \beta_3 \left[Tool_{r(i)} \times Score_{r(i)}\right] + \gamma_{s(i)} + \epsilon_i, \tag{1.2}$$

where $Y_i$ is the outcome of interest pertaining to child $i$ on referral $r(i)$. $Tool_{r(i)}$ is a binary {0,1} intervention status indicator, equal to one if the tool (score) was available to the team during their discussion of the initially randomized referral $r(i)$ and zero if the tool was not available. $\gamma_s(i)$ are randomization procedure controls for sibling group size for referrals without a mother after March 2021, which control for an implementation glitch that led to greater likelihood of tool availability for referrals with more children among the subset of referrals with missing mothers. $Score_{r(i)}$ is the maximum score of all children listed on referral $r(i)$, which is made salient to the team during a discussion when the tool is available. Finally, $\epsilon_i$ is a random error term. Standard errors are clustered conservatively at the household (family network) level.

Equation 1.1 estimates the level effect of tool availability, $\beta_1$ on outcome $Y_i$. Under conditional random assignment of tool availability, such that $E[\epsilon_i|T_{r(i)}, \gamma_{s(i)}] = 0$, our estimate $\hat{\beta}_1$ can be interpreted as the causal effect of having the tool available for the initial randomized referral on a child's outcome of interest.

Equation 1.2, in contrast, estimates the effect of tool access for children interacted by risk level: the coefficient of interest is $\beta_3$, which estimates the effect of the interaction of intervention status and risk level ($T_{r(i)} \times Score_{r(i)}$) on outcome $Y_i$. In this specification, $\beta_1$ is similar to an intercept term: the effect of the tool on outcomes for children at hypothetical risk score zero. $\beta_2$ is the linear relationship between risk score and outcome in the control (no tool) group. For example, $Y_i$ defined as screen-in and $\beta_2 = 0.01$ would imply that, in the absence of having the tool, teams are one percentage point more likely to screen in a referral for every one unit increase in predicted risk. $\beta_3$ is the relative increase in screen-in per one unit increase risk level for children in the intervention group compared to children in the control group.

Since workers were encouraged but not required to consult the tool, our policy-relevant treatment effect is the intent-to-treat (ITT) effect of giving workers access to the tool. We find evidence of high rates of tool use across the risk distribution, such that instrumental variables (IV) estimates are qualitatively similar but with slightly larger magnitudes.

### 1.5.2 Confirming Random Assignment

Random assignment of the tool is important for estimating an unbiased causal effect of tool availability on outcomes. Before estimating equations 1.1 and 1.2, it is possible to check for evidence of conditional random assignment of tool availability by running a balance test of intervention status on salient observable characteristics. Table 1.A.2 presents a test of randomization. We do not find joint significance of observable characteristics with respect to tool availability, which is evidence that the tool was randomized successfully.

## 1.6 Key Findings

### 1.6.1 Tool Validity

In order for the tool to be useful, it must be predictive of relevant outcomes during the trial period. The algorithm was trained on statewide historical data to predict a child's foster care placement (removal from home due to maltreatment) within two years of the referral. Although placement in foster care is a relatively rare outcome, a more common set of outcomes are worker risk and safety assessments that are required to be filled out by an investigator for every screened-in family.

First, Figure 1.4 presents a comparison of tool predicted risk score and investigator-assessed risk for the control group where the score was not available.[15] We find that, even conditional on investigation, the tool is predictive of the investigator's further findings on the current referral.

Is the tool predictive of negative future outcomes for screened-out (not investigated) children as well? Figure 1.A.5 shows that, even among initially screened-out referrals during the trial, the algorithmic tool is predictive of re-referral through the end of the data collection period (July 2022), ranging from 25 percent of children at the lowest risk bin to 70 percent of children at the highest risk bin. Although re-referrals do not necessarily imply abuse or neglect, they do serve as signals of concern from the community. One may be concerned that foster care placement is only observed conditional on screen-in (see Lakkaraju et al. 2017 for discussion of selective labels). In this context, however, a large share of high-risk children are eventually re-referred to CPS; the tool's margin for intervention is in part relevant to the *speed* of intervention. Figure 1.A.5 also confirms that the tool is predictive of a much more common outcome, re-referral, that only rarely leads to foster care and is not mechanically dependent on CPS policy.

[15]The investigator risk assessment does rely on several fields using past history, such that it could plausibly be correlated with the algorithmic risk score. We find a similar strong association when assigning numerical values to the other required worker assessment, the safety assessment, which has less variation but does not rely on past history (Figure 1.A.4).

Third, Figure 1.A.6 shows that the tool is also predictive of ground-truth medical risk, where having a higher tool risk score is associated with a greater incidence of preventable emergency room (ER) visits – a signal of potential neglect – over time. The figure also shows that during the month of randomization, when a child is observed by CPS, preventable visits are elevated among low-risk children to a level comparable to the high-risk group. The figure therefore shows that although the tool is predictive of potential harm, CPS agencies may struggle to observe underlying risk at the moment of contact, when both lower- and higher-risk children have similar high levels of hospitalization. The algorithmic tool has the potential to allocate attention to children at risk of harm.

Finally, we evaluate the relative predictive power of tool predictions and worker assessments side by side. Table 1.2 presents regression estimates comparing tool and worker risk scores in the same model for outcomes that the county agency wants to minimize: removals (placements in foster care), re-referrals, re-referrals with a repeat screen-in, re-referrals when a case is already open, and number of hospitalizations with injury 60+ days after the referral. Columns 3-6 show that, for all non-rare outcomes, the tool is statistically (and in terms of magnitude) more predictive than worker assessments across a variety of outcomes.

These findings confirm that the algorithmic tool is predictive of future negative outcomes – even for outcomes it was not directly trained to predict – and is even more predictive than expert assessments that use additional information from the home. It is therefore plausible that access to a predictive risk tool could improve decision quality and child outcomes.

### 1.6.2 Tool Use

The effectiveness of algorithmic tools depends both on how they are used and when they are used (e.g., Raghu et al. 2019). Because tool use was encouraged but not required, it is entirely possible that workers chose to forego consulting the study's machine learning tool. Whether or not the tool was consulted in practice affects one's interpretation of key findings.

In this study, we measured tool use by asking teams to write down the score(s) they saw

on the discussion document and to write "no tool" or "no score" if the score was not seen. If teams do not write down a comment about the tool, we are unsure if they checked the tool or not. To measure the effect of tool availability on verified use (score at least viewed), we construct an indicator for seeing the tool score that we regress on tool availability. Table 1.A.3 shows that tool availability is associated with at least a 64 percentage point increase in tool use, relative to 0.4 percent use in the control group.[16] In the control group, 77 percent of children receive a "score not seen" comment, compared to 12 percent in the intervention group. We therefore confirm that workers tend to use the tool, consulting it in at least 64 percent and no more than 88 percent of situations when it is available.

Figure 1.5 presents a local mean regression plot of confirmed tool use across the distribution of predicted risk. Workers consult the tool across all risk levels, with a slight increase in consultation for the highest-predicted-risk referrals. We find similar rates of tool use across racial categories.

### 1.6.3 Impact of Tool on Child Health and Wellbeing

Despite CPS's primary objective of protecting children, few studies have produced causal estimates of the impact of child welfare system policies on child injury and other health outcomes.[17] To measure the effect of algorithmic information round-truth outcomes, we link statewide hospitalization records from the state's hospital association.

Table 1.3 describes the primary outcomes used in the analysis, which reflect plausible maltreatment-related hospitalizations. Following initial randomization, the average number of high-priority (emergency, urgent, or trauma-related) hospital admissions is more than 0.5 per child. A subset of these high-priority visits are likely the result of harmful home environments where a child suffers from abuse or neglect. Likewise, hospital visits with an

[16]Tool use in the control group is relatively rare and may be due to team recording errors or system glitches.

[17]Doyle (2007*a*) and Chorniy and Mills 2022 are two exceptions. Each estimates the impact of changes to payments to foster care caregivers on child injuries and other indicators of wellbeing in Medicaid claims data.

injury ICD code are another proxy for potential maltreatment, for example an inflicted injury or lack of parental supervision. Avoidable emergency room (ER) visits refer to the number of child admissions for ambulatory care sensitive conditions (ASCS, defined using Appendix 8 of Carey et al. 2017), for example an ER visit for asthma, which could result from medical neglect. Admissions with confirmed maltreatment and intentional injuries (assault, self-harm) are each coded as binary outcomes due to their rarity. Finally, the number of injuries for never-screened-in children provides an outcome for children who were never investigated by CPS during the trial period, to isolate treatment effects from improved investigation targeting. The columns of Table 1.3 show that rates of hospitalization for these potential maltreatment-related outcomes are about twice as frequent for children who are predicted as above median risk by the algorithmic tool.

The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better for the child. All outcomes are measured following 60 days of randomization to accommodate a period of potential mechanical increases in child receipt of medical care during the 60-day case determination window, such as post-investigation medical care for previously untreated conditions. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. Our primary findings are robust to alternative choices for the sample period and number of days of non-measurement after the referral date.

Table 1.4 presents regression estimates of equation 1.1 using the primary hospitalization outcomes of interest. Column 1 shows that tool availability reduced a child's future harm exposure by 0.08 standard deviations compared to children in the control group. Columns

2-5 provide additional clarity on particular outcomes. In particular, the algorithmic tool reduced the average number of future injury admissions by 32 percent: 0.051 relative to a mean of 0.158. The tool is also associated with a reduction in number of high-priority visits ($p < 0.10$) and the other outcomes of interest including confirmed maltreatment (not statistically significantly). Interestingly, column 7 reports a reduction in injuries for children never screened in during the trial, which suggests the tool improved targeting and caused teams to screen out children who were at lower risk of injury.

Given potential differences in team decision patterns for active (ongoing CPS intervention with the home) and non-active referrals, could the tool have influenced child wellbeing differently for these two groups? Tables 1.A.4 and 1.A.5 consider outcomes separately for non-active and active referrals, respectively, and show that outcomes improved for children in both groups despite different decision patterns (Table 1.12, to be discussed in the proceeding sections).

One concern is if the 60+ day window for detecting effects masks any hypothetical short-run harm cause by the tool: a hypothetical 0-60 day increase in hospitalization would likely simply be consistent with increased CPS attention and treatment of previously untreated injuries, but could still raise concerns about a short-run increase in harm. We find that our estimates of improved child health in Table 1.4 are robust to including the 0-60 day period, and Table 1.A.6 shows that tool availability is not associated with a significant change in outcomes within 60 days of the child's first trial referral. A final concern is whether children in the intervention group had lower baseline rates of injury and other harm indicators. Table 1.A.7 shows that a pre-intervention harm index is not significantly different for intervention group and control group children, ruling out a difference of more than 0.05 standard deviations at the 95 percent confidence level. Injury visits were if anything greater for the intervention group prior to randomization. We also find that the results of 1.4 are robust to including pre-randomization health outcome levels, and the intervention status randomization balance test passes after including pre-randomization health outcomes.

Does the tool benefit certain children more than others? Figure 1.6 provides a visual depiction of harm reduction across the distribution of risk scores.[18] Reassuringly, children with tool access have lower harm indices for most score levels. Gaps are largest in absolute terms for children at the highest risk levels (16-20), and this is true for individual health outcomes in absolute and percentage terms as well (e.g., Figure 1.A.8 for injury-related hospitalizations).

What causes the change in outcomes for these high-risk children? For descriptive evidence, Figure 1.7 plots the harm index gap for intervention minus control children over time, split by whether the child was initially screened in or not. The connected plot with triangle markers shows that, among children who were screened in during their first trial referral, screened-in children with tool access had higher rates of past hospitalization than children who were screened in without tool access. Likewise, the connected plot with circle markers shows the opposite for screened-out children: children who were screened out using the tool had lower past (and future) histories of hospitalization, consistent with the rightmost column of Table 1.4. This figure provides visual evidence that the tool shifted the targeting of CPS investigations toward children with greater pre-existing injury (and other hospitalization) histories.

### Dynamic Trends in Hospitalization

To further explore the temporal relationship between CPS contact and hospitalization, we estimate an event study (dynamic difference-in-differences) specification:

$$\begin{aligned} Y_{it} = \beta_0 + \sum_{k=-5}^{5} \mathbb{1}\{K_{it} = k\}(\gamma_k + \rho_k Tool_{r(i)}) + \mathbb{1}\{K_{it} \leq -6\}(\gamma_{-6} + \rho_{-6} Tool_{r(i)}) \\ + \mathbb{1}\{K_{it} \geq 6\}(\gamma_6 + \rho_6 Tool_{r(i)}) + \beta_1 X_i + \lambda_t + \epsilon_{it}, \end{aligned}$$

where $Y_{it}$ is the outcome of interest, namely the harm index, for child $i$ in period $t$. $\gamma_k$ are event time coefficients for each two month period before and after the initial trial referral, $\rho_k$

[18] Figure 1.A.7 illustrates a smoothed local mean regression visualization of the same result with confidence intervals.

are event time coefficients scaled by another variable of interest (namely the impact of tool access over time relative to control), $\gamma_{s(i)}$ are randomization controls for the small subset of children that were randomized conditional on household size, $\lambda_t$ are period-of-discussion and period-of-admission fixed effects, and $\epsilon_{it}$ is the error term.

Figure 1.A.9 plots the estimated $\hat{\gamma}_k$ coefficients without any $\hat{\rho}_k$ coefficients to illustrate the mechanical relationship between CPS contact and hospitalization. The harm index gradually increases prior to a referral (robust to excluding infants) at which point potential maltreatment-related hospitalizations rise in the days prior to, and day of, a given referral. This finding is consistent with some hospitalizations triggering a CPS referral. Figure 1.A.10 shows that CPS contact (screen-in), however, is not mechanically associated with elevated hospitalization over time – for instance if CPS sent children to the hospital for untreated injuries. Although we cannot entirely rule out countervailing effects – for example if investigation increased utilization but reduced short-run harm by an offsetting amount – this finding provides further support for the interpretation that a reduction in potential-maltreatment-related-hospitalizations for children is indeed a beneficial outcome.

Finally, Figure 1.8 presents coefficients $\hat{\rho}_k$ estimating the time path of changes in the harm index for children with tool availability compared to the control group. Children for whom the tool is available do not experience an immediate reduction in harm, consistent with an extended case-planning process that may not immediately bind, but experience an improvement (reduction) in potential-maltreatment hospitalizations beginning four months and onward following their initial trial referral. Estimates suggest that the benefits from the tool persist for at least one year.

### Worker Propensities and Counterfactuals

For a more informed interpretation of findings, we consider the effects of CPS investigation (screen-in) on children in general, regardless of the tool. Estimates of the effect of screen-in on child outcomes, however, may be confounded by unobserved differences between children who

are screened in compared to those who are not. To overcome this challenge, we use natural variation in the assignment of lead supervisors to team discussions to provide suggestive estimates of CPS intervention impacts and consider counterfactual scenarios.

In the CPS agency of our study, each decision team had between three to five main supervisors who each took turns in leading team discussions. Although the decision is made as a team, supervisors exercise discretion over how information is presented and discussed. It is therefore plausible that a child's assigned supervisor has an impact on the child's likelihood being investigated that is orthogonal to child characteristics. We estimate a leave-self-out supervisor propensity to screen-in measure popularized by Kling (2006) and Doyle (2007*a*):

$$Propensity_{r(i)j} = \frac{1}{N_j - n_{r(i)}} \sum_{k \notin r(i)}^{N_j} Investigate_{r(k)j},$$

Where $Propensity_{r(i)j}$ is a supervisor $j$'s leave-one-out propensity to screen in child $i$ on referral $r(i)$, $n_{r(i)}$ is the number of children on referral $r(i)$, $N_j$ is the total number of children assigned to worker $j$, and $Investigate_{r(k)j}$ is a binary indicator for whether another child $k$ on referral $r(k)$ seen by worker $j$ was screened in or not. This leave-self-out approach is used to break mechanical correlation between a supervisor's investigation propensity and a child's own investigation (Angrist, Imbens and Krueger 1999). This measure is then residualized by discussion team, as rotational assignment is most plausibly conditionally random within a team. Supervisors may be reassigned to families they are familiar with, which leads us to focus on first-time, non-active referrals during the trial period.[19]

Table 1.A.8 evaluates the assignment of supervisors to children and presents the association of supervisor propensity with outcomes of interest. Column 1 shows that supervisor propensity to screen in is not strongly correlated with a set of observable child and referral characteristics ($p > 0.20$ with wild bootstrapped standard errors to account for variation from a small number of supervisors), providing some support for conditional random as-

[19]The propensity measure was constructed using all non-active referrals among the three primary decision teams from January 2020 through June 2022 to maximize predictive power.

signment.[20] Column 2 shows that supervisor screen-in propensity is predictive of a child's likelihood of being screened in on a given referral. Column 3 presents OLS estimates of the standardized hospitalization index ("harm index") on screen-in, conditioning on the set of observable characteristics, including the algorithmic risk score. The OLS estimate may be biased upward if children with unobservably higher risk of harm are more likely to be screened in. To address this concern, Columns 4 and 5 present two-stage least squares estimates of the harm index on screen-in with and without controls, respectively, instrumenting for screen-in using supervisor propensity. Although the magnitude of the estimated coefficient may be overstated due to other sources of bias, such as supervisor skill or other propensities (Mueller-Smith 2015; Chan, Gentzkow and Yu 2022), the estimates are suggestive that being screened in is associated with a reduction in future hospitalizations (Figure 1.A.11 for dynamic trends).

Figure 1.9 presents visuals of the first stage and the association of screen-in propensity with outcomes. Both panels show considerable variation in residualized supervisor propensity to screen in. Panel 1 presents a local linear regression of the first stage, child screen-in on supervisor propensity to screen in, revealing a monotonic relationship. Panel 2 meanwhile shows that a child's future harm index is decreasing (improving) with supervisor propensity to screen in. While past studies have focused on the relatively extreme intervention of removing a child to foster care (Doyle 2007*b*; Doyle 2008; Gross and Baron 2022; Baron and Gross 2022; Bald et al. 2022), these findings suggest that CPS *investigation* may improve child outcomes on at least the margin of hospitalization, such that better targeting of investigations could reduce injuries. Additionally, Panel 2 shows that when the tool is made available, the index falls the most for children assigned to low-screen-in propensity supervisors (imprecisely estimated, c.f., Table 1.A.8 Column 6).

Finally, as an area of ongoing work, we use variation in supervisor propensities to compare the efficiency of supervisor decision-making with alternative regimes (Figure 1.A.12). This

[20]We caution that the balance test result in Column 1 is sensitive to specification choices, but we reassuringly find that estimates using the propensity coefficient are robust to the inclusion of salient controls.

approach is similar to Angelova, Dobbie and Yang (2022) and Arnold, Dobbie and Hull (2022), who use extrapolation to estimate outcomes conditional on non-intervention. First, for each supervisor we calculate future injury-related hospitalization rates for children who are screened out. We then extrapolate to estimate hospital injury rates for a supremely lenient supervisor who screens out every child, thus approximating the mean conditional injury risk in the sample. This rate reflects the expected conditional injury rate under a random screen-in counterfactual, regardless of the chosen screen-in rate. Finally, we consider a scenario in which the algorithm makes all decisions by re-estimating conditional injury rates under different screen-out thresholds by omitting all observations below a particular algorithm risk score level. Although our counterfactual estimates are subject to strong assumptions about worker objectives to minimize injury rates, Figure 1.A.12 suggests that supervisors at least make better-than-random screen-in decisions.[21]

### 1.6.4 Racial Disparities

A significant concern regarding the implementation of algorithmic risk assessments is their potential to perpetuate and widen inequality in treatment. On the one hand, algorithms trained on past history could embed human bias that could perpetuate the oversurveillance and/or underservice of low-SES groups (Obermeyer et al. 2019).[22] Additionally, machine learning predictions can be more accurate for majority groups, with potentially less effective or even detrimental consequences for minority groups. On the other hand, given widespread complaints about systemic bias in criminal justice and social service contexts, algorithms may be *less* biased than the status quo of human decision-makers (c.f., Hoffman, Kahn and Li 2018). Machine learning can also detect and respond to disparities often unknown to human experts (Pierson et al. 2021) or cause teams to think about allegations more carefully (Agan et al. 2022). An ambiguous implication for bias is understanding in what ways users

[21] Future editions of this paper will consider the relative efficiency of human-algorithm interaction compared to worker-only and (hypothetical) algorithm-only scenarios.

[22] We also note tradeoffs in algorithmic fairness objectives (Kleinberg, Mullainathan and Raghavan 2016).

override recommendations for particular groups (Albright 2019; Cheng et al. 2022) and exhibit preferences for disadvantaged groups in general (Kisat 2021).

Black children are disproportionately likely to receive a CPS investigation and foster care placement, both nationally and in the study county (Edwards et al. 2021). The authors and the county CPS agency were therefore each particularly interested in understanding the impacts of human-algorithm interaction on racial disparities in screen-ins.

Figure 1.10 presents investigation rates by race and risk level for intervention and control children, respectively. First, low-risk Black children in the control group are *50 percentage points* more likely to be screened in than non-Black (primarily white) children, an extremely large disparity that is qualitatively robust to excluding missing race observations from the sample (Table 1.A.9 Columns 5-8), including allegation, age, and family-size controls (Columns 2, 4, 6, and 8), and including hospital-derived race (Table 1.A.10). Second, however, lower-risk Black children in the intervention group are screened in at much lower rates, at rates comparable to non-Black children. These findings suggest that, in this context, access to the algorithmic tool reduced unwarranted racial disparities among low-risk children. We do not find evidence of a reduction in disparities by Hispanic ethnicity, but there was not a statistically significant difference in screen-ins by ethnicity in the control group to begin with (Table 1.A.11 and Figure 1.A.13). Although the county has a relatively small Black population, there are still $n = 182$ Black children in the sample when using CPS-derived race alone and $n = 242$ when including hospital-derived race (Figure 1.A.14), with at least 27 children in each race-by-risk-by-tool-availability cell.[23]

The distribution of child risk is not equal across race/ethnic groups, however. Figure 1.A.17 shows that Black children are much more likely to be scored as high-risk than other children. It is therefore possible that racial disparities could increase if the tool affected screen-in rates for different risk groups differently (c.f., Lambrecht and Tucker 2019). Columns 3-4 and 7-8 of Tables 1.A.9 and 1.A.10 show that overall racial disparities in in-

[23]More granular plots are available in Figures 1.A.15 and 1.A.16, which shows screen-ins fall for most Black children but may increase for the very highest risk children.

vestigation rates do fall by a marginally significant 8-19 percentage points depending on the chosen specification. Therefore, the algorithmic tool reduced overall racial disparities, primarily through a reduction in investigations for low-risk Black families.

The downstream effects of tool use may not be equitable if the tool is not as predictive for minority families, however. Figure 1.A.18 presents differences in the harm index by risk level for Black and non-Black children, respectively. Although the harm-risk gradient is relatively flat for lower-risk Black children, hospitalization for Black children does increase substantially at the top risk quintile, suggesting the tool is at least somewhat predictive for Black children. We do not find any adverse effects of tool use on health outcomes for Black children.

The reduction in racial disparities is an additional benefit of the algorithmic tool. It is unclear, however, to what extent this reduction in disparities would generalize to municipalities with more Black families or workers who may have different attitudes and training practices. In our context, the agency trained workers to prioritize cultural competency and sensitivity to disparities by race, ethnicity, sexual orientation, and language.

### 1.6.5 Impact of Algorithmic Information on Decisions

The impact of tool availability (intent-to-treat) on decision-making can be estimated both in absolute terms (equation 1.1) and relative to a child's risk (equation 1.2). A key decision margin of interest is the choice to screen in (investigate) or screen out (not investigate) a family.

Table 1.5 presents estimates of tool availability on the primary investigation margin. The mean rate of initial screen-in is 30.3 percent in the full sample of first-time randomized children (Column 1). Column 1 shows that a one unit increase in a child's predicted risk of future removal score bin (range of 1-20) is associated with a 0.3 percentage point increase in the likelihood of screen-in for the full sample, such that a child with a referral score of 20 is 6 percentage points (20 percent) more likely to be screened in than a child with a referral

score of 1. Making the tool available to teams does not have a statistically nor economically significant effect on the overall rate of screen-in, which is consistent with the idea that many CPS agencies are constrained in the number of investigations they can administer.

Are workers more likely to screen in high-risk children in the absence of the algorithmic tool, and does the tool increase the responsiveness of screen-ins with respect to predicted risk? Table 1.5 Column 2 presents estimates of the effect of tool availability interacted with predicted risk score level. First, workers in the control group – when the score is hidden – are sensitivity to the score, with a 0.6 percentage point increase in screen-in per unit increase in risk score. Curiously, however, making the score available *reduces* worker sensitive to risk by about 0.5 percentage points per risk score unit. Tool availability is associated with a marginal 4.9 percentage point increase in screen-in for children at the lowest level of risk, such that making the tool available increases investigations for children predicted as low-risk and reduces investigations for children predicted as high-risk. This outcome in isolation suggests that access to algorithmic risk scores reduced worker sensitivity to risk and had the opposite effect of what was intended.[24]

Table 1.5 Columns 3-6 present heterogeneity in estimates by current CPS intervention status. In particular, 86 percent of children in the sample had no ongoing CPS intervention ("Non-Active") at the time of their initial randomization in the trial. 14 percent of children, however, were members of a household with an ongoing CPS intervention, such as an open case. Child Protective Services responds to referrals differently in these two situations, and it is plausible they respond to the tool differently by context as well. Column 4 shows that the patterns in the full sample are even stronger for non-active referrals, which are less dependent on recent prior decisions. For non-active referrals, teams in the control group are 1.1 percentage point more likely to screen in a child per unit increase in risk (about 20

[24]One commonly-raised hypothesis is that, for high-scoring children, workers updated their prior probability of child removal risk from greater than 0.5 to less than 0.5, since the probability scale (ranging from <0.01 to 0.48) was available for workers to reference. This theory, however, less easily explains an increase in screen-ins for low-scoring children when they were revealed to be at less than one percent risk of removal. Furthermore, workers were accustomed to using risk scores on 1-10 and 1-20 scales, and appear to have exclusively used the 1-20 risk score itself (not the probability) in their discussions.

percentage point from 1 to 20 relative to a mean of 32 percent), but are only 0.3 percentage point more likely to screen-in per risk unit when using the tool, an 0.8 percentage point per-unit reduction in risk sensitivity (Column 4). Finally, Columns 5-6 first show that the overall rate of screen-in is indeed lower for active referrals (21 percent vs. 32 percent for non-active), in part because an intervention is already ongoing.[25] For active referrals, workers are not sensitive to risk in the control group, but do increase in risk sensitivity by 1.5 percentage point per risk unit when using the tool. The tool, therefore, has heterogeneous impacts on screen-ins depending on whether the family already has an ongoing intervention.

Figures 1.11 and 1.12 provide local mean regression visualizations of the impact of tool availability on screen-in across the risk distribution.[26] Figure 1.12 shows the strong impact for non-active referrals: an increase in screen-ins for lower risk children and decline in screen-in from about 45 percent to about 33 percent for the highest risk children.

For robustness, we leverage two quasi-experimental designs, (1) discontinuities in displayed score and (2) changes in score for a child re-referred over time, to validate our experimental findings. The regression discontinuity (RD)-based design exploits the feature that risk predictions were generated as a continuous variable, but scores were shown to workers as a binned integer value, leaving variation in predicted risk within a given score level. The estimated equation is as follows:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 Tool_{r(i)} + \beta_2 Score_{r(i)} + \beta_3 \left[Tool_{r(i)} \times Score_{r(i)}\right] + \rho_1 LassoProb_{r(i)} \\ + \phi_1 LassoProb^2_{r(i)} + \sum_{\tau=2}^{19} \left[\theta_\tau \gamma_\tau + \rho_\tau (LassoProb_{r(i)} \times \gamma_\tau) + \phi_\tau (LassoProb^2_{r(i)} \times \gamma_\tau)\right] + \epsilon_i, \end{aligned} \tag{1.3}$$

where $Y_i$ is the outcome of interest (screen-in) for child $i$, $Tool_{r(i)}$ is a binary indicator of tool availability for child $i$ on referral $r(i)$, and $Score_{r(i)}$ is the 1-20 score bin appearing on

[25] For an active referral, a repeat screen-in involves an additional visit to the family along with further assessment of risk and safety.

[26] Trends are robust to other types of estimation, such as a margins plot approach (Figure 1.A.19). Trends also apply when considering interventions broadly, such as a decision to screen a family out but refer them to resources (Figure 1.A.20). The findings also hold within the subsample of children randomized during the first year of the trial.

referral $r(i)$. We then flexibly control for a quadratic relationship between the outcome and the tool's hidden predicted probability, $LassoProb_{r(i)}$, separately for each score discontinuity cutoff $\tau$, where a score cutoff is defined by placing each observation into the lower or upper half of its observed score bin. For example, $\rho_1$ is estimated from children who have an above-median risk probability in $Score_{r(i)} = 1$, or who have a below-median risk probability in $Score_{r(i)} = 2$. $\gamma_t$ are indicators for different cutoff levels ranging from $\tau = 2$ (above-median 2 and below-median 3) to $\tau = 19$ (above-median 19 and below-median 20). Finally, $\beta_3$ is the coefficient of interest, measuring the effect of a one unit increase in observed score in the intervention group relative to the control group, after controlling for trends in underlying risk.

Table 1.A.18 presents RD estimates pooled across 19 score discontinuities, showing that a one unit increase in score level is associated with a one percentage point reduction in screen-in, and a one percentage point reduction for any visit to the family in general, for children listed on non-active referrals. This estimate is consistent with the experimental estimate of a 0.8 percentage point reduction in screen-in likelihood per unit increase in risk (Table 1.A.16).

A second robustness check is a child fixed effects strategy that uses changes in child risk level over time. For children who are re-referred to CPS, their predicted risk score is updated such that the score level may increase or decrease. Table 1.A.19 presents estimates from a child fixed effects model showing that a one unit increase in risk score corresponds to a 1.2 percentage point reduction in screen-in likelihood for children on non-active referrals, and a 1.9 percentage point increase in screen-in likelihood for children on active referrals, again consistent with the experimental estimates (Tables 1.A.16 and 1.A.17). These designs provide further credibility to the conclusion that tool access decreases worker sensitivity to predicted risk for non-active referrals and increases sensitivity for active referrals.

One threat to identification is a form of contamination bias where tool availability changes worker behavior for children in the control group. Table 1.A.20 tests for the presence of short-

run, within-day spillovers by regressing a team's tool use, screen-in decisions, and discussion time on the randomization status of a team's previous referral(s) that day. We find no evidence of within-day spillovers, such as a crowd-out of decision time or change in risk sensitivity. Second, Figure 1.A.25 presents estimates of screen-in sensitivity to risk score by quarter, among non-active referrals, for the intervention (tool available) and control (tool not available) groups. Recall that for non-active referrals, our estimates imply the tool reduces risk score sensitivity such that teams screen in more low-scored referrals and screen out more high-scored referrals, relative to the control group. Under contamination, we would expect that control group risk sensitivity would converge toward the intervention group. Figure 1.A.26 shows that control sensitivity estimates do not appear to mechanically fall over time, at least for the first five quarters of the trial.[27] To the extent that there remain undetected changes in worker decision patterns, our study estimates would likely be a conservative lower bound of the true impacts of tool availability.

Finally, we consider the robustness and potential external validity of these findings by comparing impact heterogeneity across different supervisors.[28] The agency has three teams with approximately a hundred affiliated workers per team. Teams meet and interact relatively independently of one another, with workers participating alternatively in the team discussion meetings of five workers. Three to four main supervisors per team take turns overseeing team discussions. Figure 1.A.21 shows that the impact of the tool on screen-ins is qualitatively similar across different supervisors, suggesting some generalizability of the results. Each of the eight most frequent supervisors during the trial has a neutral or positive investigation-risk gradient when discussing children in the control group, and a reduction in the gradient slope when discussing children in the intervention group.

[27]Figure 1.A.26 presents similar estimates using limited statewide data provided by the implementing partner. Difference-in-differences estimates using the holdout team do not provide evidence of contamination, and are omitted due to imprecision.

[28]These findings are similar to considering heterogeneity across the three primary discussion teams.

### 1.6.6 Effect of Tool on Other Outcomes

The previous section showed that tool availability surprisingly increased screen-ins for low-risk referrals and reduced screen-ins for high-risk referrals, relative to the control group. Did the tool's impact on screen-ins coincide with other changes in team decisions? And did the tool impact downstream outcomes, such as investigator findings and future calls to CPS?

Table 1.A.12 presents the level effect of the tool on a range of CPS-related decisions and outcomes in the full sample. The *Team Decisions* panel shows that tool availability is not associated with significant changes in ever being screened in during the trial, receiving an expedited (within 72 hours or faster) investigation response time, the intensity type of investigation receive, receipt of a light-touch family visitor, or in time to arrive at a decision, which is derived from discussion timestamp data. Tables 1.A.15 and 1.A.16 likewise show no evidence of compensating for changes in screen-ins by risk level: high-risk children are no more likely to receive a high-priority investigation, expedited response time, or family visitor when the tool is available.

The *Investigation* panel of Table 1.A.12 shows the effect of tool availability on family assessment decisions, conditional on investigation. The tool does not have a strong effect – in terms of both magnitudes and statistical significance – on investigator risk assessment findings, whether services are recommended to a family, and number of days from referral until investigation closure. There are likewise few changes by risk level (Tables 1.A.15-1.A.17).

Finally the *Outcomes* panel presents two types of child outcomes: injuries discovered by CPS on the initial investigation and future calls to CPS for the same child. We find a marginally significant 3.7 percentage point increase in the rate of discovered injuries when using the tool, suggesting that the tool may have improved targeting of visits to maltreated children. Injuries that are discovered are typically sustained prior to investigation, and so would not be affected by the type of treatment a family receives as a result of the most recent referral. However, the relative rarity of this outcome ($n = 100$) and the lack of change in

risk assessment findings suggests caution in overstating the impact of the tool on targeting from this result alone. We also find that the tool did not increase re-referrals, future calls to CPS for the same child that are proxies for community concern and potential maltreatment. In fact, despite the reduction in risk sensitivity, we find a marginally significant ($p = 0.101$) 4.5 percentage point reduction in the share of children re-referred, relative to a full sample mean of 53 percent, among children randomized during the first half of the trial who had a full year post-referral for observation. The reduction in re-referrals is relatively stronger for low-risk children, but the effect is still neutral or negative for high-risk children as well (Figure 1.13).

Figures 1.14 (level effect) and 1.A.22 (by score) present visualizations of the primary estimates of tool availability on CPS-related decisions and downstream outcomes. Despite a surprising reduction in predicted risk sensitivity when using the tool, these results suggest that tool use did not harm children and potentially improved conditions.

### 1.6.7 Team Discussions

Tool access reduced worker sensitivity to predicted risk, yet improved child outcomes and reduced racial disparities. What can reconcile these findings? Workers clearly do not naively follow the tool (which would imply an increase in sensitivity), nor do they appear to make arbitrary or entirely contrary decisions when using the tool either (inconsistent with improvement in outcomes and reduction in racial disparities for low-risk children). Although investigators who visited the home did have access to the risk score if it was written in the discussion notes, we do not find any marked changes of tool access on their behavior either, such as risk assessment scores, service recommendations, or time to case closure. For further insight on mechanisms of tool use, we therefore turn to written discussion notes recorded by teams during their decision-making process.

Figure 1.15 Panel 1 shows that the tool's impact on screen-ins (and screen-outs) is primarily in reasons for screen-out that involve "no allegation of abuse or neglect", such as

a call that mentions a child complaining about their parents but does not include specific allegations of bruising, emotional abuse, or other maltreatment. However, there is no significant effect of the tool on screen-outs for other reasons, such as sending referrals to law enforcement (Figure 1.A.23). Furthermore, when considering pre-recorded information on allegation type, tool availability appears to lead to a reduction in screen-ins for referrals with alleged parental substance abuse and an increase in screen-ins for referrals with alleged sexual abuse (Figure 1.A.24). Taken together, these findings suggest that tool use changed workers' attention about particular details from the present allegation, which was not included in the algorithm's risk calculation.

Text analysis performed on worker discussion documents provides further evidence of how the tool changed discussion content. Figure 1.16 provides a descriptive example of the most distinct frequency-weighted word sequences (flexible length n-grams) for discussion fields in intervention group relative to the control group. Reassuringly, the phrases "no score generated" and "tool: none" are among the most distinctive text phrases in the control group, appearing more than five times as often as in the intervention group. Phrases most distinctive to tool availability include "no allegation of abuse or neglect", "no mark or bruise", "physical abuse", and "defined by law no [abuse or neglect]". These disproportionate word frequencies suggest that tool use is associated with greater team attention to – or at least written justification of – particular allegations and home conditions on the present allegation that complement historical and demographic risk factors summarized by the algorithmic tool.[29] These findings bolster evidence from the hospitalization data that providing workers with algorithmic information improved the targeting of interventions to children (Table 1.4 column 7 and Figure 1.7)

[29]Our partner agency has expressed interest in forthcoming text analysis that will use variation in pre-recorded reporter descriptions of the alleged incident.

## 1.7 Conclusion

Machine learning tools have drawn increasing interest from public policy practitioners, yet our understanding of the effectiveness and equity of such tools when paired with human decision-makers is limited. Child Protective Services (CPS) presents an important context for evaluation. Like many public agencies, CPS departments have (i) long-accumulated histories of administrative data that are useful for predictive risk modeling, and (ii) require high-stakes, time-intensive decisions from workers with limited experience and high turnover. Whether such tools enhance worker decisions and specifically improve equity, however, has been subject to speculation.

This paper provides one of the first randomized evaluations of an algorithmic decision tool in criminal justice or social services and advances new approaches for evaluating such tools. Our family-based randomization strategy can in principle be applied to other contexts that require estimating long-run impacts of decision tools on client outcomes. Similarities in our findings across independent teams and supervisors suggest there may be some external validity to the results of this study. Future studies in other geographic areas and service contexts, however, will be important for generalizing our knowledge of the effects of human-algorithm interaction on racial disparities and other important outcomes. An area ripe for future inquiry is bounding the gains from effectively pairing algorithmic predictions with private information in other high-stakes contexts, and estimating the extent to which technical change from ML tools is skill-biased or skill-replacing in these settings (Acemoglu 2002).

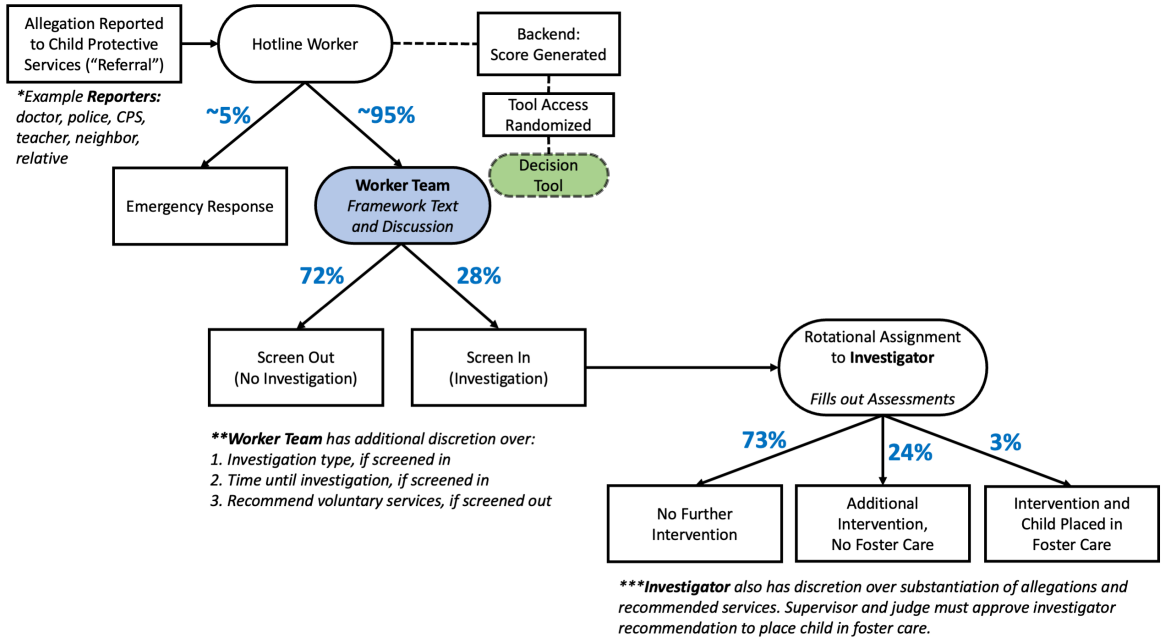
Although the use of predictive risk modeling is valuable for allocating the marginal investigation in this context, our discussions with workers highlight the limitations such tools have over discretion: Workers are statutorily required to have supporting (non-tool) evidence to screen in a referral of abuse or neglect, and particular allegations require an investigation regardless of predicted risk. An optimal agency must therefore screen in at least a fraction of predicted low-risk reports, and cannot screen in all high-risk reports. Furthermore, many

of the children and families at greatest risk are those who interact with CPS repeatedly, for whom more information is already known. Consequently, supervisors have expressed interest in tools that provide additional insight on case planning for families who are already heavily involved in the system. Identifying the most beneficial decision points for deploying new algorithmic tools, improving the explainability and presentation of tool predictions,[30] and understanding the impact of particular CPS interventions on families remain critical areas for future research.

[30]Anwar et al. (2022) show that the presence of a public defender increases judge concurrence with a pretrial risk assessment tool. In the absence of a human mediator, explainability may yield similar benefits.

## 1.8 Figures and Tables

Figure 1.1: Description of the Child Welfare System in the Evaluation County




Notes: Listed percentages are the share of children in the county assigned to the respective decision, conditional on reaching the given node. For example, three percent of investigated children are placed in foster care within 90 days.

Figure 1.2: Tool Display for Workers

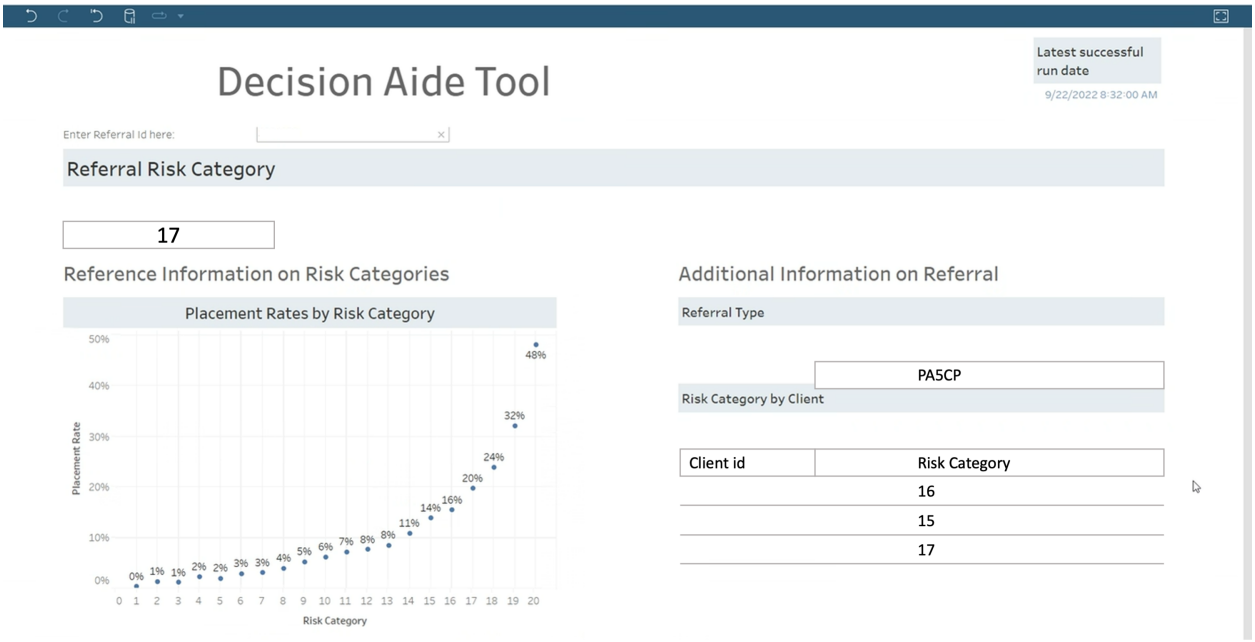
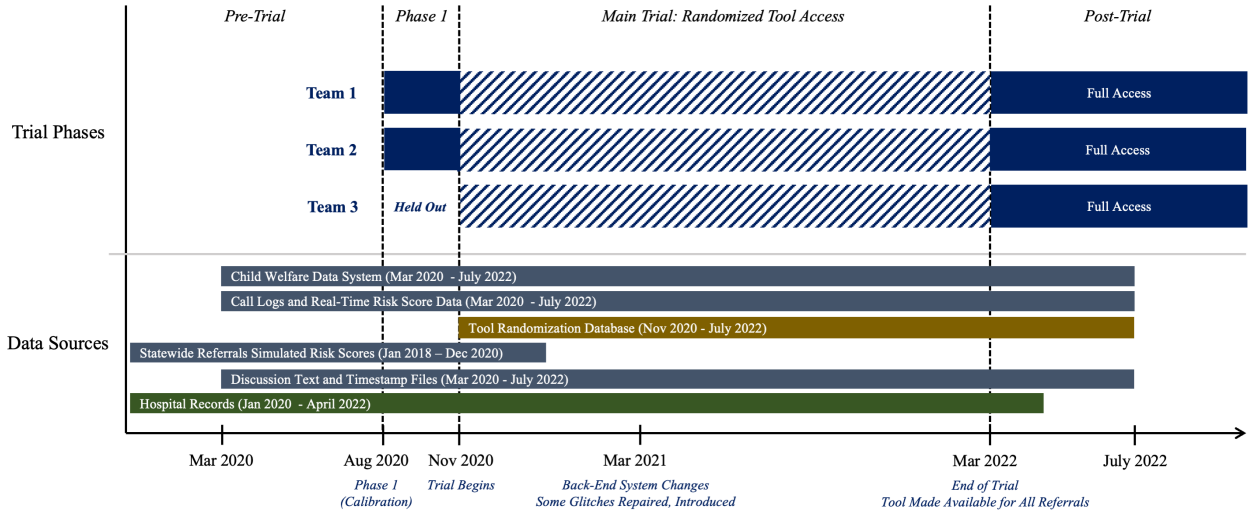
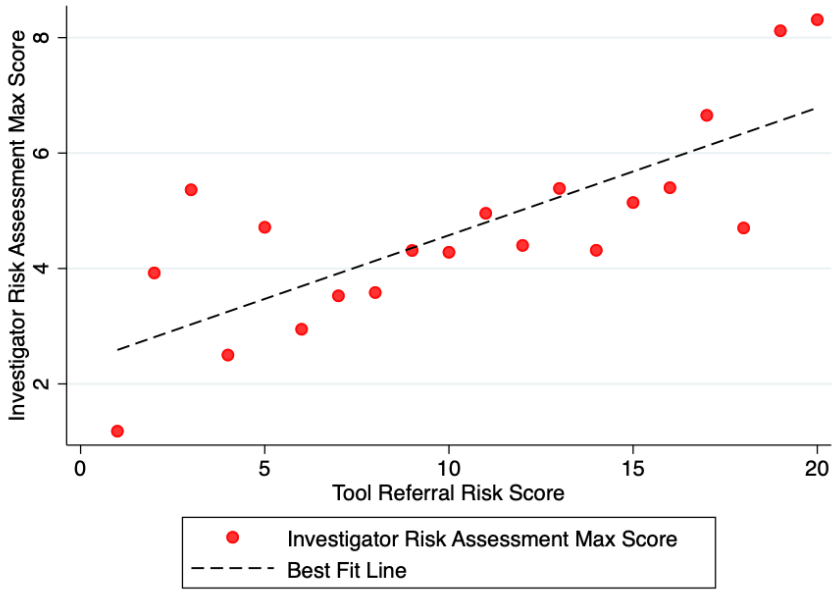



Figure 1.3: Trial Timeline and Data Sources



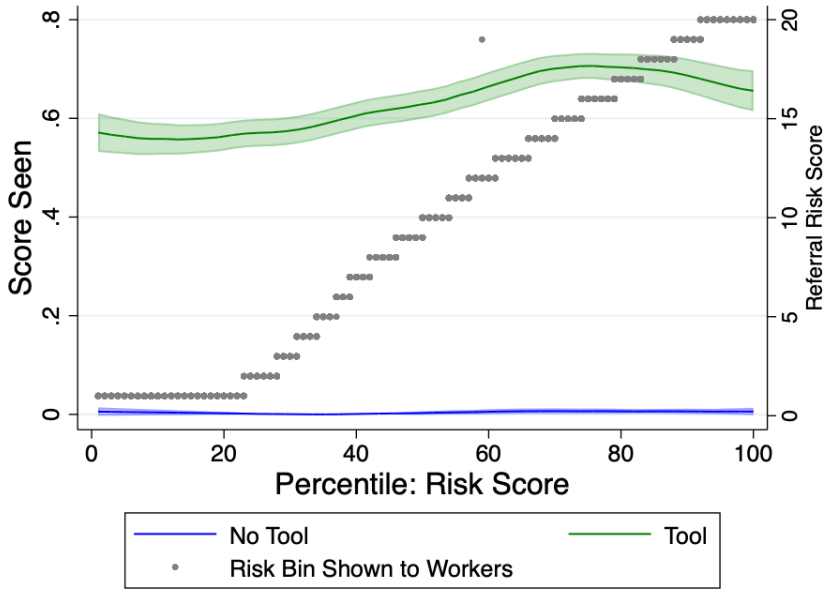

*Notes:* Phase 1 varied the timing of when the algorithm risk score was revealed to teams 1 and 2: (i) at the beginning of recording information or (ii) later, during the middle of the framework process, at the beginning of discussion. Team 3 was withheld from Phase 1 as a control group to help evaluate whether giving workers the tool changed their decisions when they did not have the tool. Following the conclusion of Phase 1, all teams opted to use the tool during the middle of the framework, at the beginning of discussion.

Figure 1.4: Tool Predicted Risk and Investigator-Assessed Risk



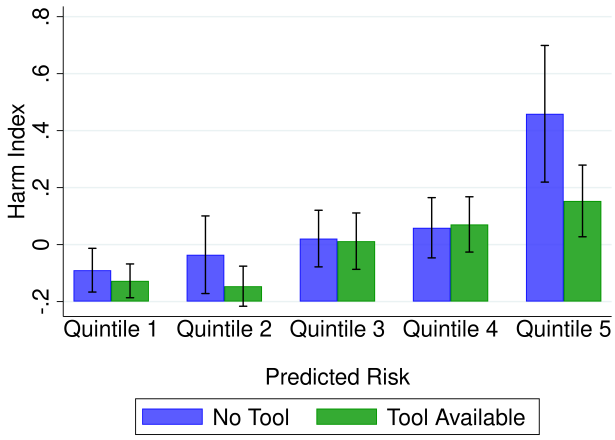

Notes: This figure presents a scatter plot of the relationship between an investigator's risk assessment numerical evaluation of a child's risk (maximum of the worker's assigned abuse and neglect risk scores) on the vertical axis and the decision tool's predicted risk of future foster care on the horizontal axis. The sample is restricted to the control group (score not seen). Worker assessments are only performed for screened-in children.

Figure 1.5: Tool Use when Tool Made Available



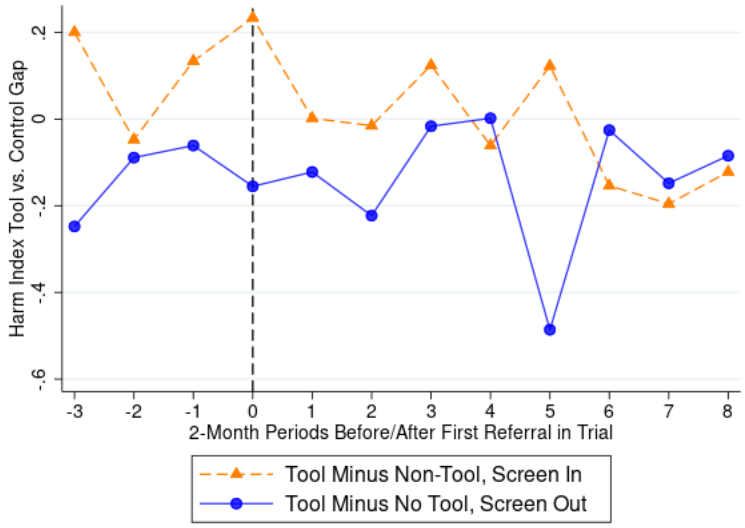

Notes: This figure presents local mean regressions of whether the team wrote down seeing a risk score (left vertical axis) on predicted referral risk percentile (horizantal axis) for the intervention and control groups, respectively. Observations where a team did not mention any comments about the tool are coded as zero, such that the levels are a lower bound on whether the score was actually seen. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.6: Harm Index by Predicted Risk Level and Intervention Status



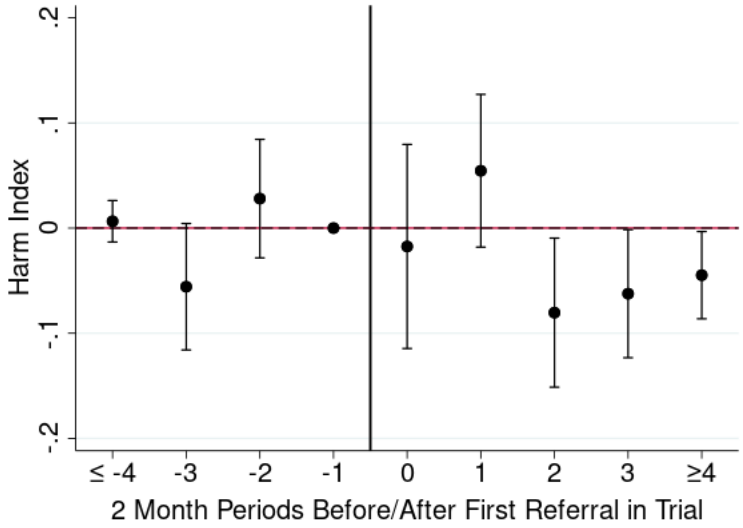

Notes: This figure presents mean child hospitalization index outcomes by score quintile and tool availability cell. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

Figure 1.7: Tool Availability and Targeting of Screen-Ins




Notes: This figure presents time trends in child health outcomes (dynamic harm index) for children in the intervention group minus the index for children in the control group, by initial referral screened-in and and screened-out status, respectively. The sample is restricted to high-risk (score 16-20) children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. The sample also omits a small number of referrals with multiple randomized clients that were more likely to be treated, for visual clarity. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

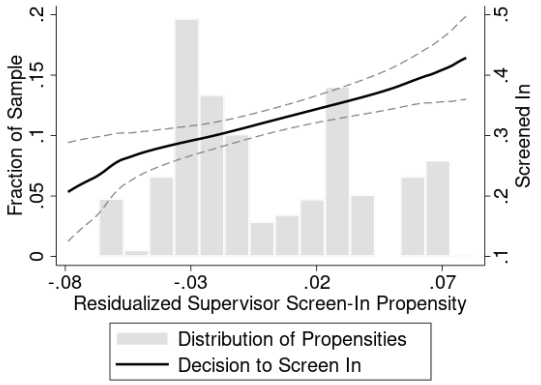
Figure 1.8: Impact of Tool Availability on Dynamic Harm Index over Time



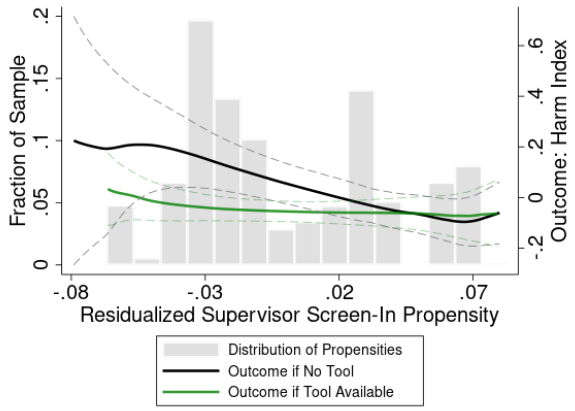

Notes: This figure presents event study estimates of child health outcomes (dynamic harm index) for children with the tool available compared to children in the control group, relative to their first referral in the trial at period 0. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

Figure 1.9: Supervisor Screen-In Propensity and Child Outcomes

(1) First Stage

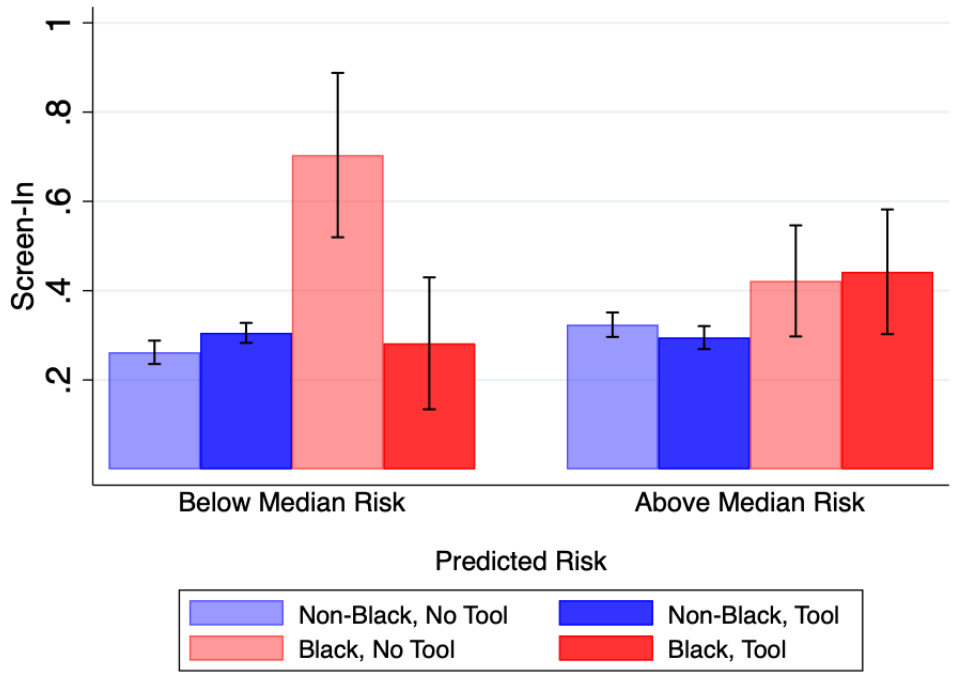



(2) Outcomes, by Treatment Status



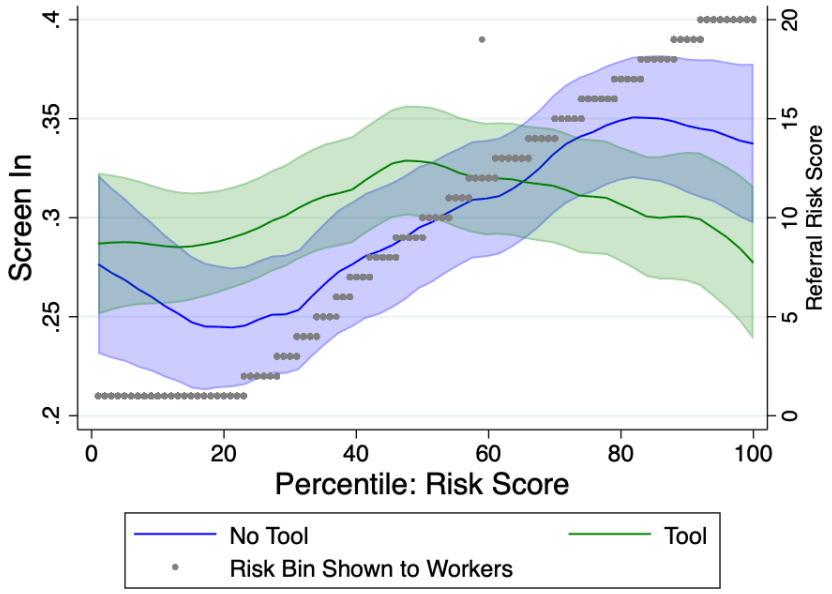

Notes: This figure presents the relationship between supervisor leave-self-out propensity to screen in, screen-in decisions, and child outcomes. The histogram in each panel shows the distribution of a supervisor's rate of screen-in for children on all other referrals, residualized after controlling for team fixed effects. The solid line in Panel 1 is a local linear regression of child screen-in on supervisor propensity to screen in. The solid lines in Panel 2 are local linear regressions of a child's harm index on supervisor propensity to screen-in, by intervention status. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. The sample consists of all first-time observations of children on non-active referrals in the first half of the trial among the three primary decisions teams and where a supervisor had at least 100 child observations in the remaining sample.

Figure 1.10: Tool Availability and Racial Disparities in Investigations




Notes: This bar plot presents mean screen-in rates by race (Black vs. other) and intervention status, for children below and above median predicted risk, respectively. Race is coded only for children for whom race is known to CPS at the time of referral.

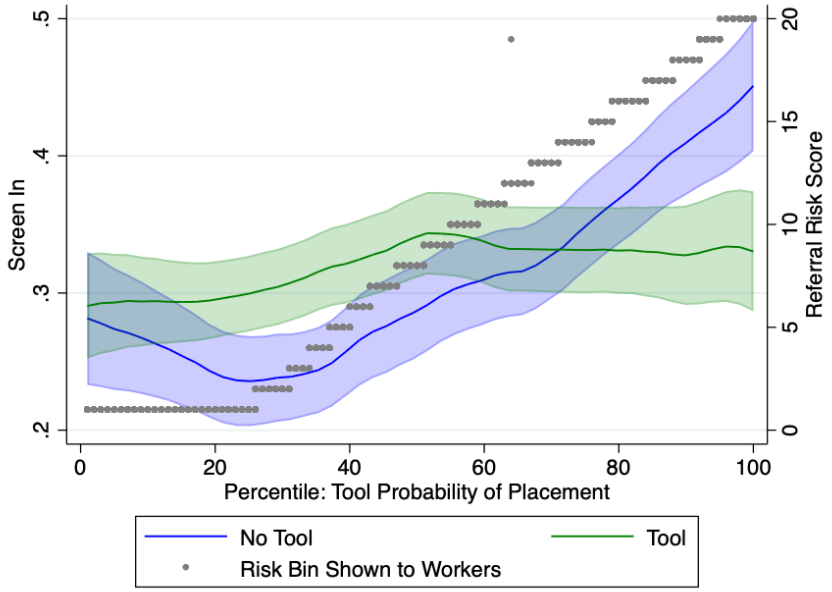
Figure 1.11: Tool Availability and Screen-In Sensitivity to Predicted Risk



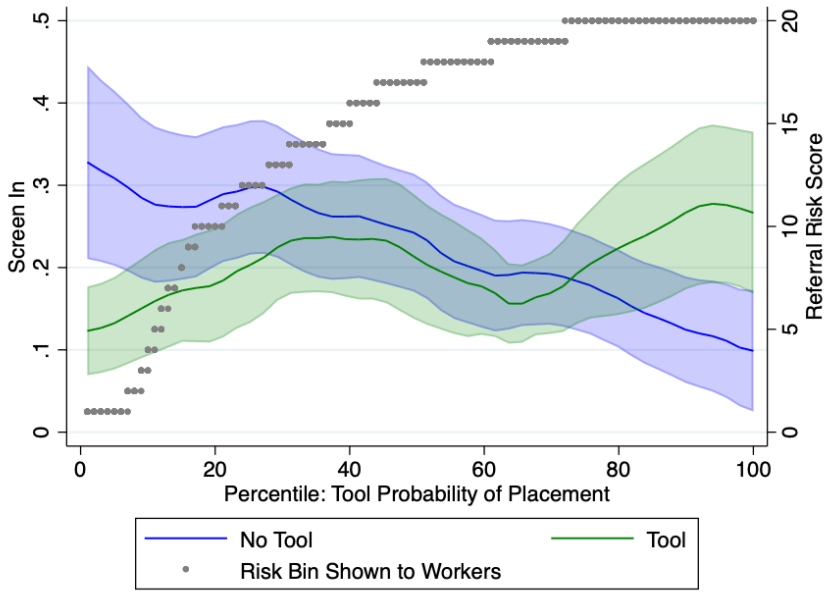

Notes: This figure presents local mean regressions of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.12: Tool Impact Heterogeneity by Ongoing Case for Family

(1) Non-Active (86% of Sample)

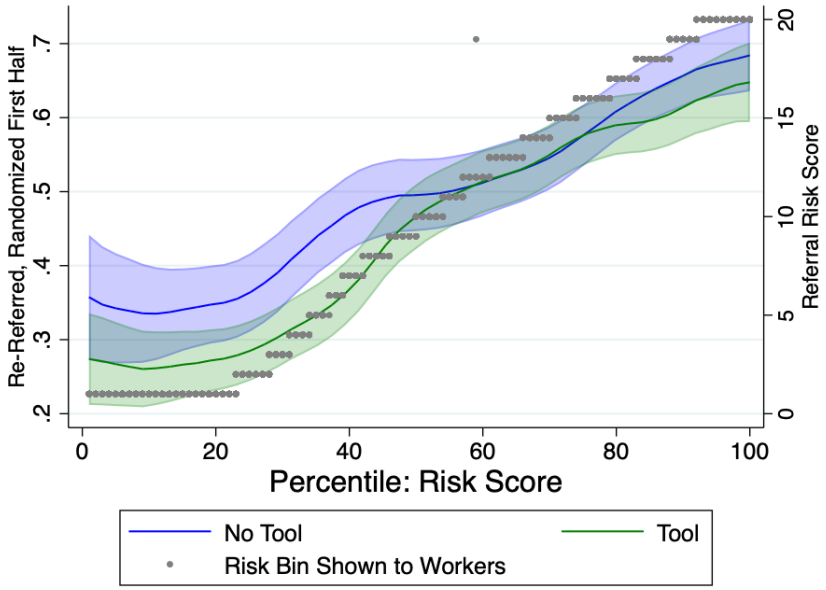



(2) Active (14% of Sample)



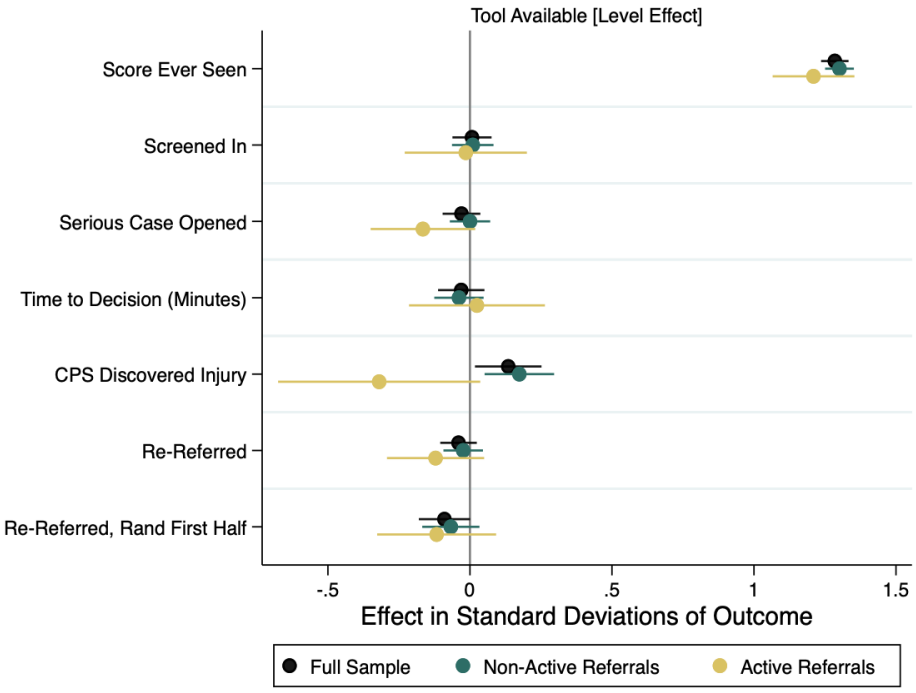

Notes: This figure presents local mean regressions of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, for active and non-active referrals, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. 86 percent of child observations in the main sample are listed on non-active referrals. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.13: Tool Availability and Re-Referrals




Notes: This figure presents local mean regressions of any future re-referral on predicted referral risk score for the intervention and control groups, respectively. Sample restricted to children randomized during the first half of the trial (November 2020 - June 2021). Re-referral outcome measured through July 2022. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

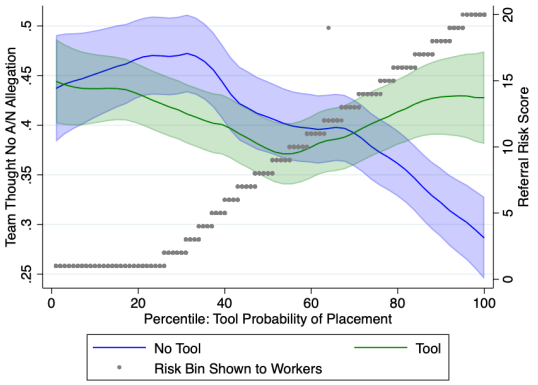
Figure 1.14: Level Effect of Tool on CPS-Related Outcomes



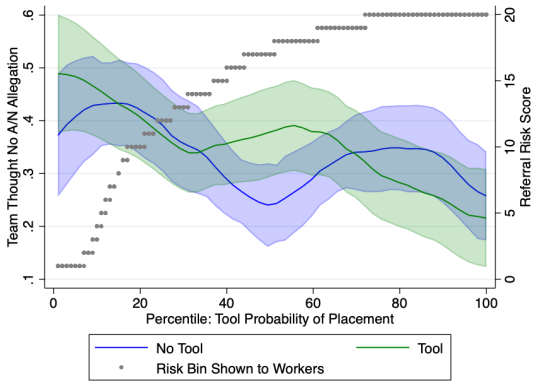

Notes: This figure presents point estimates and standard errors for the coefficient on tool availability in estimating equation 1.1 for a variety of outcomes. Estimates are presented with standardized outcome variables (mean 0, variance 1) for comparability of coefficient magnitudes across different outcomes. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. Standard errors are clustered at the household level and are shown at the 90 percent confidence level.

Figure 1.15: Impact of Tool on Screen-Out for "No Allegation of Abuse or Neglect"

(1) Non-Active

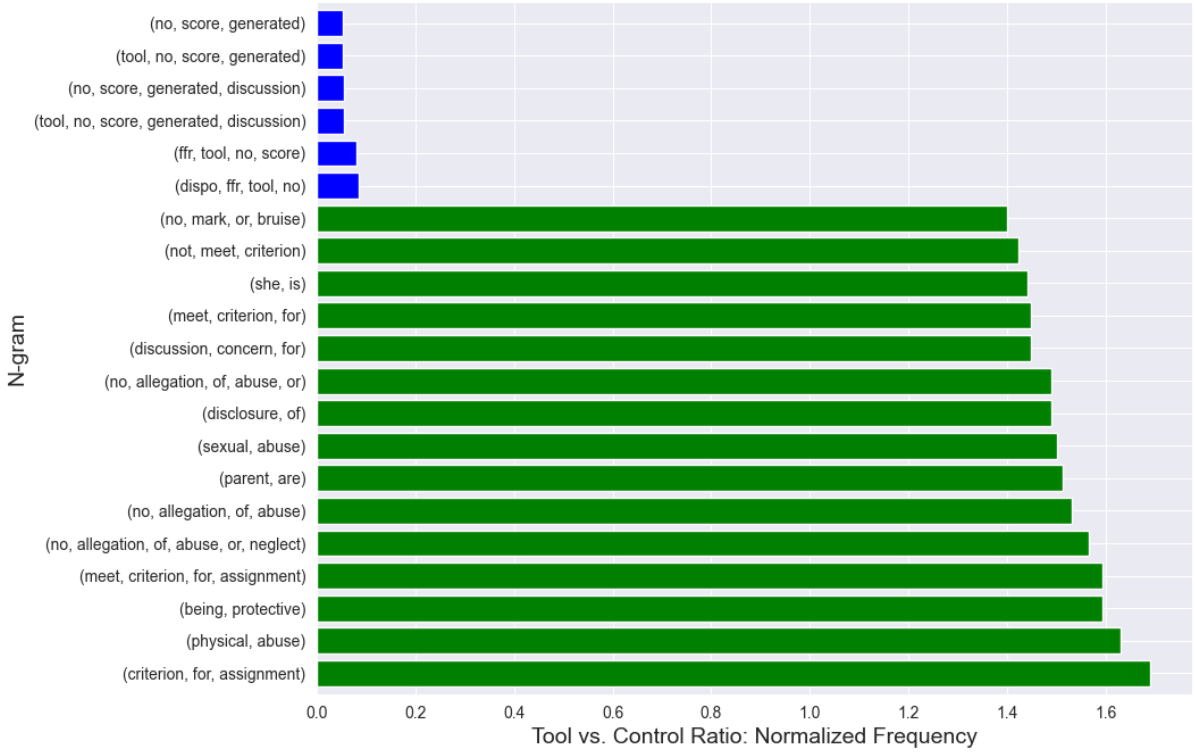



(2) Active




Notes: This figure presents local mean regressions of a child being screened out due to the allegations not being legally defined as abuse or neglect on predicted referral risk score for the intervention and control groups, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.16: Most Distinctive Discussion Phrases, Control and Intervention Groups




Notes: This figure presents $n = \{2, 3, 4, 5, 6\}$ n-grams (consecutive word sequences) found in discussion text, sorted by relative frequency in the intervention group compared to relative frequency in the control group. For example, consecutive pairs of words most unique to the tool are listed near the top with high frequency ratios. N-grams with frequency ratios between (0.3, 1.4) are omitted from the display for concision. Figures 1.A.27 and 1.A.28 provide additional distinct n-grams for the control and intervention groups, respectively.

Table 1.1: Descriptive Statistics

| | Full Sample | Hospital Outcome Sample |
|---|---|---|
| *Panel A: Child Characteristics* | | |
| Female | 0.52 | 0.51 |
| Age at Initial Randomization | 10.52 | 10.47 |
| White | 0.80 | 0.83 |
| Black | 0.05 | 0.05 |
| Missing Race | 0.14 | 0.12 |
| Hispanic | 0.17 | 0.18 |
| Missing Ethnicity | 0.36 | 0.31 |
| Number of Times Referred | 2.32 | 2.57 |
| Re-Referred and Screened In | 0.11 | 0.13 |
| Removed after Referral Date | 0.02 | 0.03 |
| *Panel B: Referral Characteristics* | | |
| N Children Listed on Referral | 1.69 | 1.70 |
| Mandated Reporter | 0.75 | 0.73 |
| Neglect Allegation | 0.77 | 0.77 |
| Abuse Allegation | 0.19 | 0.19 |
| N Unique Children | 5,225 | 3,790 |
| N Referrals | 3,180 | 2,296 |
| N Family Clusters | 3,156 | 2,284 |

Notes: This table describes characteristics of referred children in the primary analysis samples in this paper. Active clients are members of a household where at least one person had an ongoing CPS intervention at the time of referral.

Table 1.2: Tool Predictive Power Relative to Workers

| | (1) Removed | (2) Removed | (3) Re-Referred | (4) Repeat Screen In | (5) Re-Referred When Open | (6) N Injury Visits 60d+ |
|---|---|---|---|---|---|---|
| Tool Risk Score | 0.0068* | 0.0047 | 0.017** | 0.014** | 0.019*** | 0.0046 |
| | (0.0029) | (0.0028) | (0.0056) | (0.0051) | (0.0052) | (0.0044) |
| Worker Risk Score | | 0.0070 | 0.010 | -0.0021 | 0.0018 | 0.00092 |
| | | (0.0055) | (0.0086) | (0.0088) | (0.0091) | (0.0075) |
| Mean of Outcome Variable | 0.062 | 0.062 | 0.531 | 0.237 | 0.250 | 0.159 |
| Observations | 320 | 320 | 320 | 320 | 320 | 320 |

Notes: This table presents estimates of the relationship between tool risk predictions and outcomes relative to worker risk assessment scores on a similar scale. The sample is limited to screened-in referrals in the control group that were randomized during the first half of the trial. Tool risk score is defined as the maximum child score for the referral. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.3: Summary Statistics for Main Hospitalization Outcomes

| | Below Med Risk | Above Med Risk |
|---|---|---|
| N High Priority | 0.367 | 0.687 |
| N Injury Visits | 0.119 | 0.195 |
| N Avoidable ED | 0.076 | 0.194 |
| Any Maltreat ICD | 0.004 | 0.013 |
| Inten Injury | 0.009 | 0.017 |
| N Injury, Never Screened In | 0.092 | 0.196 |
| Harm Index | -0.116 | 0.112 |
| N Child Observations | 1,853 | 1,931 |

Notes: This table presents descriptive statistics of child hospitalizations in the randomization sample. Column 1 describes post-referral hospitalizations for children scored by the algorithm as below-median risk of future removal, and column 2 for children scored as above-median risk. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. Outcomes are measured beginning 60 days after initial randomization. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization.

Table 1.4: Impact of Tool Availability on Hospitalization

| | (1) Harm Index | (2) N High Priority | (3) N Injury Visits | (4) N Avoidable ED | (5) Any Maltreat ICD | (6) Inten Injury | (7) N Injury, Never Screened In |
|---|---|---|---|---|---|---|---|
| Tool Available | -0.082** | -0.087* | -0.051** | -0.021 | -0.0027 | -0.0062 | -0.058** |
| | (0.038) | (0.050) | (0.022) | (0.020) | (0.0036) | (0.0042) | (0.030) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | 0.000 | 0.530 | 0.158 | 0.136 | 0.009 | 0.013 | 0.138 |
| Observations | 3,784 | 3,784 | 3,784 | 3,784 | 3,784 | 3,784 | 1,995 |

Notes: This table presents estimates of tool availability on child injuries and other hospitalization outcomes. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. Outcomes are measured beginning 60 days after initial randomization. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 1.5: Impact of Tool Availability on Investigation

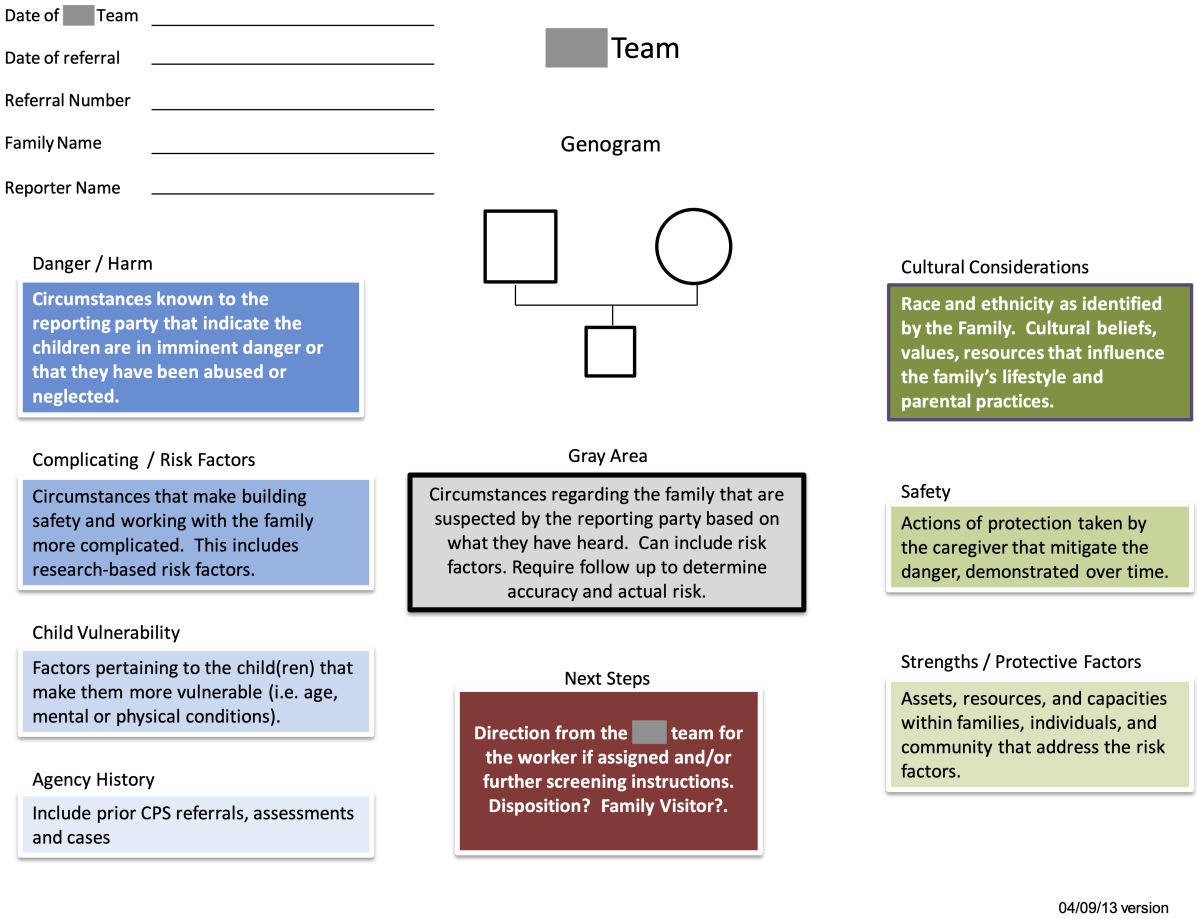
| | All | | Non-Active | | Active | |
|---|---|---|---|---|---|---|
| | (1) Screened in | (2) Screened in | (3) Screened in | (4) Screened in | (5) Screened in | (6) Screened in |
| Referral Score | 0.003** | 0.006*** | 0.006*** | 0.011*** | -0.002 | -0.010 |
| | (0.001) | (0.002) | (0.002) | (0.002) | (0.004) | (0.006) |
| Tool Available (T) | 0.004 | 0.049 | 0.005 | 0.077** | -0.007 | -0.240* |
| | (0.019) | (0.030) | (0.021) | (0.031) | (0.053) | (0.135) |
| Referral Score · T | | -0.005* | | -0.008** | | 0.015* |
| | | (0.003) | | (0.003) | | (0.008) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | 0.303 | 0.303 | 0.317 | 0.317 | 0.208 | 0.208 |
| Observations | 5,225 | 5,225 | 4,514 | 4,514 | 711 | 711 |

Notes: This table presents estimates of tool availability and tool interacted with score on a number of primary outcomes. Non-active referrals are referrals where no CPS intervention was ongoing at the time of referral. Randomization procedure controls are included. The intervention (T) is whether the tool was made available on the first referral in the trial, and the score is defined as the maximum child score for that referral. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

# Appendix

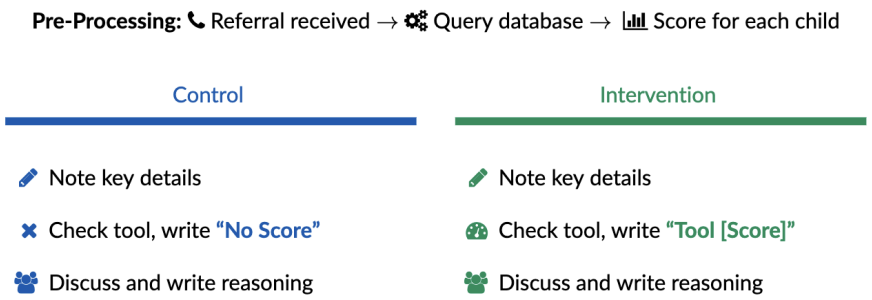
## 1.A Appendix Figures and Tables

Figure 1.A.1: Team Decision Document



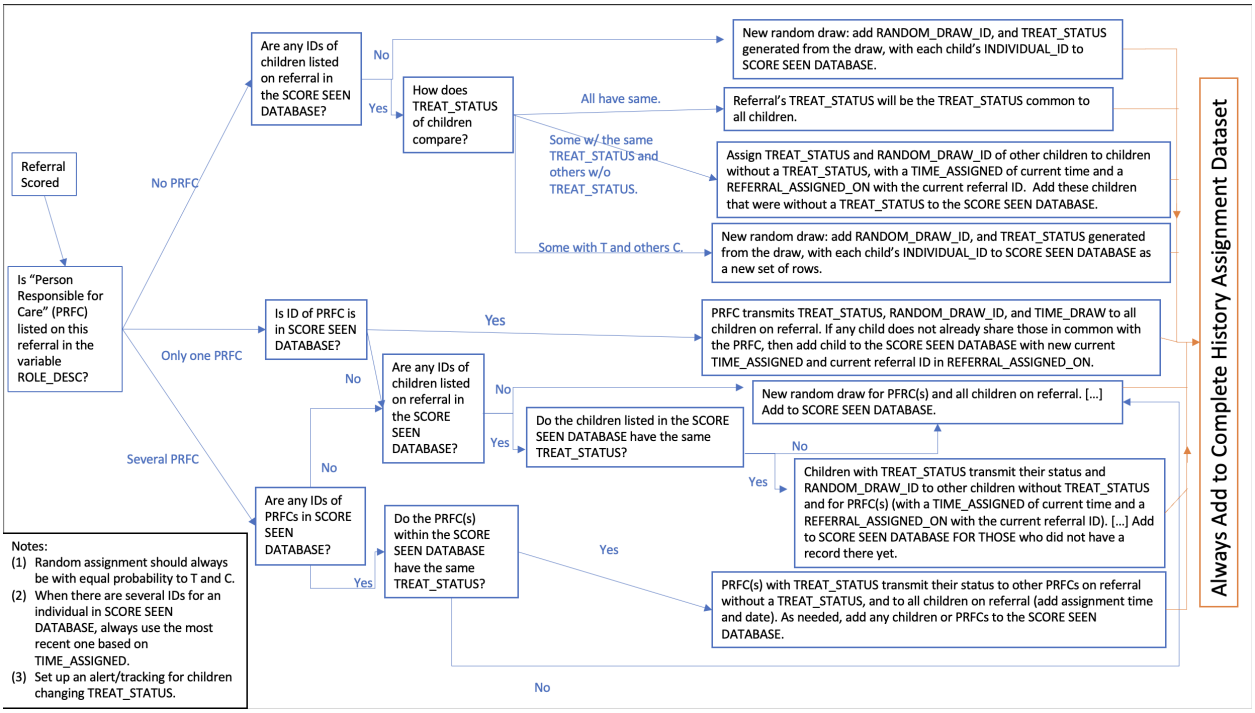

Notes: This figure is an example of the framework document that a team fills out during a team's decision-making process. The "next steps" field is filled out only after the tool is checked and a discussion is held.

Figure 1.A.2: Trial Description



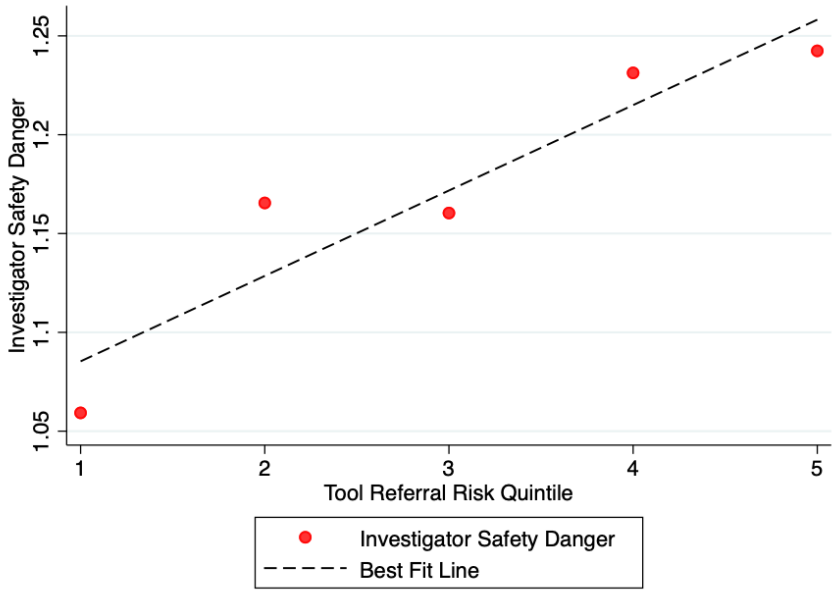

Notes: This figure shows the difference between the control and intervention groups during the team decision-making process.

Figure 1.A.3: Randomization Database Flowchart



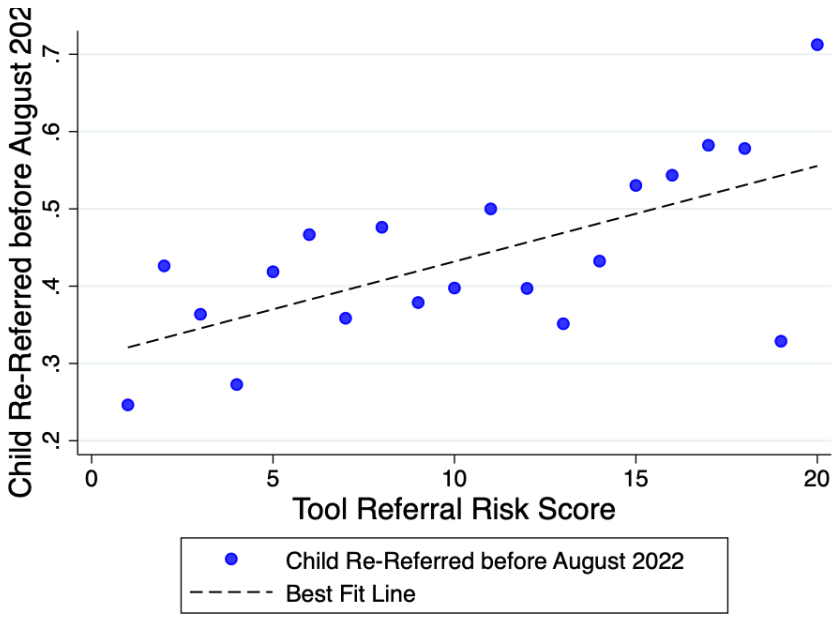

Notes: This figure is a graphical depiction of the intervention status design developed by the authors. The randomization status database was required to preserve the spirit of household-level randomization when some referrals were missing biological mothers or had previously-randomized members.

Figure 1.A.4: Tool-Predicted Risk and Investigator-Assessed Risk



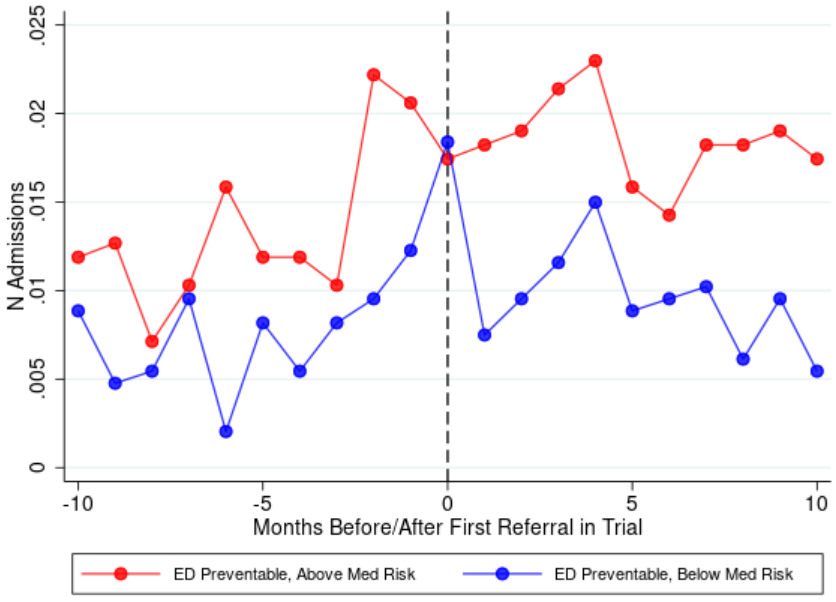

Notes: This figure presents a scatter plot of the relationship between an investigator's safety assessment evaluation of a child's current safety and the decision tool's predicted risk of future foster care. The authors have assigned numerical values to the safety assessment as follows: 1: "No current or impending danger to the child" (90 percent of observations), 2: "Current/impending danger to the child and caregiver is protective: child is safe" (four percent of observations), and 3: "Current/impending danger to the child and caregiver is not sufficiently protective" (six percent of observations). Tool risk is presented in quintiles for visual clarity. The sample is restricted to the control group (score not seen). Worker assessments are only performed for screened-in children.

Figure 1.A.5: Risk Score and Re-Referrals



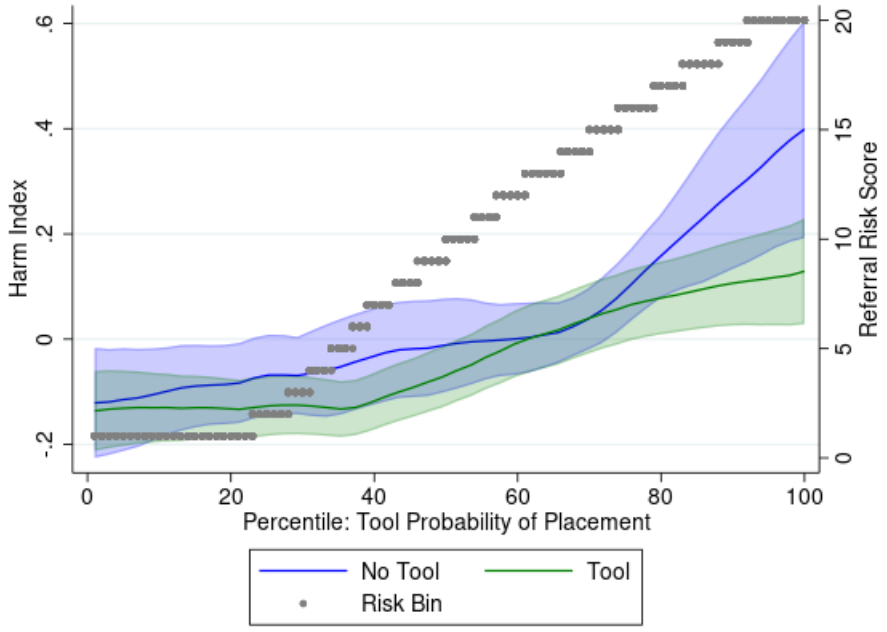

Notes: This figure presents a scatter plot of re-referral rates from initial randomization before August 2022, by algorithmic tool risk bin. The sample consists of all observations in the control group (no score revealed) that were screened out.

Figure 1.A.6: Risk Score and Preventable ER Visits over Time



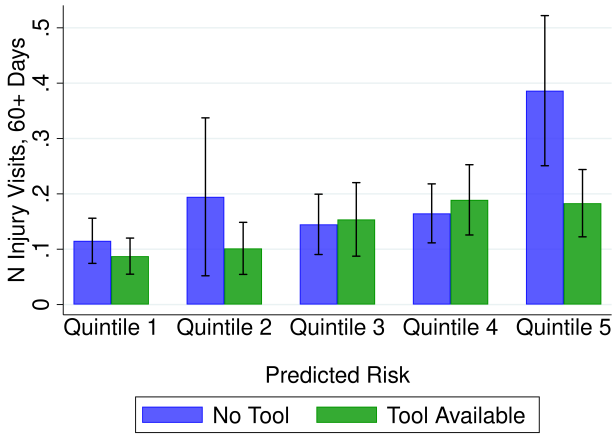

Notes: This figure shows the relationship between predicted risk and preventable emergency room visits relative to the timing of a child's first CPS appearance during the trial. Each cell of the plot presents the mean number of hospital admissions with at least one ambulatory care sensitive condition (ACSC) ICD-10 code in a given month for children above median predicted removal risk at the time of randomization (red) and below median risk (blue), respectively.

Figure 1.A.7: Harm Index by Predicted Risk Level and Intervention Status



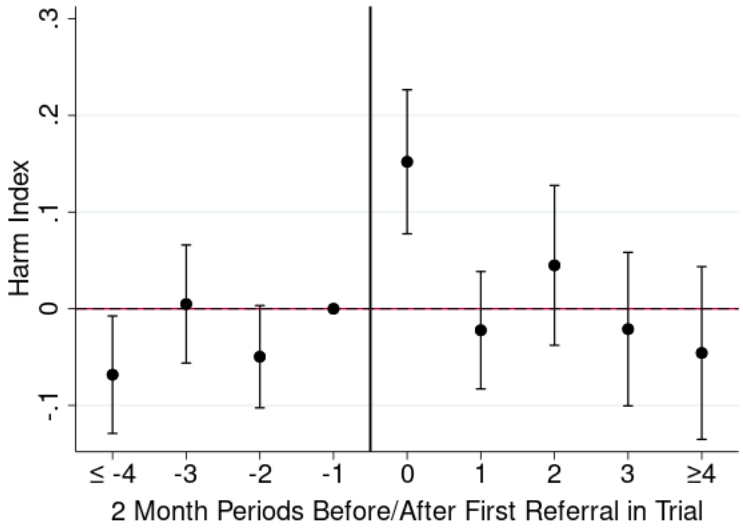

Notes: This figure presents local mean regressions of an aggregated index of potential maltreatment-related hospitalizations on predicted referral risk score for the intervention and control groups, respectively. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.A.8: Injury Hospitalizations by Predicted Risk Level and Intervention Status



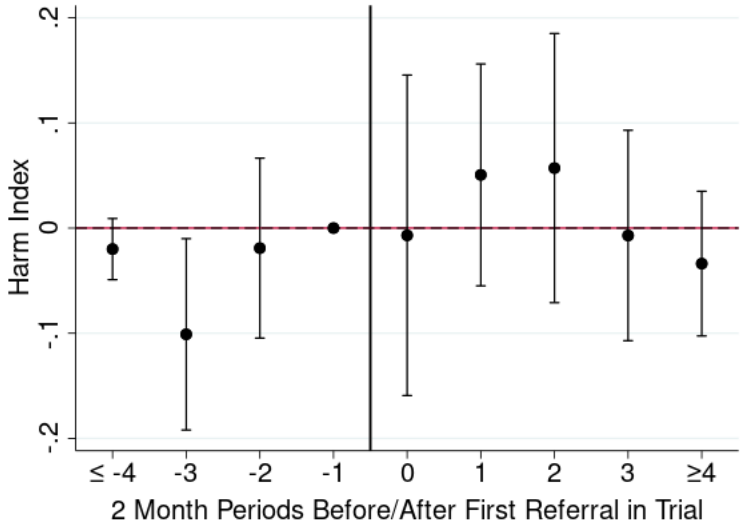

Notes: This figure presents mean rates of injury-related hospitalizations 60+ days after randomization by risk of removal quintile and tool availability cell. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization.

Figure 1.A.9: Dynamic Harm Index Relative to First Appearance in Trial



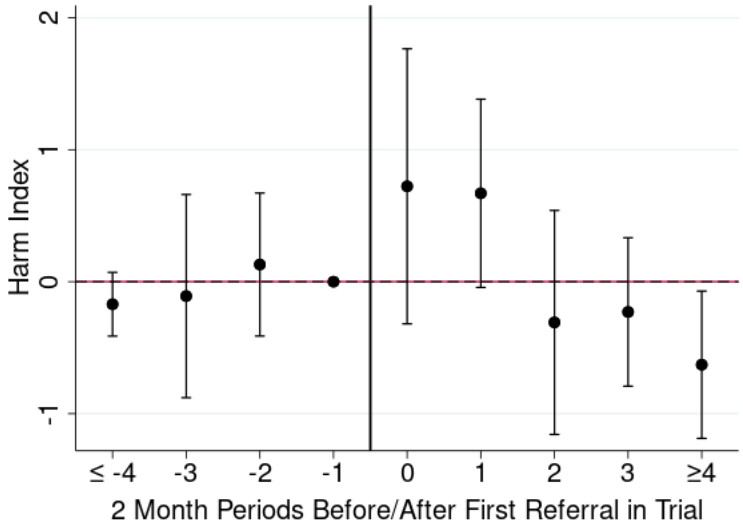

Notes: This figure presents event study estimates of child health outcomes (dynamic harm index) for children relative to their first referral in the trial at period 0. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

Figure 1.A.10: Association of Screen-In with Dynamic Harm Index



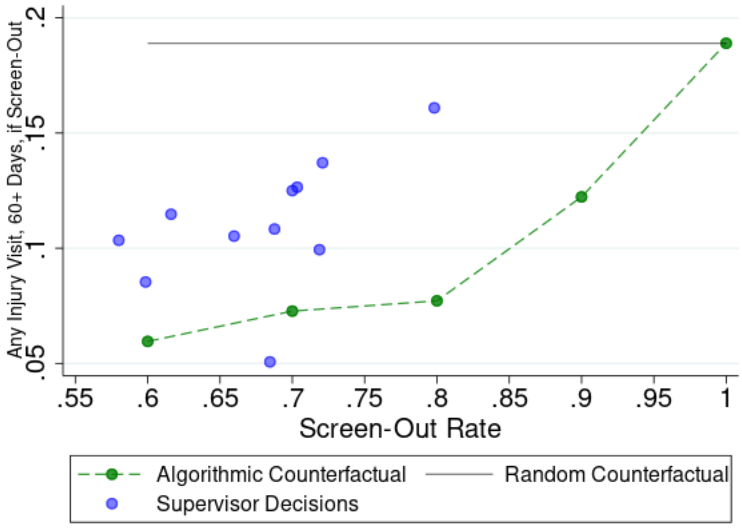

Notes: This figure presents event study estimates of child health outcomes (dynamic harm index) for children screened in compared to children screened out, relative to their first referral in the trial at period 0. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

Figure 1.A.11: Association of Supervisor Screen-In Propensity with Dynamic Harm Index



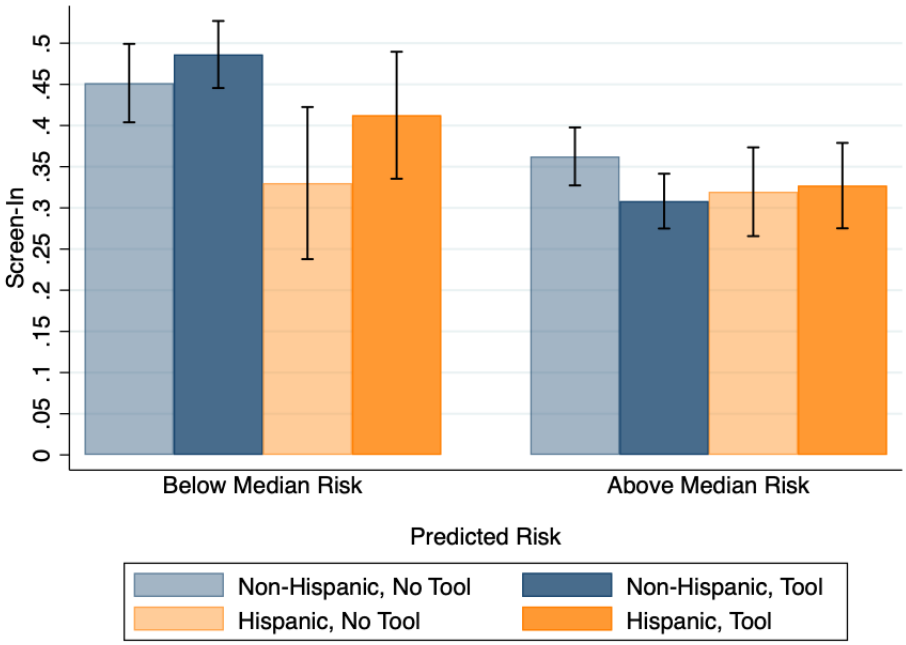

Notes: This figure presents event study estimates of child health outcomes (dynamic harm index) for children scaled by supervisor propensity to screen in. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.

Figure 1.A.12: Counterfactual Estimation: Any Injury Conditional on Screen-Out



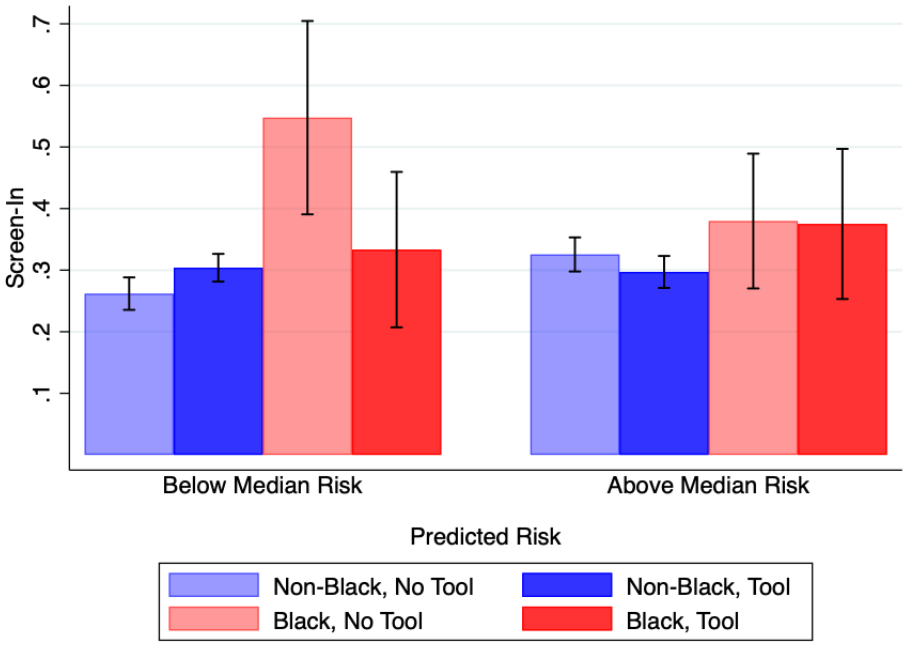

Notes: This figure presents a plot of injury rates after 60 days and screen-out rates across the 11 main supervisors (in blue). The random counterfactual plot (in black) is the mean screen-out injury rate for the entire sample of children, estimated using linear extrapolation following an approach by Angelova, Dobbie and Yang (2022). The algorithmic counterfactual spline (in green) is the estimated injury rate conditional on screen-out if the algorithm screened out a given share of children with the highest algorithmic risk scores. The sample consists of all first-time observations of children on non-active referrals in the first half of the trial among the three primary decisions teams and where a supervisor had at least 100 child observations in the remaining sample.

Figure 1.A.13: Impact of Tool on Investigations by Ethnicity



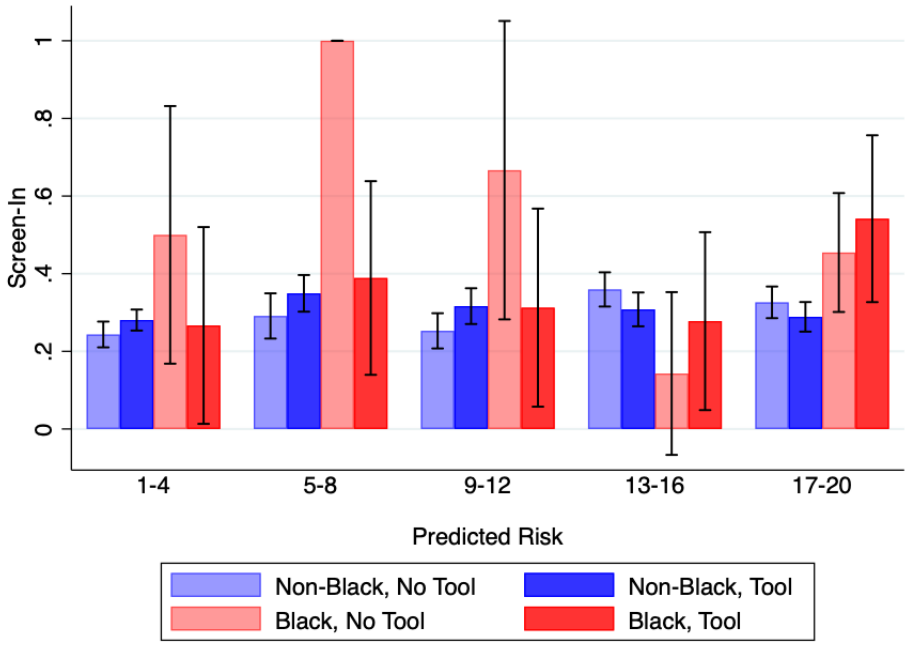

Notes: This bar plot presents mean screen-in rates by ethnicity (Hispanic vs. other) and intervention status, for children below and above median predicted risk, respectively.

Figure 1.A.14: Tool Availability and Screen-In Disparities




Notes: This bar plot presents mean screen-in rates by race (Black vs. other) and intervention status, for children below and above median predicted risk, respectively.

Figure 1.A.15: Tool Availability and Screen-In Disparities, by Quartile

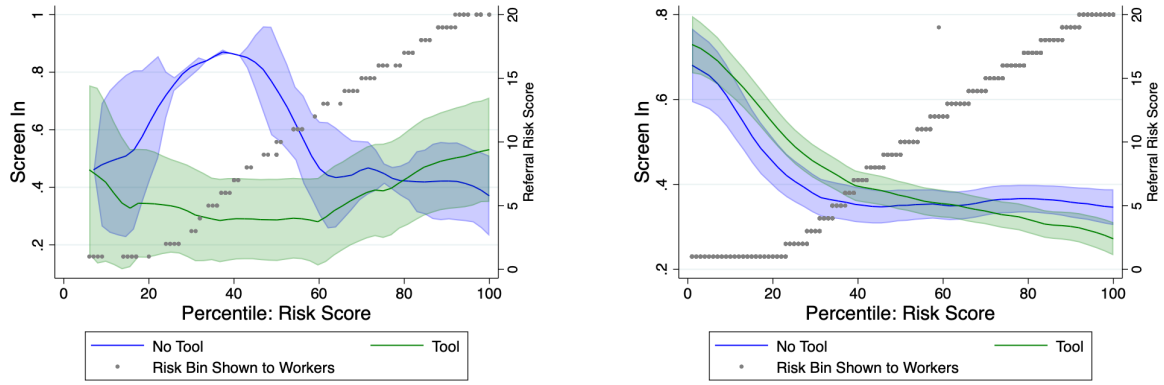



Notes: This bar plot presents mean screen-in rates by race (Black vs. other) and intervention status, for children at each risk quintile, respectively.

Figure 1.A.16: Screen-In Heterogeneity by Race

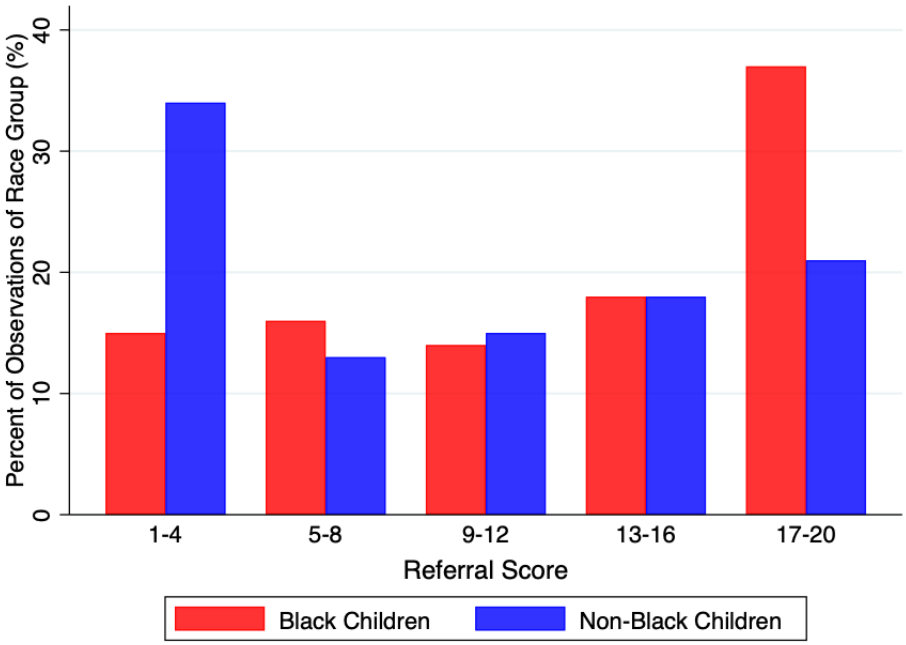
(1) Black

(2) Not Black



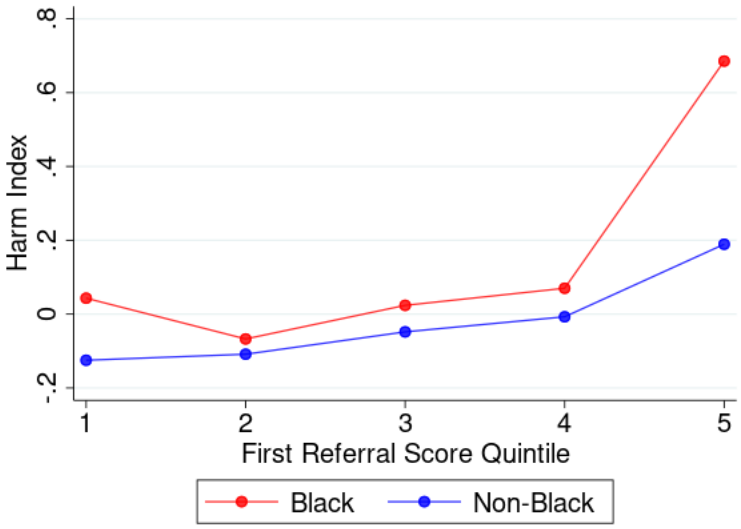

Notes: This figure presents heterogeneity in the relationship between screen-in and risk score by race (Black and non-Black). Each subfigure presents local mean regressions of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.A.17: Distribution of Predicted Risk by Race



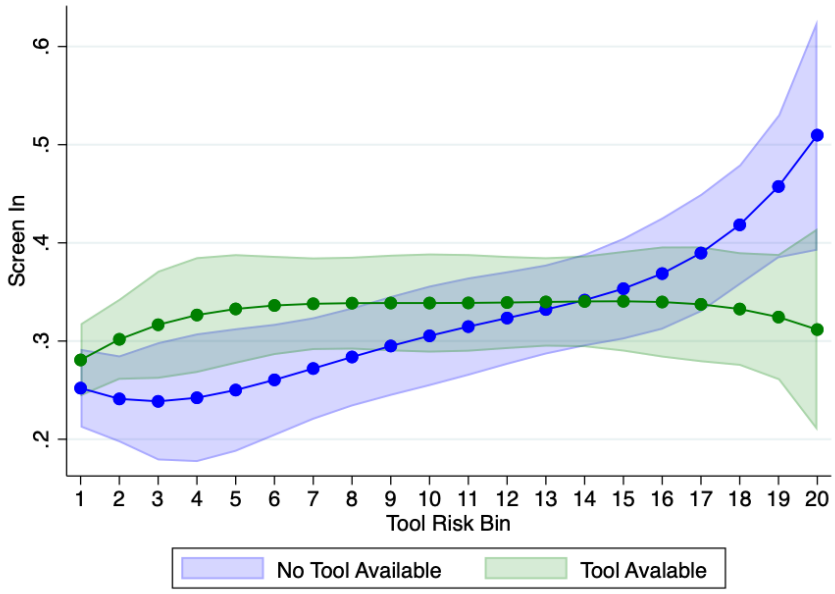

Notes: This bar plot presents the share of within-race observations across five risk quintiles. For example, 34 percent of non-Black children have a referral risk score between 1 and 4.

Figure 1.A.18: Score Predictive Power by Race




Notes: This connected plot presents mean harm index rates by race (Black vs. other) for children at each risk quintile.

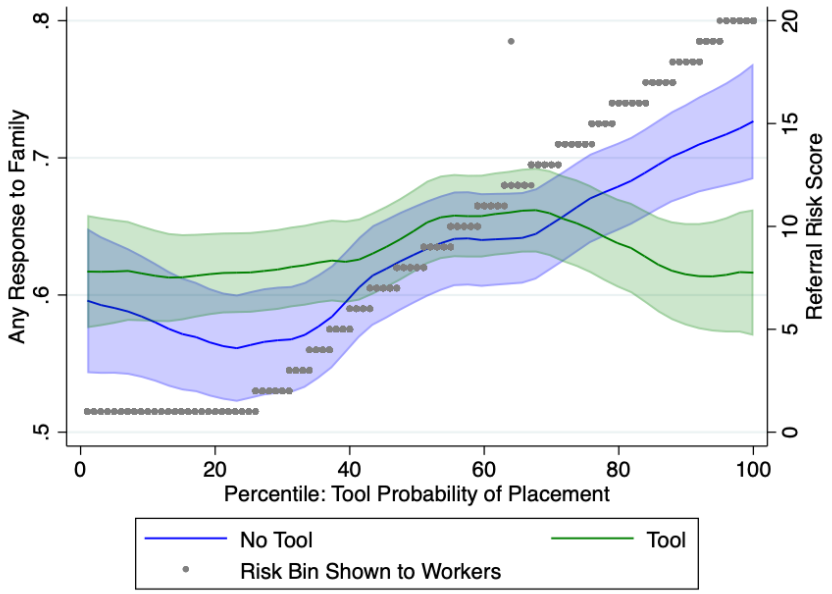
Figure 1.A.19: Tool Availability and Sensitivity to Risk – Alternative (Margins) Plot



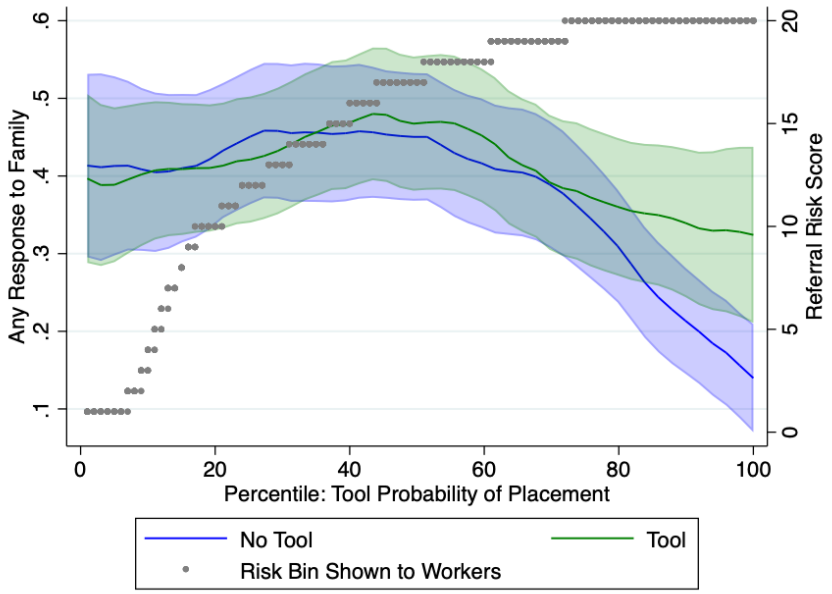

Notes: This figure presents margins plot coefficient estimates of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, respectively. Estimates are generated from margins of a quadratic polynomial of screen-in on score, including controls for randomization procedure. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile.

Figure 1.A.20: Impact of Tool on Any Response to Family

(1) Non-Active (86% of Sample)

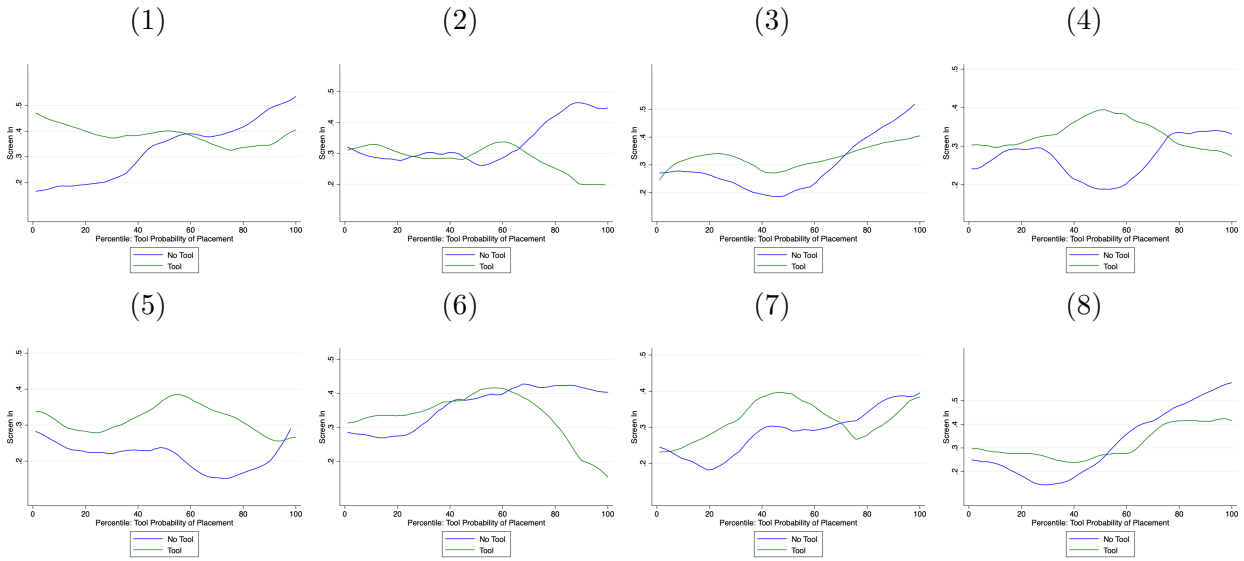



(2) Active (14% of Sample)



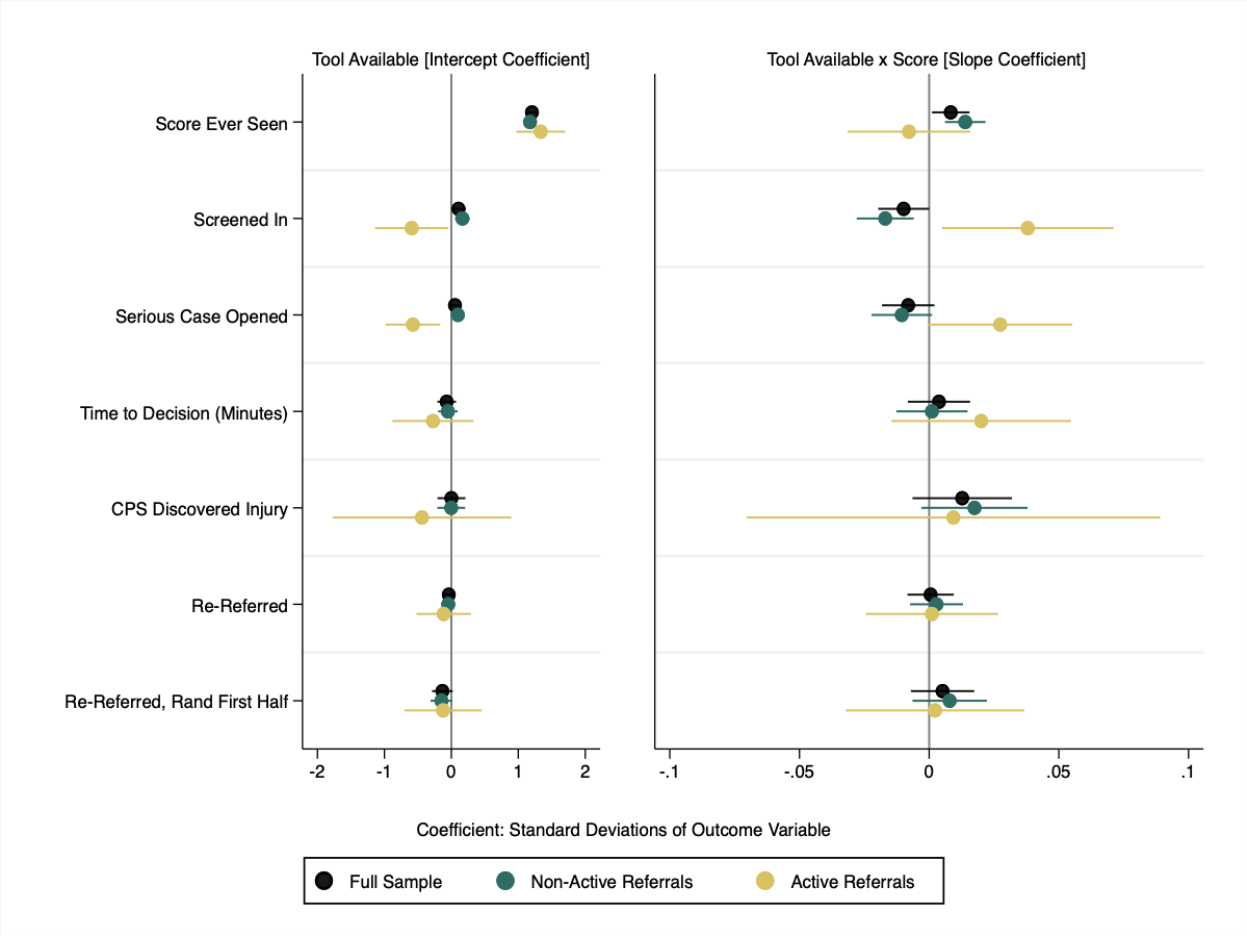

Notes: This figure presents local mean regressions of decision to respond (screen in or screen out with resources, informal visitor, or law enforcement) on predicted referral risk score for the intervention and control groups, for active and non-active referrals, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. 86 percent of child observations in the main sample are listed on non-active referrals. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.A.21: Tool Impact Across Supervisors




Notes: This figure presents heterogeneity in the relationship between screen-in and risk score by supervisor, for the eight most common supervisors during the sample period. Sample restricted to non-active referrals. Each subfigure presents local mean regressions of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

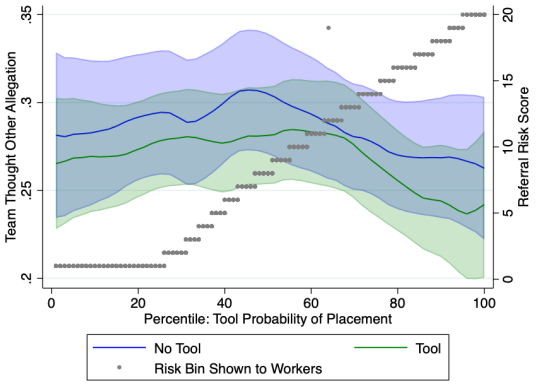
Figure 1.A.22: Impact of Tool Availability on CPS-Related Outcomes, by Risk Score



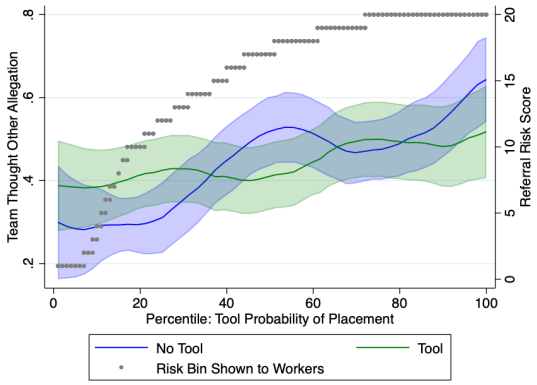

Notes: This figure presents point estimates and standard errors for the coefficient on tool availability and tool interacted with score when estimating equation 1.2 for a variety of outcomes. Estimates are presented with standardized outcome variables (mean 0, variance 1) for comparability of coefficient magnitudes across different outcomes. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. Standard errors are clustered at the household level and are shown at the 90 percent confidence level.

Figure 1.A.23: Impact of Tool on Screen-Out for Other Reasons

(1) Non-Active



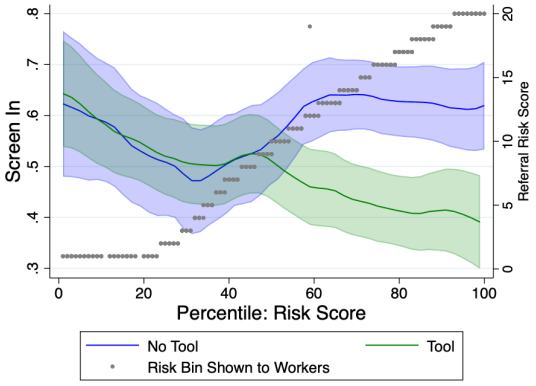

(2) Active



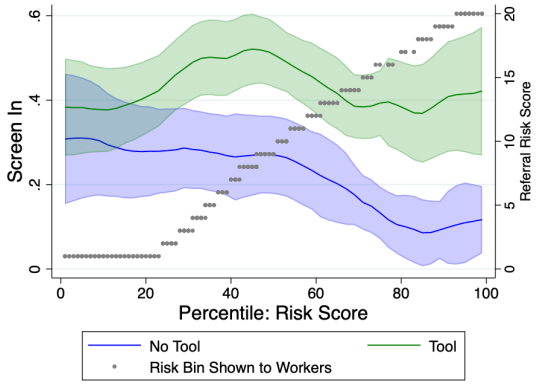

Notes: This figure presents local mean regressions of a child being screened out due to reasons unrelated to statutory definitions of maltreatment on predicted referral risk score for the intervention and control groups, respectively. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.A.24: Impact of Tool, by Select Allegation Types

(1) Alleged Parental Substance Abuse

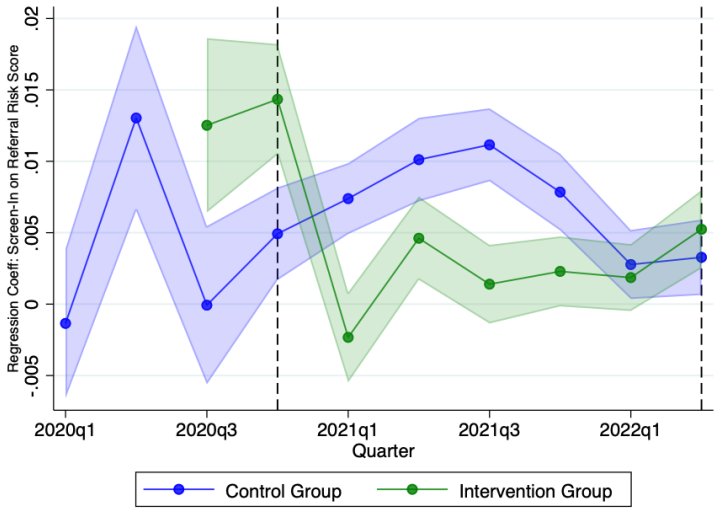



(2) Alleged Sexual Abuse



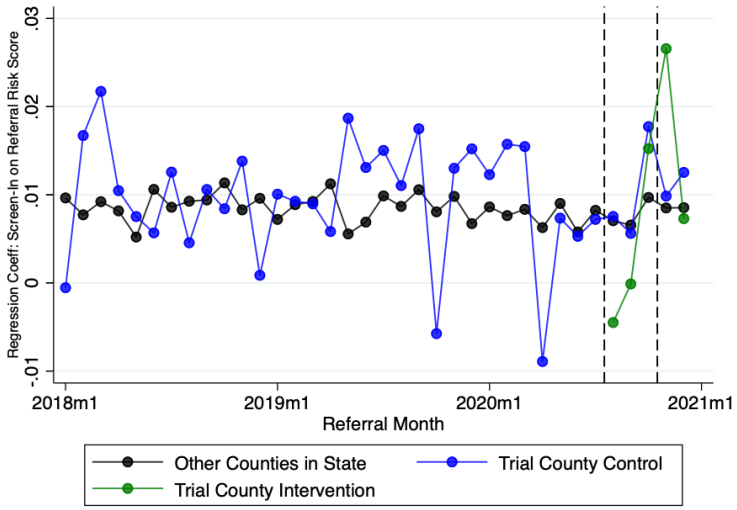

Notes: These figures present local mean regressions of decision to screen in (investigate) on predicted referral risk score for the intervention and control groups, respectively. Panel 1 uses a sample of children with parental substance abuse allegations. Panel 2 uses a sample of children who are alleged sexual abuse victims. Both samples include children on non-active and active referrals. Gray bins shown on the second vertical axis illustrate how the risk bin (1-20) shown to a team is associated with the risk probability percentile. Degree: 0. Kernel: Epanechnikov. Bandwidth: 10. Confidence level: 95%.

Figure 1.A.25: Screen-In Sensitivity to Predicted Risk in Intervention and Control Groups over Time



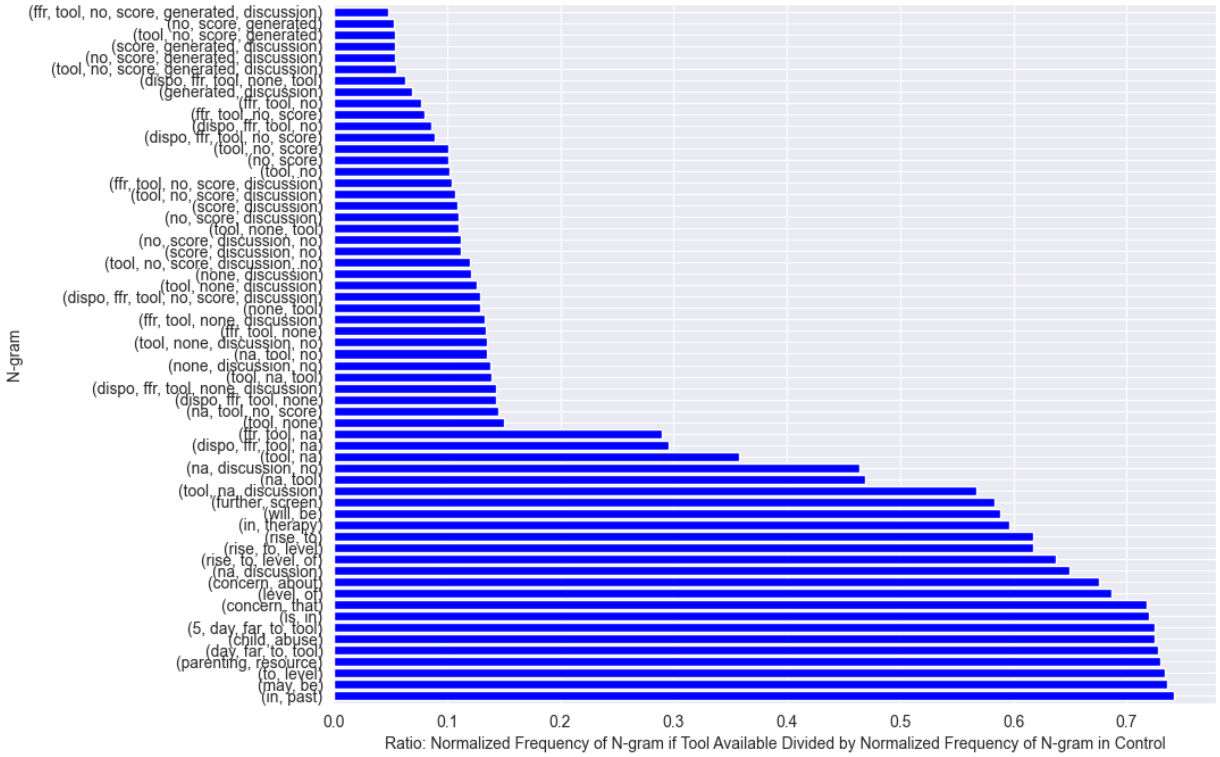

Notes: This figure plots within-cell regression coefficient estimates of screen-in on predicted referral risk, for each quarter by intervention-status cell (tool available vs. tool not available). This figure helps to visualize general equilibrium effects of tool use on county agency decisions over time, in comparison with other counties. Sample restricted to non-active referrals, where no CPS intervention was ongoing at the time of referral. Dashed vertical lines indicate the beginning of the trial (November 2020) and the end of the trial (March 30, 2022).

Figure 1.A.26: Screen-In Sensitivity to Predicted Risk over Time, Statewide



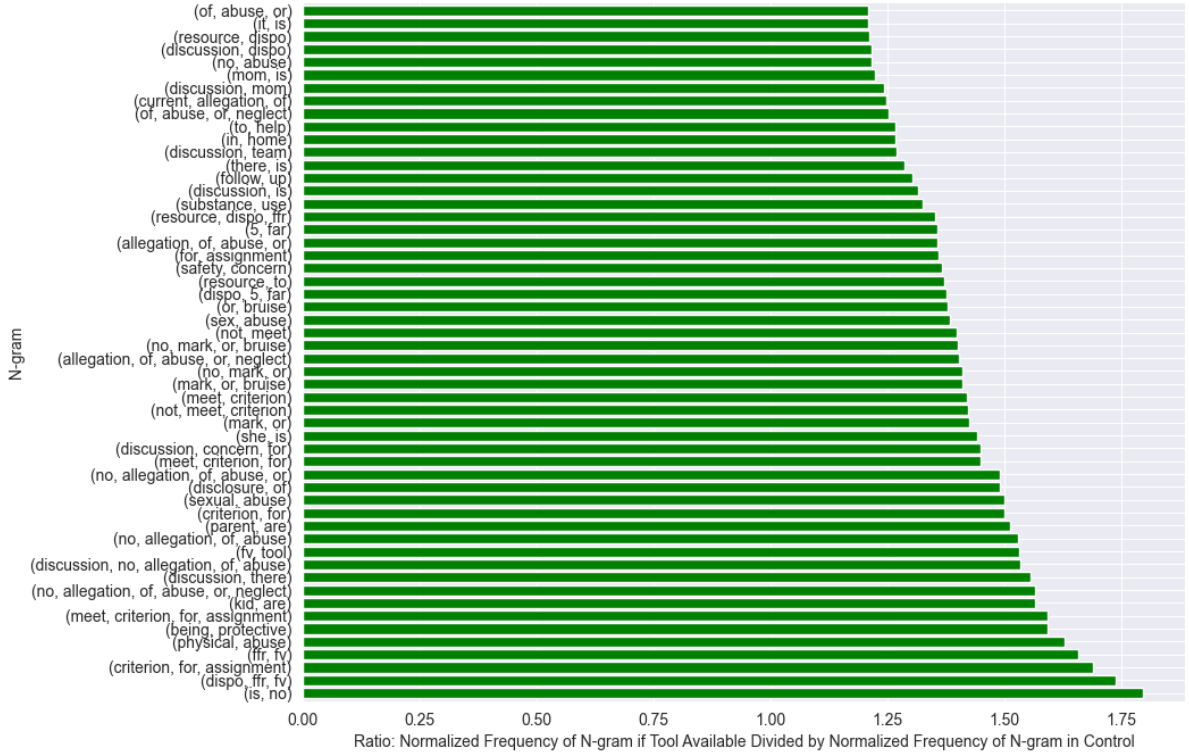

Notes: This figure plots within-cell regression coefficient estimates of screen-in on predicted referral risk, for each county by month by intervention-status cell (trial county vs. other counties in the state; tool available vs. tool not available). This figure helps to visualize general equilibrium effects of tool use on county agency decisions over time, in comparison with other counties. Dashed vertical lines indicate the beginning of Phase 1 (August 2020) and the beginning of the main trial (November 2022).

Figure 1.A.27: Most Distinctive Discussion Phrases, Control Group




Notes: This figure shows the most distinctive $n = \{2, 3, 4, 5, 6\}$ n-grams (consecutive word sequences) found in discussion text for the control group relative to the intervention (tool available) group. All n-grams found at 33 percent or greater relative frequencies in the control group are included, normalizing for discussion length.

Figure 1.A.28: Most Distinctive Discussion Phrases, Intervention Group




Notes: This figure shows the most distinctive $n = \{2, 3, 4, 5, 6\}$ n-grams (consecutive word sequences) found in discussion text for the intervention (tool available) group relative to the control group. All n-grams found at 20 percent or greater relative frequencies in the control group are included, normalizing for discussion length.

Table 1.A.1: Sample Restrictions for Trial Analysis Sample, November 2020 - March 2022

| **Restrictions** | **Observations** | **Unique Clients** | **Referrals** |
|---|---|---|---|
| 1. In Trial Period Listed on Referral | 15,023 | 7,723 | 8,309 |
| 2. Scored | 13,115 | 7,021 | 7,345 |
| 3. First Scored Referral in Trial | 6,852 | 6,852 | 4,213 |
| 4. Excluding those Seen in Testing Phase (Phase 1) of Trial | 6,601 | 6,601 | 4,073 |
| 5. Excluding those with Household Member already in Trial | 5,998 | 5,998 | 3,631 |
| 6. First Call going to Discussion Team | 5,781 | 5,781 | 3,451 |
| 7. Omit Multiple Scored Referrals before first Discussion Team Meeting | 5,503 | 5,503 | 3,299 |
| 8. Final Sample, Excluding Observations with Overwritten Status | 5,225 | 5,225 | 3,180 |

Notes: This table presents the sample restrictions used to arrive at the main analysis sample.

Table 1.A.2: Test of Randomization of Intervention Status

| | (1)<br>Tool Available |
|---|---|
| White | -0.02 |
| | [0.07] |
| Black | -0.06 |
| | [0.06] |
| Missing Race | -0.06 |
| | [0.07] |
| Hispanic | 0.03 |
| | [0.03] |
| Missing Ethnicity | 0.01 |
| | [0.04] |
| Female | -0.00 |
| | [0.01] |
| Algorithm Score (out of 20) | -0.00 |
| | [0.00] |
| Mandated Reporter | -0.02 |
| | [0.02] |
| Child in Danger | 0.07 |
| | [0.09] |
| Alleged Injurious Environment | -0.00 |
| | [0.02] |
| Alleged Physical Abuse | 0.06** |
| | [0.03] |
| Alleged Substance Allegation - Parent | 0.01 |
| | [0.03] |
| Team 3 | 0.04 |
| | [0.03] |
| Team 2 | 0.00 |
| | [0.03] |
| Auxiliary Team | 0.05 |
| | [0.03] |
| Number of Persons Listed on Referral in CSDA data | -0.00 |
| | [0.01] |
| Number of Alleged Victims on Referral | 0.00 |
| | [0.02] |
| Active | 0.01 |
| | [0.03] |
| Number of referrals of client since we have data | -0.02 |
| | [0.01] |
| Randomization Procedure Controls | × |
| N Randomization Clusters | 3,042 |
| F-statistic | 0.973 |
| p-value of F-test | 0.491 |

Notes: This table presents a balance test supporting the conditional random assignment of intervention status to children in the sample. The $p$-value of the joint F-test is presented at the bottom, showing the joint significance of the listed observable characteristics on being assigned to intervention status, conditioning on number of children for a subset of referrals with no mother and received after March 2021. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.3: Frequency of Score Seen when Tool Available

| | (1) Control Means | (2) Tool Available (T) |
|---|---|---|
| Reported Seeing Score | 0.004 | 0.640*** |
| | | (0.013) |
| Reported Not Seeing a Score | 0.769 | -0.649*** |
| | | (0.016) |
| Score Reported Seen on Initial Referral or Future Referral | 0.107 | 0.640*** |
| | | (0.015) |
| N Observations | 2,300 | 5,225 |
| N Clusters | | 3,156 |

Notes: This table presents the first stage relationship between being assigned to intervention (tool available) and the worker team reportedly seeing a score from the tool. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.4: Impact of Tool on Child Health, Non-Active Referrals

| | (1) Harm Index | (2) N High Priority | (3) N Injury Visits | (4) N Avoidable ED | (5) Any Maltreat ICD | (6) Inten Injury | (7) N Injury, Never Screened In |
|---|---|---|---|---|---|---|---|
| Tool Available | -0.085** | -0.087 | -0.040* | -0.030 | -0.0049 | -0.0047 | -0.047 |
| | (0.041) | (0.053) | (0.024) | (0.022) | (0.0041) | (0.0042) | (0.032) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | -0.020 | 0.497 | 0.146 | 0.131 | 0.009 | 0.011 | 0.124 |
| Observations | 3,188 | 3,188 | 3,188 | 3,188 | 3,188 | 3,188 | 1,667 |

Notes: This table presents estimates of tool availability on child injuries and other hospitalization outcomes. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. The sample consists of non-active referrals, where no CPS intervention was ongoing at the time of referral. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.5: Impact of Tool on Child Health, Active Referrals

| | (1) Harm Index | (2) N High Priority | (3) N Injury Visits | (4) N Avoidable ED | (5) Any Maltreat ICD | (6) Inten Injury | (7) N Injury, Never Screened In |
|---|---|---|---|---|---|---|---|
| Tool Available | -0.053 | -0.056 | -0.097* | 0.030 | 0.0081 | -0.013 | -0.11 |
| | (0.10) | (0.14) | (0.058) | (0.049) | (0.0073) | (0.014) | (0.080) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | 0.107 | 0.710 | 0.220 | 0.163 | 0.007 | 0.023 | 0.213 |
| Observations | 596 | 596 | 596 | 596 | 596 | 596 | 328 |

Notes: This table presents estimates of tool availability on child injuries and other hospitalization outcomes. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. The sample consists of active referrals, where a CPS intervention was already taking place with the household at the time of a referral. Outcomes are measured beginning 60 days after initial randomization. The sample is restricted to children first randomized before November 2021, to allow for at least six months of potential hospitalization after the 60 days following randomization. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.6: Impact of Tool on Child Health, Short Run

| | (1) N High Priority | (2) N Injury Visits | (3) N Avoidable ED | (4) Any Maltreat ICD | (5) Inten Injury | (6) N Injury, Never Screened In |
|---|---|---|---|---|---|---|
| Tool Available | -0.0043 | -0.0064 | 0.0018 | 0.0026 | -0.0035* | -0.0036 |
| | (0.015) | (0.0067) | (0.0069) | (0.0023) | (0.0021) | (0.0082) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | 0.085 | 0.022 | 0.019 | 0.003 | 0.003 | 0.017 |
| Observations | 3,188 | 3,188 | 3,188 | 3,188 | 3,188 | 1,667 |

Notes: This table presents estimates of tool availability on child injuries and other hospitalization outcomes in the 60 days following randomization. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better. The sample is restricted to children first randomized before November 2021. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.7: Tool and Child Health – Placebo Outcomes Prior to Randomization

| | (1) Harm Index | (2) N High Priority | (3) N Injury Visits | (4) N Avoidable ED | (5) Any Maltreat ICD | (6) Inten Injury |
|---|---|---|---|---|---|---|
| Tool Available | -0.019 | -0.018 | 0.014 | -0.0031 | 0.00015 | -0.0079* |
| | (0.036) | (0.042) | (0.018) | (0.017) | (0.0038) | (0.0042) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | -0.000 | 0.518 | 0.166 | 0.124 | 0.010 | 0.015 |
| Observations | 3,784 | 3,784 | 3,784 | 3,784 | 3,784 | 3,784 |

Notes: This table presents estimates of tool availability on child injuries and other hospitalization outcomes. The sample is restricted to children first randomized before November 2021, and looks at hospitalization rates *prior* to randomization. The harm index is constructed as a sum of standardized outcomes including number of high-priority admissions, number of admissions with listed injury, number of avoidable ER visits, any admission with a confirmed maltreatment code, and any visit with an intentional injury (assault, self-harm) code. The index is standardized (mean 0, variance 1) across the entire sample and a lower value is considered better.Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.8: Supervisor Propensity Instrument Details

| | Balance Test | First Stage Instrument Strength | OLS | IV | | Tool Impact |
|---|---|---|---|---|---|---|
| | Sup Propensity: Screen-In | Screened In | Harm Index | Harm Index | Harm Index | Harm Index |
| Sup Propensity: Screen-In | | 0.728** | | | | -1.967** |
| | | (0.282) | | | | (0.934) |
| | | [0.06] | | | | [0.08] |
| Screened In | | | 0.0183 | -1.578*** | -1.942*** | |
| | | | (0.0545) | (0.601) | (0.673) | |
| | | | [0.76] | [0.05] | [0.05] | |
| Tool Available (T) | | | | | | -0.417 |
| | | | | | | (0.416) |
| | | | | | | [0.38] |
| Tool Available (T)=1 × Sup Propensity: Screen-In | | | | | | 1.025 |
| | | | | | | (1.232) |
| | | | | | | [0.42] |
| Referral and Demographic Controls | × | × | × | | × | × |
| Discussion Team FEs | × | × | × | × | × | × |
| F-statistic | 6.05 | 6.64 | | | | |
| p-value of F-test (bootstrap) | 0.26 | 0.06 | | | | |
| Mean of Outcome Variable | .314 | .322 | -7.96e-09 | -7.96e-09 | -7.96e-09 | -7.96e-09 |
| Observations | 2,246 | 2,246 | 2,246 | 2,246 | 2,246 | 2,246 |

Notes: Column 1 shows that leave-self-out supervisor propensity to screen in is uncorrelated with a number of salient characteristics that could predict screen-in (F-test p-value ¿ 0.10), conditioning on discussion team. Characteristics include: predicted referral risk score, indicator for mandated reporter, indicators for the three most common allegation types (injurious environment, physical abuse, alleged parental substance), child gender, child race, and number of children listed on the referral. Column 2 shows that decisions for a given child-referral are associated with supervisor decisions for other referrals. Columns 3 presents OLS estimates an association of screen-in with a standardized post-60-day hospitalization harm index, conditioning on discussion team and characteristics of the referral and child. Columns 4-5 present IV estimates using supervisor propensity measures with and without controls. Column 6 presents heterogeneity by intervention status. The sample consists of all first-time observations of children on non-active referrals in the first half of the trial among the three primary decisions teams and where a supervisor had at least 100 child observations in the remaining sample. Standard errors are clustered by supervisor. Due to a small number of supervisor clusters (11) in the remaining sample, wild-bootstrapped standard error p-values are reported in brackets. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 1.A.9: Effect of Tool on Racial Disparities in Screen-Ins, CPS-Known Race

| | Black=0 if Missing Race | | | | Excluding Missing Race | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) Screened in | (2) Screened in | (3) Screened in | (4) Screened in | (5) Screened in | (6) Screened in | (7) Screened in | (8) Screened in |
| Referral Risk Score | 0.0059*** | 0.0023 | 0.0028* | -0.00066 | -0.0059** | -0.0070** | -0.010*** | -0.011*** |
| | (0.0020) | (0.0020) | (0.0014) | (0.0014) | (0.0029) | (0.0028) | (0.0020) | (0.0019) |
| Black | 0.53*** | 0.52*** | 0.20*** | 0.23*** | 0.28* | 0.29 | 0.14* | 0.18** |
| | (0.16) | (0.17) | (0.076) | (0.074) | (0.17) | (0.19) | (0.073) | (0.072) |
| Score · Black | -0.025** | -0.022** | | | -0.011 | -0.0087 | | |
| | (0.011) | (0.011) | | | (0.011) | (0.012) | | |
| Treatment Status: Score Available | 0.063** | 0.053* | 0.0098 | 0.0030 | 0.10** | 0.092* | -0.0062 | -0.013 |
| | (0.030) | (0.028) | (0.019) | (0.018) | (0.051) | (0.047) | (0.024) | (0.023) |
| Referral Score · T | -0.0054* | -0.0051** | | | -0.0089** | -0.0087** | | |
| | (0.0028) | (0.0026) | | | (0.0039) | (0.0036) | | |
| T · Black | -0.59*** | -0.54** | -0.13 | -0.14 | -0.64*** | -0.58** | -0.16 | -0.17* |
| | (0.22) | (0.23) | (0.11) | (0.10) | (0.23) | (0.24) | (0.11) | (0.10) |
| Score · T · Black | 0.036** | 0.032** | | | 0.041** | 0.035** | | |
| | (0.016) | (0.015) | | | (0.016) | (0.016) | | |
| Detailed Controls | | × | | × | | × | | × |
| Observations | 5225 | 5205 | 5225 | 5205 | 3614 | 3613 | 3614 | 3613 |

Notes: This table presents estimates of tool availability on screen in for Black children relative to non-Black children. Race is defined only using race indicators recorded by Child Protective Services. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. Standard errors are clustered at the household level. A permutation test of the race variable Black using specification (1) suggests that the likelihood of finding such results with this sample is less than one in 1,000. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.10: Effect of Tool on Racial Disparities in Screen-Ins, CPS-Known Race and Hospital-Known Race

| | Black=0 if Missing Race | | | | Excluding Missing Race | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) Screened in | (2) Screened in | (3) Screened in | (4) Screened in | (5) Screened in | (6) Screened in | (7) Screened in | (8) Screened in |
| Referral Risk Score | 0.0060*** | 0.0024 | 0.0028** | -0.00057 | -0.0058** | -0.0069** | -0.010*** | -0.011*** |
| | (0.0021) | (0.0020) | (0.0014) | (0.0014) | (0.0029) | (0.0028) | (0.0020) | (0.0019) |
| Black | 0.34** | 0.33** | 0.14** | 0.16** | 0.26 | 0.29 | 0.12* | 0.16** |
| | (0.14) | (0.14) | (0.065) | (0.065) | (0.16) | (0.18) | (0.068) | (0.068) |
| Score · Black | -0.017* | -0.014 | | | -0.010 | -0.0099 | | |
| | (0.0092) | (0.0092) | | | (0.011) | (0.011) | | |
| Treatment Status: Score Available | 0.062** | 0.051* | 0.0091 | 0.0021 | 0.10** | 0.095** | -0.0045 | -0.011 |
| | (0.030) | (0.028) | (0.019) | (0.018) | (0.052) | (0.047) | (0.025) | (0.023) |
| Referral Score · T | -0.0053* | -0.0049* | | | -0.0090** | -0.0089** | | |
| | (0.0028) | (0.0026) | | | (0.0039) | (0.0036) | | |
| T · Black | -0.33* | -0.27 | -0.084 | -0.087 | -0.60*** | -0.57** | -0.17* | -0.19* |
| | (0.18) | (0.18) | (0.091) | (0.089) | (0.21) | (0.22) | (0.098) | (0.095) |
| Score · T · Black | 0.021 | 0.016 | | | 0.036** | 0.032** | | |
| | (0.013) | (0.013) | | | (0.015) | (0.015) | | |
| Detailed Controls | | × | | × | | × | | × |
| Observations | 5225 | 5205 | 5225 | 5205 | 3614 | 3613 | 3614 | 3613 |

Notes: This table presents estimates of tool availability on screen in for Black children relative to non-Black children. Race is defined using race indicators known either to Child Protective Services or a hospital. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. Standard errors are clustered at the household level. A permutation test of the race variable Black using specification (1) suggests that the likelihood of finding such results with this sample is less than one in 1000. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.11: Effect of Tool on Ethnic Disparities in Screen-Ins

| | Hispanic=0 if Missing Ethnicity | | | | Excluding Missing Ethnicity | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) Screened in | (2) Screened in | (3) Screened in | (4) Screened in | (5) Screened in | (6) Screened in | (7) Screened in | (8) Screened in |
| Referral Risk Score | 0.0065*** | 0.0028 | 0.0027* | -0.00051 | -0.0079** | -0.0083*** | -0.0096*** | -0.011*** |
| | (0.0022) | (0.0021) | (0.0015) | (0.0014) | (0.0033) | (0.0032) | (0.0021) | (0.0020) |
| Hispanic | 0.11 | 0.068 | 0.018 | 0.015 | -0.15 | -0.14 | -0.053 | -0.045 |
| | (0.088) | (0.078) | (0.036) | (0.034) | (0.095) | (0.087) | (0.039) | (0.037) |
| Score · Hispanic | -0.0078 | -0.0047 | | | 0.0068 | 0.0069 | | |
| | (0.0063) | (0.0057) | | | (0.0067) | (0.0062) | | |
| Treatment Status: Score Available | 0.042 | 0.033 | -0.0035 | -0.0072 | 0.035 | 0.044 | -0.035 | -0.036 |
| | (0.031) | (0.029) | (0.020) | (0.019) | (0.057) | (0.053) | (0.028) | (0.026) |
| Referral Score · T | -0.0048* | -0.0042 | | | -0.0060 | -0.0068* | | |
| | (0.0029) | (0.0027) | | | (0.0043) | (0.0040) | | |
| T · Hispanic | 0.088 | 0.096 | 0.039 | 0.018 | 0.079 | 0.068 | 0.051 | 0.030 |
| | (0.12) | (0.11) | (0.050) | (0.048) | (0.13) | (0.12) | (0.054) | (0.051) |
| Score · T · Hispanic | -0.0029 | -0.0051 | | | -0.00091 | -0.0016 | | |
| | (0.0086) | (0.0079) | | | (0.0091) | (0.0084) | | |
| Detailed Controls | | × | | × | | × | | × |
| Observations | 5225 | 5205 | 5225 | 5205 | 3340 | 3339 | 3340 | 3339 |

Notes: This table presents estimates of tool availability on screen in for Hispanic children relative to non-Hispanic children. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.12: Level Effect of Tool on CPS-Related Outcomes, All Referrals

| | (1) Control Mean | (2) Tool Available T | (3) Frequency |
|---|---|---|---|
| *First Stage* | | | |
| Score Ever Seen | 0.11 | 0.64*** (0.015) | 2383 |
| *Team Decisions* | | | |
| Screened in | 0.30 | 0.0033 (0.019) | 1581 |
| Ever screened in | 0.45 | -0.0027 (0.020) | 2286 |
| Expedited Response Time | 0.029 | 0.0043 (0.0073) | 176 |
| Low-Intensity Investigation | 0.22 | -0.0067 (0.017) | 1145 |
| High-Intensity Investigation | 0.091 | -0.0083 (0.011) | 437 |
| Screen-Out: Community Response | 0.16 | -0.021 (0.015) | 797 |
| Screen-Out: Family Visitor | 0.093 | 0.0078 (0.012) | 524 |
| Screen-Out: Family Given Resources | 0.14 | 0.016 (0.015) | 832 |
| Screen-Out: Sent to Law Enforcement | 0.041 | -0.0047 (0.0076) | 195 |
| Any Response to Family | 0.59 | -0.00038 (0.020) | 3131 |
| Time to Decision (Minutes) | 11.8 | -0.28 (0.37) | 3964 |
| *Investigation* | | | |
| Investigator Risk Assessment Score | 4.56 | 0.17 (0.31) | 1124 |
| Risk Assessment: Services Recommended | 0.54 | -0.069* (0.041) | 671 |
| Days Until Investigation Closed | 49.9 | -2.20 (1.37) | 1579 |
| *Outcomes* | | | |
| CPS Discovered Injury | 0.062 | 0.037* (0.020) | 100 |
| Re-Referred, First Half of Trial | 0.53 | -0.045 (0.027) | 1237 |
| Removed, Randomized First Half | 0.033 | -0.0048 (0.0093) | 73 |
| Future CPS Index, Randomized First Half | 0.18 | -0.082 (0.063) | 422 |

Notes: This table presents estimates of tool availability on a number of primary outcomes. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.13: Level Effect of Tool on CPS-Related Outcomes, Non-Active Referrals

| | (1) | (2) | (3) |
|---|---|---|---|
| | | Tool Available | |
| | Control Mean | T | Frequency |
| *First Stage* | | | |
| Score Ever Seen | 0.090 | 0.65*** | 2027 |
| | | (0.015) | |
| *Team Decisions* | | | |
| Screened in | 0.32 | 0.0046 | 1433 |
| | | (0.021) | |
| Ever screened in | 0.45 | -0.00039 | 1988 |
| | | (0.022) | |
| Expedited Response Time | 0.030 | 0.0083 | 169 |
| | | (0.0083) | |
| Low-Intensity Investigation | 0.24 | -0.013 | 1069 |
| | | (0.019) | |
| High-Intensity Investigation | 0.081 | 0.000053 | 349 |
| | | (0.012) | |
| Screen-Out: Community Response | 0.17 | -0.024 | 720 |
| | | (0.017) | |
| Screen-Out: Family Visitor | 0.11 | 0.0035 | 502 |
| | | (0.013) | |
| Screen-Out: Family Given Resources | 0.16 | 0.0082 | 786 |
| | | (0.017) | |
| Screen-Out: Sent to Law Enforcement | 0.046 | -0.0074 | 181 |
| | | (0.0085) | |
| Any Response to Family | 0.63 | -0.0099 | 2854 |
| | | (0.021) | |
| Time to Decision (Minutes) | 11.8 | -0.38 | 3445 |
| | | (0.39) | |
| *Investigation* | | | |
| Investigator Risk Assessment Score | 4.48 | 0.0090 | 1012 |
| | | (0.32) | |
| Risk Assessment: Services Recommended | 0.53 | -0.070 | 601 |
| | | (0.043) | |
| Days Until Investigation Closed | 49.6 | -1.55 | 1431 |
| | | (1.42) | |
| *Outcomes* | | | |
| CPS Discovered Injury | 0.056 | 0.049** | 93 |
| | | (0.021) | |
| Re-Referred, First Half of Trial | 0.50 | -0.034 | 961 |
| | | (0.031) | |
| Removed, Randomized First Half | 0.025 | -0.0014 | 47 |
| | | (0.0094) | |
| Future CPS Index, Randomized First Half | 0.17 | -0.067 | 329 |
| | | (0.070) | |

Notes: This table presents estimates of tool availability on a number of primary outcomes. The sample consists of non-active referrals, where no CPS intervention was ongoing at the time of referral. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.14: Level Effect of Tool on CPS-Related Outcomes, Active Referrals

| | (1) Control Mean | (2) Tool Available T | (3) Frequency |
|---|---|---|---|
| *First Stage* | | | |
| Score Ever Seen | 0.20 | 0.61*** (0.044) | 356 |
| *Team Decisions* | | | |
| Screened in | 0.22 | -0.0060 (0.053) | 148 |
| Ever screened in | 0.44 | -0.012 (0.059) | 298 |
| Expedited Response Time | 0.020 | -0.020* (0.011) | 7 |
| Low-Intensity Investigation | 0.099 | 0.026 (0.041) | 76 |
| High-Intensity Investigation | 0.15 | -0.055 (0.037) | 88 |
| Screen-Out: Community Response | 0.11 | -0.0043 (0.037) | 77 |
| Screen-Out: Family Visitor | 0.014 | 0.028 (0.023) | 22 |
| Screen-Out: Family Given Resources | 0.037 | 0.054* (0.030) | 46 |
| Screen-Out: Sent to Law Enforcement | 0.014 | 0.0083 (0.014) | 14 |
| Any Response to Family | 0.37 | 0.045 (0.060) | 277 |
| Time to Decision (Minutes) | 11.5 | 0.34 (1.00) | 519 |
| *Investigation* | | | |
| Investigator Risk Assessment Score | 5.21 | 1.53 (1.06) | 112 |
| Risk Assessment: Services Recommended | 0.61 | -0.086 (0.15) | 70 |
| Days Until Investigation Closed | 52.3 | -6.28 (4.56) | 148 |
| *Outcomes* | | | |
| CPS Discovered Injury | 0.11 | -0.083 (0.055) | 7 |
| Re-Referred, First Half of Trial | 0.65 | -0.057 (0.062) | 276 |
| Removed, Randomized First Half | 0.064 | -0.010 (0.028) | 26 |
| Future CPS Index, Randomized First Half | 0.16 | -0.11 (0.14) | 93 |

Notes: This table presents estimates of tool availability on a number of primary outcomes. The sample consists of active referrals, where a CPS intervention was already in progress with the household at the time of a referral. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.15: Effect of Tool on CPS-Related Outcomes, by Score

| | (1) Mean if Score==1 | (2) Tool Available T | (3) Score | (4) Interaction T · Score | (5) Frequency |
|---|---|---|---|---|---|
| *First Stage* | | | | | |
| Score Ever Seen | 0.089 | 0.60*** | 0.0040*** | 0.0041* | 2383 |
| | | (0.024) | (0.0014) | (0.0022) | |
| *Team Decisions* | | | | | |
| Screened in | 0.28 | 0.049 | 0.0055*** | -0.0045* | 1581 |
| | | (0.030) | (0.0020) | (0.0027) | |
| Ever screened in | 0.39 | 0.032 | 0.013*** | -0.0032 | 2286 |
| | | (0.032) | (0.0021) | (0.0029) | |
| Expedited Response Time | 0.027 | 0.015 | 0.00092 | -0.0011 | 176 |
| | | (0.012) | (0.00072) | (0.0011) | |
| Low-Intensity Investigation | 0.22 | 0.025 | 0.0012 | -0.0032 | 1145 |
| | | (0.027) | (0.0018) | (0.0024) | |
| High-Intensity Investigation | 0.059 | 0.015 | 0.0059*** | -0.0022 | 437 |
| | | (0.015) | (0.0013) | (0.0017) | |
| Screen-Out: Community Response | 0.17 | -0.032 | -0.0017 | 0.0011 | 797 |
| | | (0.025) | (0.0016) | (0.0021) | |
| Screen-Out: Family Visitor | 0.10 | -0.0056 | -0.0031*** | 0.0013 | 524 |
| | | (0.021) | (0.0012) | (0.0017) | |
| Screen-Out: Family Given Resources | 0.16 | -0.0011 | -0.0041*** | 0.0016 | 832 |
| | | (0.026) | (0.0014) | (0.0021) | |
| Screen-Out: Sent to Law Enforcement | 0.049 | 0.0013 | -0.0013* | -0.00064 | 195 |
| | | (0.013) | (0.00079) | (0.0011) | |
| Any Response to Family | 0.59 | 0.027 | 0.0015 | -0.0027 | 3131 |
| | | (0.034) | (0.0022) | (0.0030) | |
| Time to Decision (Minutes) | 11.2 | -0.65 | 0.11*** | 0.036 | 3964 |
| | | (0.62) | (0.038) | (0.054) | |
| *Investigation* | | | | | |
| Investigator Risk Assessment Score | 3.34 | 0.086 | 0.27*** | 0.024 | 1124 |
| | | (0.41) | (0.027) | (0.037) | |
| Risk Assessment: Services Recommended | 0.46 | -0.029 | 0.016*** | -0.0029 | 671 |
| | | (0.068) | (0.0041) | (0.0057) | |
| Days Until Investigation Closed | 48.4 | 0.93 | 0.31** | -0.29 | 1579 |
| | | (2.43) | (0.15) | (0.20) | |
| *Outcomes* | | | | | |
| CPS Discovered Injury | 0.069 | 0.00024 | -0.0024 | 0.0035 | 100 |
| | | (0.035) | (0.0018) | (0.0032) | |
| Re-Referred, First Half of Trial | 0.43 | -0.067 | 0.018*** | 0.0026 | 1237 |
| | | (0.047) | (0.0027) | (0.0037) | |
| Removed, Randomized First Half | 0.0075 | -0.0049 | 0.0039*** | 0.00015 | 73 |
| | | (0.010) | (0.0010) | (0.0015) | |
| Future CPS Index, Randomized First Half | 0.034 | -0.17* | 0.021*** | 0.0083 | 422 |
| | | (0.099) | (0.0069) | (0.0090) | |

Notes: This table presents estimates of tool availability and tool interacted with score on a number of primary outcomes. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.16: Effect of Tool on CPS-Related Outcomes for Non-Active Referrals, by Score

| | (1) Mean if Score==1 | (2) Tool Available T | (3) Score | (4) Interaction T · Score | (5) Frequency |
|---|---|---|---|---|---|
| *First Stage* | | | | | |
| Score Ever Seen | 0.088 | 0.58*** | 0.0013 | 0.0069*** | 2027 |
| | | (0.025) | (0.0013) | (0.0024) | |
| *Team Decisions* | | | | | |
| Screened in | 0.28 | 0.077** | 0.011*** | -0.0079** | 1433 |
| | | (0.031) | (0.0023) | (0.0031) | |
| Ever screened in | 0.39 | 0.051 | 0.016*** | -0.0054* | 1988 |
| | | (0.033) | (0.0023) | (0.0032) | |
| Expedited Response Time | 0.026 | 0.018 | 0.0015* | -0.0011 | 169 |
| | | (0.013) | (0.00086) | (0.0013) | |
| Low-Intensity Investigation | 0.22 | 0.042 | 0.0061*** | -0.0060** | 1069 |
| | | (0.028) | (0.0021) | (0.0029) | |
| High-Intensity Investigation | 0.056 | 0.026* | 0.0063*** | -0.0028 | 349 |
| | | (0.015) | (0.0016) | (0.0019) | |
| Screen-Out: Community Response | 0.18 | -0.041 | -0.00095 | 0.0019 | 720 |
| | | (0.026) | (0.0017) | (0.0024) | |
| Screen-Out: Family Visitor | 0.11 | -0.0019 | -0.0019 | 0.00057 | 502 |
| | | (0.022) | (0.0015) | (0.0020) | |
| Screen-Out: Family Given Resources | 0.17 | -0.0020 | -0.0020 | 0.0011 | 786 |
| | | (0.027) | (0.0017) | (0.0025) | |
| Screen-Out: Sent to Law Enforcement | 0.053 | -0.00059 | -0.0011 | -0.00077 | 181 |
| | | (0.014) | (0.00098) | (0.0013) | |
| Any Response to Family | 0.61 | 0.048 | 0.0085*** | -0.0063** | 2854 |
| | | (0.034) | (0.0023) | (0.0032) | |
| Time to Decision (Minutes) | 11.2 | -0.50 | 0.15*** | 0.010 | 3445 |
| | | (0.65) | (0.044) | (0.062) | |
| *Investigation* | | | | | |
| Investigator Risk Assessment Score | 3.31 | 0.035 | 0.28*** | 0.024 | 1012 |
| | | (0.42) | (0.027) | (0.039) | |
| Risk Assessment: Services Recommended | 0.45 | -0.021 | 0.017*** | -0.0035 | 601 |
| | | (0.070) | (0.0043) | (0.0060) | |
| Days Until Investigation Closed | 48.3 | 0.64 | 0.31** | -0.20 | 1431 |
| | | (2.46) | (0.16) | (0.22) | |
| *Outcomes* | | | | | |
| CPS Discovered Injury | 0.065 | -0.00032 | -0.0027 | 0.0049 | 93 |
| | | (0.035) | (0.0017) | (0.0035) | |
| Re-Referred, First Half of Trial | 0.43 | -0.074 | 0.016*** | 0.0040 | 961 |
| | | (0.049) | (0.0033) | (0.0043) | |
| Removed, Randomized First Half | 0.0084 | -0.011 | 0.0028** | 0.00092 | 47 |
| | | (0.012) | (0.0012) | (0.0017) | |
| Future CPS Index, Randomized First Half | 0.060 | -0.22** | 0.018** | 0.015 | 329 |
| | | (0.11) | (0.0081) | (0.011) | |

Notes: This table presents estimates of tool availability and tool interacted with score on a number of primary outcomes. The sample consists of non-active referrals, where no CPS intervention was ongoing at the time of referral. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

Table 1.A.17: Effect of Tool on CPS-Related Outcomes for Active Referrals, by Score

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Tool Available | | Interaction | |
| | Mean if Score==1 | T | Score | T · Score | Frequency |
| *First Stage* | | | | | |
| Score Ever Seen | 0.10 | 0.67*** | 0.0096* | -0.0039 | 356 |
| | | (0.11) | (0.0052) | (0.0072) | |
| *Team Decisions* | | | | | |
| Screened in | 0.28 | -0.24* | -0.010 | 0.015* | 148 |
| | | (0.13) | (0.0062) | (0.0081) | |
| Ever screened in | 0.38 | -0.18 | 0.0057 | 0.012 | 298 |
| | | (0.14) | (0.0068) | (0.0089) | |
| Expedited Response Time | 0.035 | -0.042 | -0.0015 | 0.0015 | 7 |
| | | (0.033) | (0.0021) | (0.0020) | |
| Low-Intensity Investigation | 0.22 | -0.060 | -0.010** | 0.0054 | 76 |
| | | (0.11) | (0.0046) | (0.0061) | |
| High-Intensity Investigation | 0.10 | -0.19** | -0.0015 | 0.0090 | 88 |
| | | (0.082) | (0.0045) | (0.0055) | |
| Screen-Out: Community Response | 0.087 | 0.12 | 0.0018 | -0.0086 | 77 |
| | | (0.099) | (0.0036) | (0.0060) | |
| Screen-Out: Family Visitor | 0.035 | -0.0069 | -0.00060 | 0.0023 | 22 |
| | | (0.040) | (0.0012) | (0.0030) | |
| Screen-Out: Family Given Resources | 0.070 | 0.094 | -0.0033 | -0.0029 | 46 |
| | | (0.093) | (0.0025) | (0.0057) | |
| Screen-Out: Sent to Law Enforcement | 0 | 0.057* | 0.0012 | -0.0033 | 14 |
| | | (0.035) | (0.00075) | (0.0023) | |
| Any Response to Family | 0.42 | -0.060 | -0.0074 | 0.0068 | 277 |
| | | (0.16) | (0.0071) | (0.0096) | |
| Time to Decision (Minutes) | 11.5 | -2.02 | -0.019 | 0.16 | 519 |
| | | (2.53) | (0.12) | (0.14) | |
| *Investigation* | | | | | |
| Investigator Risk Assessment Score | 3.64 | 4.23 | 0.18 | -0.19 | 112 |
| | | (3.79) | (0.14) | (0.22) | |
| Risk Assessment: Services Recommended | 0.54 | -0.42 | 0.0065 | 0.020 | 70 |
| | | (0.41) | (0.016) | (0.027) | |
| Days Until Investigation Closed | 50.1 | 17.6 | 0.68 | -1.59** | 148 |
| | | (11.0) | (0.48) | (0.66) | |
| *Outcomes* | | | | | |
| CPS Discovered Injury | 0.12 | -0.11 | -0.0051 | 0.0024 | 7 |
| | | (0.21) | (0.012) | (0.012) | |
| Re-Referred, First Half of Trial | 0.48 | -0.060 | 0.020*** | 0.0011 | 276 |
| | | (0.17) | (0.0065) | (0.010) | |
| Removed, Randomized First Half | 0 | 0.023 | 0.0070*** | -0.0018 | 26 |
| | | (0.031) | (0.0022) | (0.0033) | |
| Future CPS Index, Randomized First Half | -0.12 | 0.16 | 0.036** | -0.016 | 93 |
| | | (0.31) | (0.014) | (0.020) | |

Notes: This table presents estimates of tool availability and tool interacted with score on a number of primary outcomes. The sample consists of active referrals, where a CPS intervention was already in progress with the household at the time of a referral. Randomization procedure controls are included. Intervention is original tool availability status, and score is defined as the maximum child score for the referral. the rightmost column provides a list of counts for when the outcome variable occurred with a value of at least one. Standard errors are clustered at the household level. Significance reported as: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 1.A.18: Effect of Tool Availability on Screen-In for Non-Active Referrals: Discontinuity Design

| | (1) Screened in | (2) Any Response to Family |
|---|---|---|
| Tool Available=1 | 0.124** | 0.102* |
| | (0.050) | (0.057) |
| Displayed Risk Bin (Discrete) | -0.062 | -0.010 |
| | (0.046) | (0.048) |
| Tool Available=1 × Displayed Risk Bin (Discrete) | -0.010** | -0.010** |
| | (0.004) | (0.005) |
| Hidden Risk Probability (Continuous) | 31.272 | -19.304 |
| | (32.178) | (32.700) |
| Hidden Risk Probability Squared | 6.601 | 0.542 |
| | (25.660) | (26.046) |
| Randomization Controls | Yes | Yes |
| RD Group FEs | Yes | Yes |
| RD Group FEs x Risk Probability | Yes | Yes |
| RD Group FEs x Risk Probability Squared | Yes | Yes |
| Mean of Outcome Variable | 0.32 | 0.64 |
| Observations | 3,329 | 3,329 |

Notes: This table presents discontinuity-based estimates of the effect of score seen on decisions to screen in for CPS investigation (Column 1) and to provide any intervention (screen-in, voluntary resources, informal visitor, or law enforcement; Column 2). Workers can observe the discrete risk score bin (1-20), but cannot observe the underlying continuous risk probability that varies within a bin. The coefficient on tool availability interacted with discrete risk bin is a discontinuity-based estimate of the impact of seeing a score that is one unit greater relative to the control group where the score is not seen. The sample consists of non-active referrals, where no CPS intervention was ongoing at the time of referral. Randomization procedure controls are included. Standard errors are clustered at the household level. Significance reported as: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 1.A.19: Effect of Tool Availability on Screen-In: Child Fixed Effects

| | (1)<br>Screened in |
|---|---|
| Referral Risk Score | -0.004<br>(0.004) |
| Original Status: Tool Available=1 × Referral Risk Score | -0.012**<br>(0.006) |
| Active=1 | -0.326***<br>(0.114) |
| Original Status: Tool Available=1 × Active=1 | -0.471***<br>(0.153) |
| Active=1 × Referral Risk Score | -0.004<br>(0.006) |
| Original Status: Tool Available=1 × Active=1 × Referral Risk Score | 0.031***<br>(0.009) |
| Mean of Outcome Variable<br>Observations | 0.24<br>6,708 |

Notes: This table presents person-level fixed effects model estimates of screen-in on tool availability and prediction score. The design follows changes in score for a child referred to CPS multiple times during the trial. Randomization procedure controls are included, as well as individual fixed effects. Randomization status is the original status assigned to the child (tool available, or no tool), and score is defined as the maximum child score for the referral. The original status base coefficient is omitted because it is constant for an individual over time. The sample consists of both active and non-active referrals. Active referrals are those for which a CPS intervention was already taking place with the household at the time of a referral. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

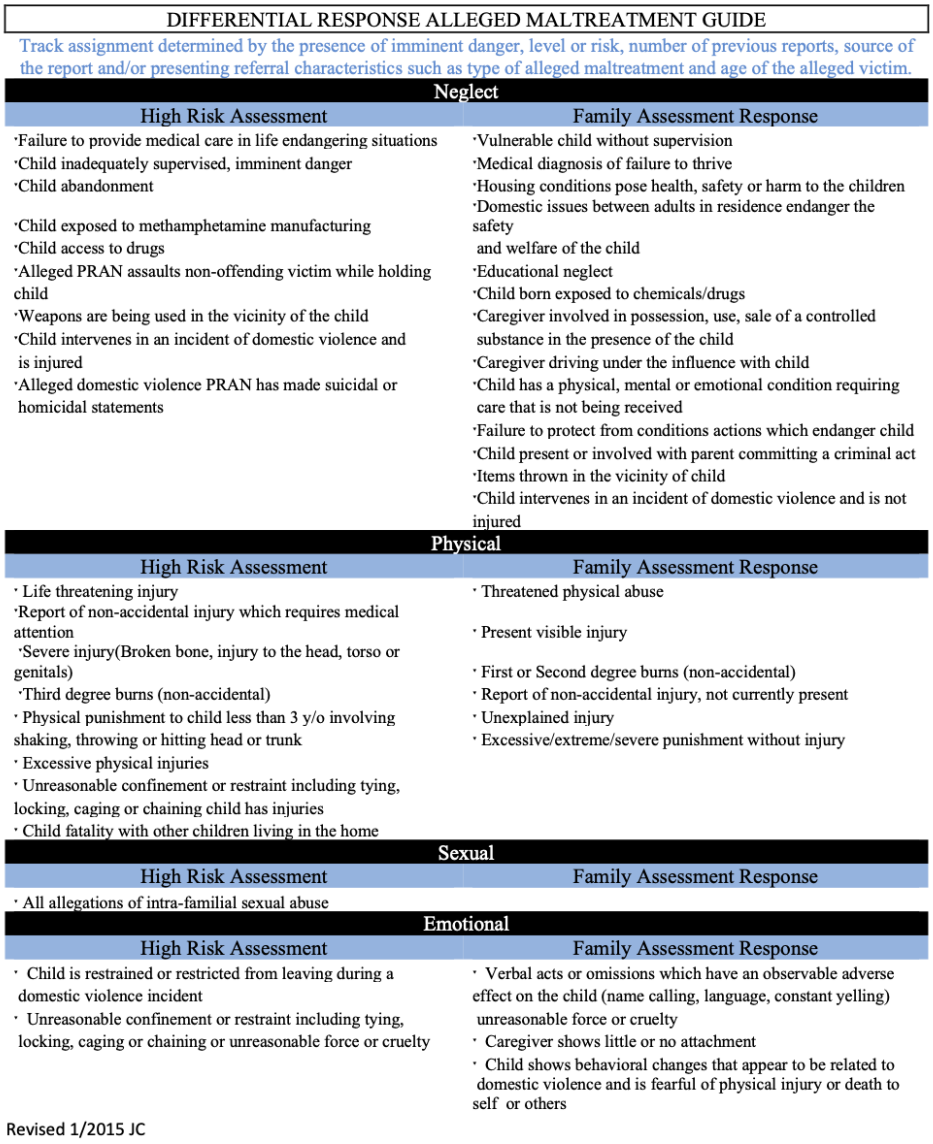
Table 1.A.20: Test of Within-Day Spillovers from Tool Availability

| | Reported Score Seen | | | Screen-In | | | Time to Decision (Minutes) | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Previous Discussion: Tool Available | -0.006<br>(0.024) | 0.017<br>(0.038) | | 0.001<br>(0.022) | -0.004<br>(0.035) | | 0.210<br>(0.350) | 0.049<br>(0.543) | |
| Referral Risk Score | | 0.003<br>(0.003) | | | 0.004*<br>(0.002) | | | 0.091**<br>(0.038) | |
| Previous Discussion: Tool Available=1 × Referral Risk Score | | -0.002<br>(0.003) | | | 0.001<br>(0.003) | | | 0.019<br>(0.052) | |
| Share of Prior Same-Day Discussions with Tool Avail | | | -0.045<br>(0.036) | | | -0.003<br>(0.033) | | | -0.191<br>(0.535) |
| Randomization Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean of Outcome Variable | 0.38 | 0.38 | 0.38 | 0.30 | 0.30 | 0.30 | 11.49 | 11.49 | 11.49 |
| Observations | 3,764 | 3,764 | 3,764 | 3,764 | 3,764 | 3,764 | 3,764 | 3,764 | 3,764 |

Notes: This table presents estimates of tool availability during a team's prior referral on decisions for the current referral. Sample restricted to morning discussions where a prior discussion took place. Randomization procedure controls are included. Score is defined as the maximum child score for the referral. Standard errors are clustered at the household level. Significance reported as: * p<0.1, ** p<0.05, *** p<0.01.

# 1.B Decision Protocol

Figure 1.B.1: Decision Protocol



DIFFERENTIAL RESPONSE ALLEGED MALTREATMENT GUIDE

Track assignment determined by the presence of imminent danger, level or risk, number of previous reports, source of the report and/or presenting referral characteristics such as type of alleged maltreatment and age of the alleged victim.

**Neglect**

| High Risk Assessment | Family Assessment Response |
|---|---|
| ·Failure to provide medical care in life endangering situations | ·Vulnerable child without supervision |
| ·Child inadequately supervised, imminent danger | ·Medical diagnosis of failure to thrive |
| ·Child abandonment | ·Housing conditions pose health, safety or harm to the children |
| ·Child exposed to methamphetamine manufacturing | ·Domestic issues between adults in residence endanger the safety and welfare of the child |
| ·Child access to drugs | ·Educational neglect |
| ·Alleged PRAN assaults non-offending victim while holding child | ·Child born exposed to chemicals/drugs |
| ·Weapons are being used in the vicinity of the child | ·Caregiver involved in possession, use, sale of a controlled substance in the presence of the child |
| ·Child intervenes in an incident of domestic violence and is injured | ·Caregiver driving under the influence with child |
| ·Alleged domestic violence PRAN has made suicidal or homicidal statements | ·Child has a physical, mental or emotional condition requiring care that is not being received |
| | ·Failure to protect from conditions actions which endanger child |
| | ·Child present or involved with parent committing a criminal act |
| | ·Items thrown in the vicinity of child |
| | ·Child intervenes in an incident of domestic violence and is not injured |

**Physical**

| High Risk Assessment | Family Assessment Response |
|---|---|
| · Life threatening injury | · Threatened physical abuse |
| ·Report of non-accidental injury which requires medical attention | · Present visible injury |
| ·Severe injury(Broken bone, injury to the head, torso or genitals) | · First or Second degree burns (non-accidental) |
| ·Third degree burns (non-accidental) | · Report of non-accidental injury, not currently present |
| · Physical punishment to child less than 3 y/o involving shaking, throwing or hitting head or trunk | · Unexplained injury |
| · Excessive physical injuries | · Excessive/extreme/severe punishment without injury |
| · Unreasonable confinement or restraint including tying, locking, caging or chaining child has injuries | |
| · Child fatality with other children living in the home | |

**Sexual**

| High Risk Assessment | Family Assessment Response |
|---|---|
| · All allegations of intra-familial sexual abuse | |

**Emotional**

| High Risk Assessment | Family Assessment Response |
|---|---|
| · Child is restrained or restricted from leaving during a domestic violence incident | · Verbal acts or omissions which have an observable adverse effect on the child (name calling, language, constant yelling) unreasonable force or cruelty |
| · Unreasonable confinement or restraint including tying, locking, caging or chaining or unreasonable force or cruelty | · Caregiver shows little or no attachment |
| | · Child shows behavioral changes that appear to be related to domestic violence and is fearful of physical injury or death to self or others |

Revised 1/2015 JC

Notes: This image shows recommended protocol for investigation responses (high priority = "High Risk Assessment"; low priority = "Family Assessment Response") depending on allegation details.

# Chapter 2

# Safe from Harm? Peers and Criminal Capital Formation in Congregate Foster Care[1]

[1]Sarah Font is an associate professor of sociology in the Department of Sociology and Criminology at Penn State University. Christopher Mills is a PhD candidate in economics in the Department of Economics at Princeton University (chris.mills@princeton.edu). We thank Lonnie Berger, Janet Currie, Matthew Lindquist, Matthew Stagner, and conference and seminar participants at the AEA Annual Meetings, APPAM, and Princeton Industrial Relations Section for their valuable comments and suggestions. We also thank the Wisconsin Department of Children and Families, Department of Health Services, Department of Public Instruction, Department of Workforce Development, and Department of Corrections for the use of data for this analysis, but these agencies do not certify the accuracy of the analyses presented. This research was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development grants R21HD091459 (University of Wisconsin-Madison) R01HD095946, P50HD089922, and P2CHD041025 (Penn State University). The Adoption and Foster Care Analysis and Reporting System (AFCARS) data used for motivating statistics were made available by the National Data Archive on Child Abuse and Neglect, Cornell University, Ithaca, NY, and have been used with permission. Data from AFCARS were originally collected by the Children's Bureau. Funding for the project was provided by the Children's Bureau, Administration on Children, Youth and Families, Administration for Children and Families, U.S. Department of Health and Human Services. The collector of the original data, the funder, the Archive, Cornell University and their agents or employees bear no responsibility for the analyses or interpretations presented here. This paper uses confidential data from the Wisconsin Institute for Research on Poverty. Replication files are available in the Online Appendix (forthcoming). Both authors received IRB approval for the analysis of restricted use secondary data from their respective institutions. The authors have no additional disclosures.

## 2.1 Introduction

More than 20 percent of young adult prison inmates in the United States have spent time in foster care (Survey of Prison Inmates, 2016).[2, 3] Although the pipeline from foster care to incarceration, homelessness, and other costly outcomes has challenged policymakers for decades, little is known about how different features of the foster care system influence life trajectories. For youth residing in congregate (group-based) placements in particular, it has long been posited that peer effects are a critical mechanism that may enhance risk; that by grouping children who are delinquent or have serious behavioral problems together, antisocial norms and behaviors may be reinforced or enhanced.[4] Yet the extent to which children in foster care are adversely impacted by exposure to high-risk peers has yet to be demonstrated. In many states, children in foster care due to abuse and neglect live alongside children in state custody for delinquency. Since approximately one million children have lived in congregate care settings since 2000 (authors' calculation, AFCARS 2000-2017) and approximately 46 percent of current teens in foster care have experienced at least one congregate care placement, understanding the effects of peers in these settings is urgent.

Peer effects in congregate care are difficult to study rigorously because youth are placed in facilities based on various constraints and preferences, including the supply of beds, the availability of less restrictive alternative settings (family foster homes), proximity to one's family of origin, and one's own characteristics and treatment needs. Facilities are also

[2]Authors' calculation using the 2016 Survey of Prison Inmates, a nationally representative survey of incarcerated adults in state and federal prisons. 22 percent of adults under 30 (22 percent of males, 24 percent of females) report having spent time in foster care before age 18. The majority of these respondents report having lived in a foster care "agency or institution", such as a congregate care setting. Rates are similar using administrative data from Wisconsin.

[3]Foster care is a temporary living arrangement for children and youth who are removed from their homes due to imminent risk of abuse or neglect or, in a smaller number of cases, child behavior problems. Foster care providers include state-licensed family homes (kin and non-kin) and congregate care facilities (group homes, residential facilities). Children and youth who leave foster care may be reunified with their biological parents, live with kin, live with an non-kin adoptive family, or age out of the system.

[4]The narrative that placing children is restrictive settings induces deviant behavior is common in news coverage: "'I knew it was wrong to lie,' said Geard, 17, sitting in a courthouse in this town of shuttered coal fields and boarded-up shops. 'But I was locked up with kids who were rapists and murderers. One kid even beat up a judge like it was nothing. That's just not me'" (Wax-Thibodeaux, 2019).

not required to accept all placement requests, such that fewer placement options exist for children with more challenging behavioral or health concerns. Our study relies on the natural variation in peer characteristics *within* a facility over time to identify the causal impacts of co-residence with delinquent and high-risk peers. We leverage 14 years of foster care placement data from Wisconsin and more than 30 years of linked administrative data on family history to estimate pre-placement imprisonment risk for youth in congregate care and address various threats to the identification of peer effects.

We show that being exposed to delinquent peers in congregate care has no significant effect on one's likelihood of future imprisonment. Children in congregate care do have a high risk of being exposed to delinquent peers: 63 percent of peers (children in the same care settings) were placed for delinquency-related reasons and approximately one in 10 peers are predicted to be sentenced to prison as a teenager. However, we find that a one standard deviation increase in average peer delinquency for a youth in congregate care has no significant effect on the likelihood of entering prison by age 20. Moreover, peer delinquency has no effect on an aggregated index of future wellbeing that includes deviant behavior such as delinquency-related terminations in placement and running away. Concerning other dimensions of wellbeing, our results show that a one standard deviation increase in peer imprisonment risk is associated with a modest one percentage point (2.5 percent) increase in the likelihood of high school dropout, but the estimated magnitude is not robust across all peer measures. The effects of peer composition on female teen parenthood are on average negligible. Overall, our results suggest that, on average, peers are not a primary channel for shaping youth justice system contact and wellbeing beyond their time in foster care. Whereas prior research has largely focused on peer effects within correctional facilities – among individuals whose criminal careers have already commenced – this study sheds light on the potential role of peer effects in the onset of criminal activity among a population of children at heightened risk.

A second goal of this paper is to understand heterogeneity in peer influences by facil-

ity characteristics and child demographics, complementing prior empirical research on peer influences and human development in other settings (for example, Sacerdote, 2001; Aizer, 2009; Bayer at al., 2009; Stevenson, 2017). Prior research on peer effects has largely focused either on juvenile detention facilities (wherein all sampled youth have initiated their criminal careers) or on youth in general settings, such as schools (where the number of peers is larger and risk distribution of peers is broader). In contrast, we examine peer effects in a population that faces heightened risk of criminal activity but has largely yet to have formal involvement with justice systems. Thus, our study considers the generalizability (rather than repeatability; Freese and Peterson, 2017) of prior peer effects studies. First, we find that (lack of) peer influence is relatively consistent across youth disaggregated by race, gender, and age, and delinquency status, with two exceptions: the effects of peers on high school dropout are primarily concentrated among Black youth and youth who had a parent imprisoned before age 10. Second, we ask whether facility-level characteristics enhance or mitigate peer influence. Our peer effect estimates are statistically insignificant from zero across all facility types and we find no evidence that variation in placement-related experiences within a facility over time, such as a youth's number of peers and length of stay, enhance peer influence. Third, we consider the structure of peer effects, finding that future outcomes are relatively invariant to nonlinear models of peer delinquency and delinquency of peers of the opposite gender.

In addition to providing new insights on the relationship between peers, crime, and risky behavior, this study contributes to an emerging empirical literature on the effects of foster care placement on later-life outcomes. Doyle (2007; 2008) leverages quasi-random variation in the assignment of child protective services (CPS) investigators in Illinois, finding that at the margin, foster care placement is associated with a more than doubling of the likelihood of adult incarceration. Recent studies from different states, however, find ambiguous or beneficial impacts of foster care placement on juvenile delinquency (Gross and Baron, 2022), adult criminal justice contact (Baron and Gross, 2022) and test scores (Bald et al., 2019;

Gross and Baron, 2022). One potential explanation for the different findings is that the experience of foster care may vary across time and place (Font and Gershoff, 2020) or that the marginal child removed varies. Our study differs from previous work focusing on the marginal child removed in that we focus specifically on the higher-risk subset of children who experience congregate care – a setting typically used only for older youth and youth who have difficulty achieving stability within a less-restrictive environment. In addition, within that higher-risk group, we measure outcomes for a wide range of youth in congregate care, and focus on a child's experiences *conditional* on being placed on congregate care, as illustrated in Figure 2.A.1.[5] On the whole, our results suggest that, conditional on a child having been placed in congregate care, variability in the types of peers a child receives has little effect on a range of outcomes.[6]

Finally, this paper provides policy implications for child welfare during a period of significant change. Bipartisan concerns about the purported adverse impacts of congregate care on children's development (No Place to Grow up: How to Safely Reduce Reliance on Foster Care Group Homes; Senate Hearing 114-273, 2015) and the high costs of congregate care relative to family foster care settings led to significant federal reforms under the Family First Prevention Services Act (FFPSA; Bipartisan Budget Act, 2018). Under FFPSA, federal reimbursement for congregate care is time-limited except in special circumstances, and states are expected to use qualified residential treatment centers when congregate care is required beyond two weeks. This legislation is designed to move a greater share of children with behavioral and mental health challenges into family foster homes. This has led to the closure of many facilities and may also increase the need for remaining facilities to co-house CPS-placed youth with juvenile corrections -involved youth. Our findings suggest that large negative peer effects are unlikely even in settings that combine non-delinquent and delin-

[5]Our work complements the approach of Taylor (2021), who leverages county variation in family home capacity to suggest that initial placement in a non-kin family home improves long-run youth outcomes relative to congregate care placement.

[6]There is a very limited amount of rigorous empirical research on the effectiveness of interventions in foster care (c.f., Bergström et al., 2020).

quent youth. Although peer effects are one of several possible mechanisms through which congregate care may adversely impact youth, it is important to consider because peer effects are less malleable to policy or programmatic reforms. That is, congregate care is positioned at the far end of the continuum of care – a specialized setting for high needs children - and thus, absent unethical and likely damaging isolation of high needs children (e.g., solitary confinement or other exclusion), some exposure to high-risk peers is an inevitable feature of congregate care. In contrast, other child-centered criticisms of congregate care pertain largely to implementation (e.g., lack of treatment and skills programs, staffing issues) or can be mitigated with intensive efforts (e.g., programs and policies that facilitate access to stable attachment figures during and after placement). The lack of evidence for negative peer effects of congregate care may provide cautious optimism for the malleability of adverse outcomes during and following congregate care.

The paper is organized as follows. Section 2 briefly describes the context of foster care. Section 3 discusses the sources of data used in the paper and descriptive statistics. Section 4 introduces the methodological framework. Section 5 presents results and discussion. Section 6 concludes.

## 2.2 Background

Six percent of children in the United States are placed in foster care before age 18 (Wildeman and Emanuel, 2014). [7] Foster care is a temporary living arrangement for children and youth who are removed from their homes, typically due to elevated risk of abuse or neglect identified through an investigation from CPS, as illustrated in Figure 2.A.1. However, recent estimates suggest that 8-35 percent of children in foster care did not come into care following a CPS investigation (Drake et al., 2021), some of whom may have been placed for mental health or delinquency-related reasons. Children involved with the juvenile justice system may be

[7] This does not include children who are placed in out-of-home care settings under the sole jurisdiction of the juvenile justice system.

rerouted into child welfare because it allows states to draw down federal (versus state and local) funds [8].

Regardless of reason for entry, all children in foster care are subject to state and federal guidelines that dictate children's placement in the least-restrictive, most family-like setting available that meets their developmental needs. Within that framework, kinship care (placement with relatives) is most preferred, followed by non-relative family foster homes. Policies and best practice guidelines discourage congregate care for children who can be safely and appropriately cared for in relative or non-relative family-based settings (National Conference of State Legislatures, 2016; Figure 2.A.2). States are also discouraged from using congregate care where possible by cost constraints: congregate care settings are substantially more expensive than family foster homes and, in 2016, federal funding only covered about 1/3 of congregate care expenditures [9]. The enactment of the Family First Prevention Services Act (2018) further limited federal reimbursements for such facilities. In addition to policy preferences related to setting type, federal guidelines further encourage placement in facilities near the child's family and school of origin (42 U.S.C. Sect. 675(5), 2010) in order to facilitate relational and educational continuity.

Congregate care includes group homes (semi-residential settings designed for three to eight youth), residential facilities (more intensive settings for larger numbers of youth), and temporary settings such as emergency shelters which are designed to accommodate youth until another setting is available but may in practice house youth for extended periods of time. Congregate care facilities can be for-profit or nonprofit corporations, and they may specialize in particular demographic groups (e.g., males, older teens) and conditions (e.g., pregnant youth). In Wisconsin, there are three groups of children who may require out-of-home placement in a congregate care setting. Children in Need of Protection (referred to as "CHIPS") are typically removed through CPS due to abuse or neglect. Juveniles in

[8] https://www.aei.org/op-eds/the-horrifying-misuse-of-foster-care/

[9] https://www.childtrends.org/wp-content/uploads/2018/12/StateAndLocalSFY2016_ChildTrends_December2018.pdf

Need of Protection (referred to as "JIPS") are typically cases initiated by parents when their child is uncontrollable, including children who are running away or refusing to attend school. The final group is delinquent youth, who are adjudicated through the juvenile justice system. Congregate care facilities may accept children from all groups but may also restrict eligibility based on group designation [10] [11]. Facilities may also reject a placement request if they believe the child's needs will not b adequately met or if the child in need of placement may adversely affect other children already at the facility. [12]

Placement in congregate care is intended for when a child cannot safely function in any available non-congregate (kin or non-relative) home; however, congregate care placement may also occur due to a lack of alternative options (no kin or non-relative foster homes have been identified or approved for the child; Alpert and Meezan, 2012). Because many states face substantial barriers to recruiting foster families for older children and children with more severe mental or behavioral challenges (Government Accountability Office, 2015), those placed in congregate care are typically teens (U.S. Children's Bureau, 2015): many have substantial mental health and behavior problems and previous unsuccessful kinship or non-relative placements (Baker et al., 2005; James et al., 2006, 2012; Barth et al., 2007). Children placed through JIPS or delinquency proceedings are most likely to require a congregate care setting due to more limited alternatives and greater supervision and monitoring requirements; they may also remain in such placements for a longer period of time if their parents are unwilling to assume custody.

Approximately 33 percent of teens in foster care currently reside in a group home (14 percent) or institution (19 percent) (authors' calculation, AFCARS 2017).[13] Overall, approx-

[10]Group homes: https://dcf.wisconsin.gov/files/cwlicensing/pdf/gh.pdf

[11]Residential facilities: https://dcf.wisconsin.gov/files/cwlicensing/pdf/rcc.pdf

[12]Facility contract template describes the following acceptable reasons for rejecting a placement request: : a) Needs of the child are greater than the staff in the Contractor's home can safely manage. The Contractor must describe the specific needs that they are unable to meet. b) The child does not fit within the age requirements that can be placed at the home. c) The histories and experiences of a child already placed at the facility are contrary to the best interests of the child seeking placement (i.e. child is a member of a rival gang). https://dcf.wisconsin.gov/files/mcps/contracts-grants/dmcps-gh-scope-of-services-2022.pdf.

[13]Thousands of children and youth in non-kin family homes also live with one or more non-related foster care peer(s).

imately 46 percent of teens currently in foster care have had at least one congregate care placement, and approximately one million children and youth have lived in a congregate care setting while in foster care since 2000 (authors' calculation, AFCARS 2000-2017).

Congregate care may impose negative consequences on youth for at least three reasons. First, the staff working in congregate care facilities tend to be poorly trained and underpaid, and turn over at high rates [14]. Although other forms of foster care are not uniformly stable and high quality, round-the-clock foster parents are better situated to provide stable adult attachment figures and the developmentally normative experience of having a family than shift-based staff (Kelly, Anthony, and Krysik, 2019). Further, and by reducing opportunities for a permanent home (Elgin et al., 2015). Second, grouping children with antisocial peers provides ample opportunity to develop criminal networks and to internalize antisocial beliefs and expectations (McGloin and Thomas, 2019). Congregate care facilities are expected to ensure children receive health and education services, but those can be provided on or off site as appropriate [15]. Many children in congregate settings attend regular schools, although their ability to participate in formal extracurricular activities and informal social gatherings may be more limited, such that their peer interactions during non-school hours (e.g., meals, social activities, and group-based skills or treatment services) are predominantly with co-resident children. Third, congregate care facilities may provide a low quality of care and supervision, with inadequate health care, safety protections, and educational or vocational opportunities (Freundlich, Avery, and Padgett, 2005; 2007). On this last dimension, congregate care facilities are highly variable and our data do not contain pertinent details. Together, these factors – lack of stable adult attachment figures, socialization by antisocial peers, and diminished access to conventional pathways into adulthood—would suggest increased risk of criminal involvement. Although observational studies generally indicate a positive association between congregate care placement and delinquency relative to other forms of foster care (Ryan et al., 2008; Robst et al., 2011), the mechanisms underlying such associations

[14] https://pccyfs.org/wp-content/uploads/2021/11/Workforce-Shortage-Final-11.17.21.pdf
[15] https://dcf.wisconsin.gov/files/mcps/contracts-grants/dmcps-gh-scope-of-services-2022.pdf

are poorly delineated. Studies generally find little evidence that congregate care has adverse effects on maltreatment victimization (DeSena et al., 2005; Font, 2015), behavior problems (McCrae et al., 2010; James et al., 2012), legal involvement (Lee and Thompson, 2008), or academic progress (McCrae et al., 2010).

Peer influence is an understudied dimension of foster care. Prior research has found that the characteristics of peers in juvenile incarceration facilities are strongly predictive of a youth's own incarceration after release (Bayer et al., 2009; Stevenson, 2017), especially peers who are from unstable homes, are themselves victims of abuse, and who have high aggression and poor social skills (Stevenson, 2017). Negative spillovers from youth exposed to antisocial relationships has also been documented in other contexts. Carrell and Hoekstra (2010) find that children from homes with domestic violence cases increase the classroom misbehavior of their peers, and Carrell et al. (2018) find that these negative effects persist and are reflected in young adult earnings. Figlio (2007), Aizer (2009), and Neidell and Waldfogel (2010) likewise find negative spillovers from potentially disruptive students on the academic achievement of classroom peers. Social peer effects for adolescents have been found in other residential settings, such as the influence of randomly-assigned college roommates on binge drinking (Duncan et al., 2005), college peers on academic cheating (Carrell at et al., 2008), and primary and secondary school peers on fitness and obesity (Carrell et al., 2011; Lim and Meer, 2018). Finally, there exists a vast empirical literature relating neighborhood influences to crime (Case and Katz, 1991; Ludwig et al., 2001; Kling et al., 2005; Damm and Dustmann, 2014; Bernasco et al., 2017).[16] These studies provide reasons to suspect that peer composition could influence a youth's development in congregate care.

Peer effects may be particularly strong for children in congregate care due to separation from the primary unit of socialization, the family. However, negative peer effects of congregate care may not materialize for several reasons. First, a person is most influenced by associates with whom they have a strong attachment relationship and frequent and long-term

[16]Lindquist and Zenou (2019) present additional examples of empirical work on social influence and crime.

contact (McGloin and Thomas, 2019). Yet, children spend a relatively short proportion of their childhoods in congregate care and there is a constant churn of children in and out of congregate care facilities (youth in our sample spend an average of 50 days with a given peer, consistent with multi-state estimates that the median span of congregate care placement is of 55-60 days [17]. As such, children's exposure to any particular set of peers is especially short and perhaps unlikely to result in meaningful peer connections that extend beyond the end of the placement. Second, children in foster care often have extensive criminogenic influences in their families (Austin, 2016; Giffords et al., 2020) and neighborhoods (Gracia et al., 2018) of origin. Given that most children in congregate care entered foster care in late childhood and return to their families (and communities) of origin, the relatively brief exposure to unrelated deviant peers may be a proverbial drop in the bucket.

## 2.3 Data and Descriptive Statistics

This study uses data from the Wisconsin Administrative Data Core (WADC), housed at the Institute for Research on Poverty at the University of Wisconsin-Madison. WADC is an integrated data system in which individuals and families involved in state-administered social welfare programs – including child protective services, foster care, state prisons, public benefits (for example, cash and food assistance), and unemployment insurance – can be linked by inter- and intra-generational family relationships, household membership, and social welfare program participation, and can be followed over time. These data are also linked to public school education records from the Wisconsin Department for Public Instruction (DPI).

Our primary analysis sample consists of all congregate care placements in Wisconsin that lasted at least two weeks for youth who were born before 2000 and who averaged at least one same-gender peer during their stay. Wisconsin counties had fully converted foster care records to electronic format by July 1, 2004. We therefore include all placements beginning on or after this date in our analysis sample. The primary sample excludes Milwaukee County due

[17] https://fcda.chapinhall.org/wp-content/uploads/2021/08/Congregate-Care-Report-Final-8.9.2021.pdf

to the county's incomplete placement records for delinquent youth. For ease of interpretation, we also exclude a small number of placements with siblings (less than three percent of sample; n=279), placements of youth older than age 17 (n=14) and youth who were in prison prior to placement (n=841), although we include these types of placements as part of the measure of peer characteristics. Finally, we exclude a small number of youth who had no co-resident peer at the beginning of their placement (n=81). The final congregate care sample consists of more than 8,000 placements involving approximately 6,500 youth across 89 group homes, 58 residential facilities, and 50 temporary living centers.

Youth in the sample can be placed in congregate care for child maltreatment reasons (via Child Protective Services), sub-delinquency behavioral or family problems (e.g., JIPS), or delinquency. Peer measures are constructed using the complete available sample of children and youth in congregate care from 2004-2017, even if these peer placements do not themselves appear in the analysis sample. In addition to group homes and residential facilities, we also include placements in settings labeled as "shelter" because these settings exhibit many features similar to traditional congregate care placements, operate as potential placements when traditional congregate care or family home options are unavailable, and may become more representative of congregate care under changes imposed by the Family First Prevention Services Act.

Although foster care systems can vary widely state to state, Wisconsin is not atypical from national averages on relevant observable measures, including entry rate (4 per 1,000 in 2014-2018), average length of stay in foster care (WI 20.2 months; US 21.0 months), percentage placed in congregate care (WI 9 percent; US 10 percent), or age distribution (WI 32 percent ages 11-20 years; US 36 percent)[18]. Of possible exception, Wisconsin does have a larger share of children in foster care for child behavioral issues and family concerns other than abuse and neglect (Child Trends link). However, we cannot discern whether this is a reflection of funding categories (whether children are served through child welfare

[18]gwww.childtrends.org/publications/state-level-data-for-understanding-child-welfare-in-the-united-states

and thus included in federal foster care data) or a true difference in the total number of children living away from home for non-maltreatment reasons. Wisconsin is also similar to the national average and trendlines for juvenile justice placements: WI had 260 per 100,000 youth residing in detention, corrections, or residential facilities in 2007 (US: 272 per 100,000), declining to 93 by 2019 (US:114).

Table 2.7.1 provides an overview of outcomes and sample characteristics. The primary dependent variable is entry into a Wisconsin state prison before age 20, which is derived from Department of Corrections records. Over 10 percent of youth who experienced congregate care (columns 1-3) sample spent time in state prison as a teenager, about twice the rate of same-aged youth in family homes.[19] Supplemental outcome measures include running away, and juvenile detention as the reason for the discontinuation of the current placement. The future wellbeing index is a standardized (mean 0, variance 1 within the main analysis sample) index of outcomes that is created by first standardizing each individual outcome – prison before 20, placement ending in detention, placement ending in a runaway, and high school dropout – with mean 0 and variance 1, summing the standardized outcomes together, and then re-standardizing the sum.[20] A higher score indicates a more beneficial aggregate outcome.

Peer attributes are measured two ways: (1) a child's share of delinquent peers in a specific placement and (2) the mean teen imprisonment risk of a child's peers. Risk scores are constructed using a least absolute shrinkage and selection operator (LASSO) logit procedure predicting future imprisonment, described below. Across placement types, residential and detention facilities are the most institutional form of foster care and family homes the least, with youth in facilities facing higher pre-placement imprisonment risk (own risk score,

[19]Youth ages 17 and above in Wisconsin are tried as adults. Juvenile courts may waive jurisdiction to adult courts for youth ages 14 and above facing particularly serious allegations (c.f., https://www.grgblaw.com/wisconsin-trial-lawyers/wisconsin-laws-and-policies-for-juvenile-offenders-charged-as-adults#:~:text=Most%20children%2010%20years%20and,be%20charged%20as%20an%20adult; https://docs.legis.wisconsin.gov/statutes/statutes/938).

[20]We omit teen parenthood from the index because we apply that outcome only to female youth in our sample. The dropout indicator includes school dropout prior to ninth grade, but 99 percent of dropout observations take place during high school grades.

average peer risk score) than those in other settings. Youth in our residential facilities sample are placed with on average of 14 non-sibling peers, compared to 4.5 and 4.7 for group home and temporary settings, respectively. Youth in residential facilities have higher rates of Supplemental Security Income (SSI) receipt (proxy for disability; $p < 0.01$), consistent with residential facilities sometimes catering to specialized medical or behavioral needs for youth.

Our analyses include a wide range of youth characteristics, birth family characteristics, and foster care experiences as covariates, which can be found in Table 2.7.1. Youth characteristics include year of birth, gender, race/ethnicity, receipt of SSI, age at entry into foster care, and pre-removal referrals. Race/ethnicity is coded as White (non-Hispanic), Black (non-Hispanic), Hispanic, Native American (alone or in combination with another racial or ethnic category), and other (including unknown). Approximately a fifth of the sample is Black, even though Black residents made up five to seven percent of Wisconsin's population during the sample period. For children enrolled in a public school prior to congregate care placement, the presence of an IDEA qualifying disability is grouped into the following categories, based on the primary condition that required special education services: Autism, learning disability, speech or language impairment, intellectual disability or traumatic brain injury, emotional or behavior disturbance, physical impairment, significant developmental delay (only for children under age 9), and other health impairment (includes ADHD).

Birth family characteristics extend back to the late 1980s and early 1990s and include a youth's birth mother's age at the time of her first birth and whether the youth was born to married parents. We also construct measures of youth's birth mothers' and fathers' (when identified) cash assistance receipt, food assistance receipt, prison history, average wages, and employment spanning the youth's birth through age 10.[21]. 32 percent of the youth in our sample were born to married parents, and more than 86 percent received food stamps before entering foster care.

[21]Nearly all placements in our analysis occur after age 10.

Finally, Table 2.7.1 describes foster care experiences for youth in the sample. Youth in congregate care experience on average more than seven foster care placements during their lifetime, and the mean placement duration in our sample is approximately 114 days with a median of 62 days, ranging from 14 days to multiple years.

Our sample is distinct from prior studies, focusing on child-welfare-involved individuals. Compared to prior samples of children from juvenile incarceration facilities (Bayer et al., 2009; Stevenson 2017) youth in our sample are younger on average, are less likely to have previously been incarcerated, and are more likely to be female and white. This paper's sample is placed in comparatively more recent years (2004-2017, compared to 1997-1999 and 2007-2011, respectively) and in Wisconsin instead of Florida. These contextual differences suggest that the results in this paper generalize better to the population of child-welfare-involvement children, for whom peer dynamics and influence may differ relative to those among juvenile delinquents. The children in our sample are are more at-risk of future incarceration than the general population and less at-risk than those in juvenile incarceration facilities. Given lower levels of baseline antisocial behavior but unstable family background, it is not ex-ante clear whether peer influence would be weaker or stronger in congregate foster care compared to detention-related settings.

### 2.3.1 Risk Score Construction

One challenge in estimating the causal effect of peer influence is bias from reflection, caused by the possible simultaneous influence of a youth's own criminality on peer criminality and vice-versa. To address this problem, we construct two exogenous measures of predetermined peer risk (Manski, 1993). First, we use the share of peers who are in the congregate care setting due to a juvenile delinquency case. In Wisconsin, children entering foster care through the child welfare system (due to risk of abuse or neglect) may share placements with children entering foster care through a juvenile delinquency case. Using delinquency-related placements as a measure of peer criminality has the advantage of capturing peer behavior prior

to placement but still very close in time to the placement itself, ameliorating concerns about underestimating peer effects when using past behavior to approximate for present behavior (Hanushek et al., 2003). Second, we predict own and peer risk of imprisonment using pre-placement risk factors that are not influenced by the current placement. Our risk score is estimated using a LASSO logit machine learning procedure trained on the full sample of congregate placements for youth in foster care from 2004-2019 and born prior to 2000. The model is trained to predict a child's likelihood of entering prison before age 20. The model is also used to predict out-of-sample risk for children born during and after 2000, who are too young to have an imprisonment outcome but may appear as peers. Figure 2.A.3 shows the accuracy of the risk score prediction in the full sample.

The risk score is estimated using a rich set of pre-placement child and family characteristics. Child demographic and history characteristics are child gender; year of birth; SSI receipt; SNAP receipt; indicator for placement due to own delinquency; number of previous child welfare referrals and foster care placements; number of previous days in foster care; and indicators for any previous placement in a group home, residential facility, temporary placement, or detention center, respectively. Pre-placement parent characteristics are paternity; whether parents were unmarried when child was born; number of children born to mother; mother's first birth was a teen birth; and the mean value or frequency of mother's and father's imprisonment, food stamp (SNAP) receipt, public welfare benefits, employment, and wages (UI-eligible jobs), respectively, measured monthly prior to the child turning 10. The set of characteristics also includes indicators for unidentified fathers and mothers. All missing values are imputed as the sample median, and the large majority of characteristics are non-missing except in the case of children with unidentified fathers.[22] To improve prediction accuracy, we include pairwise interactions of all features in the model. The LASSO procedure selects a sparse subset of features used to predict imprisonment risk, yielding a relatively

[22]Of 84,850 placements for youth born on or before 1998, approximately 7,000 are missing paternal history. These youth are comparable in terms of teen imprisonment rates. 85 percent of the estimation sample has no missing attributes.

simple yet accurate prediction model. Notably, we do not explicitly train on race/ethnicity in our prediction model, although given racial disproportionality in criminal justice contact it is possible that features selected by LASSO may be correlated with race/ethnicity. Supplemental regression specifications in the results section include controls for own race/ethnicity and peer race/ethnicity composition as an attempt to safeguard against disparities that may arise when using proxies for imprisonment risk (Obermeyer et al., 2019).

The features empirically selected for our model are listed in Table 2.A.1, in order of greatest to least "importance" (standardized penalty weight magnitude). The model weights gender interactions heavily, given the disproportionate imprisonment risk for males. Other leading risk indicators include past detention placement, current delinquency-related placement, whether a youth's mother was a teen at first birth, births to unmarried parents, public assistance receipt, and past incidence of CPS contact.

## 2.4 Empirical Strategy

Our identification strategy relies on the natural variation in delinquent peers after controlling for the placement provider and time of entry. Congregate care providers typically have an inelastic supply of beds. Therefore, the natural entry and exit of youth into a given setting has a meaningful impact on peer composition. A youth who entered a given placement in March, for example, may be exposed to a different level of peer risk than a similar youth who entered the same setting in November of the previous year. Figure 2.A.4 provides an example of the assignment process where a youth is placed in the first suitable provider with an available bed, where the peer composition changes as a result of that youth's placement.

Leveraging variation in the share of delinquent peers in a setting over time, we estimate the following model:

$$Y_{ijt} = \alpha + \beta_1 sharedelinqpeers_{ijt} + \beta_2 X_{ijt} + \lambda_j + \eta_t + \epsilon_{ijt}, \tag{2.1}$$

and similarly for mean peer imprisonment risk:

$$Y_{ijt} = \alpha + \beta_1 peerscore_{ijt} + \beta_2 X_{ijt} + \lambda_j + \eta_t + \epsilon_{ijt}, \tag{2.2}$$

where $Y_{ijt}$ is the outcome for youth $i$ entering provider $j$ in year $t$, for example teen imprisonment. We estimate peer effects separately for delinquent and non-delinquent youth. The coefficients of interest, $\beta_1$, can be interpreted as the percentage point increase in the likelihood of going to prison before age 20 associated with a one unit increase in the average peer measure over the duration of a youth's placement, for maltreated (non-delinquent) youth.

The model includes provider-by-gender fixed effects $\lambda_j$, which control for time-invariant differences across different placement settings (for example, some providers cater to youth with specific behavioral needs). $\eta_t$ are additive placement year and placement quarter fixed effects, which control for statewide trends by placement year and season. $X_{ijt}$ is a vector of time-variant placement characteristics and individual history controls, including own risk score, provider-by-gender fixed effects, year and quarter of entry fixed effects, and average number of non-sibling peers during placement. In robustness checks to evaluate the stability of the peer effect estimates after controlling for individual and peer characteristics we also include individual demographic controls (birth year, race/ethnicity, gender, and SNAP and SSI eligibility), individual history controls (past CPS investigations, previous number of days in foster care, number of previous placements, age of first foster care placement), family controls (mother teen birth and nonmarital birth, mother's total number of kids, and mother's prison history before the child turned 10), and demographics of peers at the beginning of placement. $\epsilon_{ijt}$ is the random error term. All standard errors are two way clustered by provider and youth.

Because many providers are gender-specific and others typically group youth together by gender, we calculate peer characteristics within gender. The share of delinquent peers measure is calculated as the fraction of non-sibling peers of the same gender who were placed due to delinquency, weighted by the number of days in the same setting as the index youth, and then standardized. Specifically, the same-gender measures of peer delinquency and peer

imprisonment risk, respectively, are calculated as follows:

$$sharedelinqpeers_{ijt} = z\left(\frac{\sum_{k\neq i} days_with_peer_{i,k} * \mathbb{1}\{same_gender_{i,k}\} * delinquent_k}{\sum_{k\neq i} days_with_peer_{i,k} * \mathbb{1}\{same_gender_{i,k}\}}\right) \tag{2.3}$$

$$peerscore_{ijt} = z\left(\frac{\sum_{k\neq i} days_with_peer_{i,k} * \mathbb{1}\{same_gender_{i,k}\} * riskscore_k}{\sum_{k\neq i} days_with_peer_{i,k} * \mathbb{1}\{same_gender_{i,k}\}}\right), \tag{2.4}$$

where $sharedelinqpeers_{ijt}$ is the standardized fraction of non-sibling delinquent peers, and $peerscore_{ijt}$ the standardized average non-sibling peer imprisonment risk score, for youth $i$ entering provider $j$ at date $t$. Each measure places more weight on peers with whom a youth spent more time, as in Stevenson (2017). Our main findings are robust to using average peer composition on a youth's first day of placement, which offers the advantage of being unrelated to a youth's own length of stay but the disadvantage of measurement error in peer composition. Figure 2.A.5 illustrates the general similarity between beginning-of-placement peer risk and full-placement peer risk weighted by shared duration. Finally, peer measures are standardized within gender to have a mean of zero and a standard deviation of one within gender and across the full analysis sample, as indicated by $z(\cdot)$.

The identifying assumption in interpreting $\beta_1$ is:

$$E[\epsilon_{ijt}|sharedelinqpeers_{ijt}, X_{ijt}, \lambda_j, \eta_t] = 0, \tag{2.5}$$

where, conditional on provider and date of entry, variation in one's peers' risk at the start of placement is assumed to be as good as random. We provide support for this assumption in the paragraphs below.

To identify mechanisms of peer influence, we study peer effect heterogeneity by individual and provider characteristics using a similar model:

$$\begin{aligned} Y_{ijt} = \alpha &+ \beta_1 sharedelinqpeers_{ijt} + \beta_2 Z_{ijt} \\ &+ \beta_3(sharedelinqpeers_{ijt} \times Z_{ijt}) + \beta_4 X_{ijt} + \lambda_j + \eta_t + \epsilon_{ijt}, \end{aligned} \tag{2.6}$$

where $Z_{ijt}$ is a characteristic of interest such as a youth's gender, placement duration, or placement provider type. We also consider whether the structure of peer influence operates through mean risk or tail risk, for example having one particularly high-risk peer:

$$Y_{ijt} = \alpha + \beta_1 peerscore_{ijt} + \beta_2 maxminscore_{ijt} + \beta_3 X_{ijt} + \lambda_j + \eta_t + \epsilon_{ijt}, \tag{2.7}$$

where $maxminscore_{ijt}$ is the maximum or minimum peer risk score during a placement spell, respectively.

Finally, we check for peer influence that is not linear in means by estimating a flexible nonlinear model of peer effects using an approach similar to Hoxby and Weingarth (2006) and Imberman et al. (2012). We first partition all congregate care placements during the sample period, including those used only for peer characteristics, into four quartiles by level of predicted imprisonment risk. Quartiles are calculated separately for each gender so that an equal number of males and females appear in each quartile. We then estimate four separate regressions for outcomes of youth in each risk quartile on the fraction of peers in each quartile. The nonlinear model is as follows:

$$\begin{aligned} Y_{ijt} = \alpha &+ \beta_1 fraction_Q2_peers_{ijt} + \beta_2 fraction_Q3_peers_{ijt} \\ &+ \beta_3 fraction_Q4_peers_{ijt} + \beta_4 X_{ijt} + \lambda_j + \eta_t + \epsilon_{ijt}, \end{aligned} \tag{2.8}$$

where $fraction_Q2_peers_{ijt}$ is the duration-weighted fraction of youth $i$'s peers in the second quartile of imprisonment risk, and the model is estimated for youth in each quartile of own risk. Both non-delinquent and delinquent youth are included in these regressions to maximize statistical power, and the fraction of peers in the first (lowest) risk quartile is omitted due to collinearity.

The literature on peer effects has identified two primary threats to identification: the reflection problem and selection bias. The reflection problem is a well-known challenge in estimating the causal effect of peers (Manski, 1993). An implication of the problem is that peers' future outcomes cannot be used as a peer risk measure because they may be influenced by the index youth's own latent criminality and vice-versa. To address this problem, we estimate an exogenous peer effects model using two predetermined peer attributes: a peer delinquency measure using peers' reason-for-placement delinquency status, and an additional peer imprisonment risk measure calculated using pre-placement risk factors that are not

influenced by the current placement. As discussed in the Data section above, this risk score is estimated using a LASSO logit machine learning procedure trained on the full sample of placements of children and youth who were in foster care from 2004-2019 and born prior to 2000. Both of these measures have the advantage of relying only on information prior to placement but also using information close to the time of placement that could accurately signal peer behavior or influence in congregate care.

Selection bias may arise if youth are non-randomly grouped with certain types of peers, for example peers with similar levels of imprisonment risk. The empirical strategy therefore focuses on variation in peer characteristics within a provider to control for possible non-random assignment of youth into different types of facilities. An additional form of selection, however, is possible non-random, time-variant grouping of youth within a provider over time. For example, peer effects estimates could be biased upward in magnitude if certain providers selectively accept youth who are similar to the present composition of youth in the facility. Our interviews with several congregate care providers in Wisconsin lead us to believe that provider selection of youth is negligible, conditional on the youth meeting minimal facility criteria (for example, some providers have age or gender restrictions). Although providers do specialize in treating youth with particular needs, nearly all providers interviewed noted it was highly exceptional not to admit a youth based on the composition of other youth in the setting. Furthermore, although placement in congregate care typically involves an interview and visit process, providers often have little control over the exact timing of a youth's entry and exit from a setting, which are constrained by case manager decisions and available beds. Waitlists of several months are common.

We empirically quantify potential non-random grouping within a provider over time by regressing own imprisonment risk on peer risk and controlling for provider, year of entry, and quarter of entry. Each column of Table 2.7.2 presents the results of four separate regressions of own risk on peer risk or peer delinquency, for all youth or maltreated youth, respectively. The first coefficient of column 1 shows that a one standard deviation increase in mean

peer imprisonment risk is associated with a tenth of a percentage point reduction in own risk for the full sample of youth, suggesting that selection bias is not a primary concern. The remainder of column 1 shows that the coefficients on peer risk and delinquency are economically small and insignificant for both the full sample and maltreated-youth-only sample. Columns 2-4 show no strong systematic relationship between own and peer risk across different types of placement settings.[23]

A related concern in the estimation of peer effects is exclusion bias, which arises due to mechanical negative correlation between own characteristics and peer characteristics because a youth cannot be their own peer (Guryan et al., 2009; Angrist, 2014; Caeyers and Fafchamps, 2016). Estimates in panel A of Table 2.7.2 do not show large negative associations for the full sample, and we are also able to break potential negative mechanical correlation by separately observing delinquent and non-delinquent youth in our data, using an approach similar to Carrell et al. (2018). Panel B shows that among exclusively non-delinquent youth there is no relationship between own imprisonment risk and share of delinquent peers. For the main results that follow, we also present supplemental estimates for maltreated (non-delinquent) youth as their own category. Finally, in all models to follow, we control for a youth's own risk score to attenuate unforeseen selection concerns.

## 2.5 Results

To confirm the relevance of our measures of criminality, we first examine the association between a youth's own delinquency and later life outcomes. Table 2.7.3 panel A presents the relationship between individual pre-placement delinquency and outcomes, and panel B presents the relationship between individual imprisonment risk and outcomes. We find that the official delinquency indicator and the machine-learning-constructed risk score are each highly correlated with lower levels of future wellbeing and higher levels of teen imprisonment.

[23]To the extent that there is positive delinquency-related selection in this context, our estimates can be interpreted as upper bounds on the true peer effect coefficients.

Column 2 shows that, conditional on placement controls such as provider and year fixed effects, having a delinquency-related placement is associated with a 4.2 percentage point (37 percent) increase in the likelihood of entering prison before age 20, compared to a sample mean of 11.3 percent teen imprisonment. Similarly, a one unit increase in the teen imprisonment risk score is associated with a 93 percentage point increase in the likelihood of prison before age 20. These associations provide hard evidence that the delinquency placement indicator and the imprisonment risk score capture risk factors for deviant behavior during a placement, and therefore justifies using these variables as peer characteristics of interest.

Figure 2.7.1 provides a visual depiction of the variation in a youth's share of delinquent peers within a provider over time, and a preview of the results estimating the relationship between peer delinquency and own future imprisonment. The histogram in this figure – the residualized distribution of a youth's share of delinquent peers after controlling for provider-by-gender fixed effects, year-of-entry fixed effects, and quarter-of-entry fixed effects – shows that there is considerable variation in policy variable of interest. The solid line, a local linear regression of residualized imprisonment before age 20 on peer imprisonment risk, however, suggests that mean peer delinquency is not strongly related to own imprisonment. Figure 2.A.6 presents similar findings when considering peer imprisonment risk as the right-hand-side variable.

Table 2.7.4 presents estimates of the effect of peer delinquency on a youth's own deviant behavior and teen imprisonment. panel A presents estimates of the effect of delinquent peers and panel B estimates of average peer imprisonment risk. Panels A and B show a surprising zero effect of same-gender peer delinquency and peer risk on outcomes. For non-delinquent youth, panel A column 2 shows that a one standard deviation increase in the fraction of delinquent peers is associated with a 0.43 percentage point reduction in the likelihood of entering prison by age 20, and the estimates rule out increases in prison likelihood of more than 0.05 percentage points at the 95 percent confidence level. Across each panel for both delinquent and non-delinquent youth, point estimates for future prison are economically

small, not statistically different from zero, and often signed away from detrimental outcomes. In addition to the results presented in the table, we find no effect of peer influence on other intermediate outcomes such as future placement for delinquency. One exception to the null findings is a marginally increase in high school dropout per standard deviation increase in peer risk (panel B, column 6). We consider this finding below.

Our primary findings remain for a variety of sample choices. Table 2.A.2 presents peer effects estimates separately for youth who entered the placement due to risk of being maltreated and for youth who entered the placement due to delinquency-related reasons. This choice is motivated in part by Lindquist and Santavirta (2014), who find that the association between foster care and adult crime is much larger for youth placed in care due to their own anti-social behavior. We are unable to detect a statistically significantly different effect of peer delinquency or imprisonment risk between any of the presented groups. Further, the findings are robust to the omission of a youth's own risk score or the inclusion of additional controls such as peer demographics (Table 2.A.3). The findings are also robust to restricting the sample only to males, for whom imprisonment is more likely (Table 2.A.4). The findings apply as well to youth who are in their first observed congregate care placement, who have less of a foster care history and may be particularly sensitive to peers (Table 2.A.5).

Finally, many youth in congregate care in Wisconsin are placed in, or under the custody of, Milwaukee County, the largest urban area in the state. Unfortunately, Milwaukee County does not track delinquent youth placed in foster care settings in the state's child welfare information system (E-WISACWIS). Our analysis sample excludes providers in Milwaukee County; however there may still be occasional delinquent peers under the custody of Milwaukee County who are placed outside the county and missing from the sample.[24] To address this concern, Table 2.A.7 reports coefficients after omitting the four counties bordering Milwaukee County and an additional fifth closest county (Kenosha). The estimates

[24]An early draft of this paper reported positive associations between peer risk and own imprisonment. These findings were the result of incomplete placement records for Milwaukee County, which was corrected when brought to our attention.

are largely unchanged.

The findings presented in Table 2.7.4 and accompanying figures suggest that increasing a youth's share of delinquent or high-risk peers in a foster care setting does not contribute to a youth's immediate deviant behaviors or future criminal justice system involvement. The following section proceeds to understand this finding in greater depth.

### 2.5.1 Exploring Possible Mechanisms of Peer Influence

Although peers in congregate care settings do not seem to exert systematic influence on youth on average in the context of crime, peer effects may still be prevalent for certain kinds of youth or in specific types of settings. Understanding the mechanisms of peer influence is important both for policy prescriptions to increase the welfare of children in foster care and for our broader understanding of human development. We consider several dimensions of mechanisms: heterogeneity in peer effects by placement experiences, personal characteristics, and channels of peer influence.

Providers differ in potentially meaningful ways: residential facilities tend to group a greater number of youth together and average stays in group homes (162 days) and residential facilities (115 days) are longer than in temporary placements (46 days). Hence, one might expect peer effects to be largest in group home and residential settings. Table 2.7.5 examines heterogeneity across different congregate care provider types (group home, residential facility, and temporary) with respect to the effect of peer characteristics on teen imprisonment. We find no clear systematic evidence of significant effects of peer delinquency or risk in any particular placement setting. Point estimates for temporary placements are particularly small, suggesting that peer influence with respect to criminal justice system involvement is negligible in these settings. Table 2.A.8 also presents peer effect coefficients for placements with above- and below-median placement duration providers, with no significant difference in the peer effect coefficient between these settings. The imprecise and inconsistently signed estimates suggest a relative lack of crime peer effects across all types of providers.

Table 2.7.6 presents estimates of peer effect heterogeneity within a provider over time by specific placement-related experiences including number of peers, placement duration in days, average number of days residing with a given peer, and age at the start of placement. Given that child welfare placements can be shorter and include fewer peers compared to other contexts such as classrooms or juvenile jails, it may be that peer effects are strongest in contexts where children have many peers (more precise measurement and less oversight), have longer stays (greater peer exposure), spend more time with a given peer, and are younger (most subject to influence). Table 2.7.6, however, shows that association between peer delinquency and future imprisonment does not become significantly more positive with longer stays (column 1), placements with more peers (column 2), placements with more time spent with a given peer (column 3), or starting age (column 4). These findings suggest that marginal changes to these placement features do not contribute to crime peer effects. The findings are similar when using mean peer risk as the peer characteristic of interest instead of delinquency (Table 2.A.9).

Table 2.7.7 presents demographic heterogeneity in the effect of delinquent peers on teen imprisonment. Several studies, including Carrell and Hoekstra (2010), have found disparities in susceptibility to peer influence by gender, family background, and other youth characteristics. Although less precisely estimated, our results suggest no substantive changes in peer effect coefficients on the basis of gender or own imprisonment risk. The interactions between peer delinquency and being Black (column 1) and between peer delinquency and past parental imprisonment (column 4) are marginally significant, but are not robust to using peer imprisonment risk as the peer effects measure (Table 2.A.10).

Black et al. (2013) find evidence of school peer demographic impacts on long-run outcomes for youth in Norway, in particular that a higher share of female peers advantages females but disadvantages males.[25] Table 2.A.12 presents estimates of the effect of peer race, gender, and age on teen imprisonment. Table 2.A.13 presents similar estimates with

[25]Peer gender composition effects have also been documented with respect to short-run achievement in other classroom settings (for example, Hoxby, 2000; Whitmore, 2005).

heterogeneity by gender. We find no evidence that the share of female congregate care peers or mean peer age influences imprisonment.

Finally, we examine alternative models of peer influence. The structure of peer effects, for example whether a youth is influenced primarily by one disruptive peer or many peers on average, is important for designing policies to improve human capital formation and optimally allocating youth across placement settings. On the whole, prior research indicates substantial heterogeneity in the nature of peer effects and raises questions about the importance of mean peer characteristics versus the characteristics of the most problematic peer (Sacerdote, 2014). Table 2.A.14 considers alternative peer effect specifications: max vs. mean peer delinquency (risk) and own-gender vs. opposite-gender peer delinquency (risk), respectively. Columns 1-2 do not show strong evidence that teen imprisonment depends more on maximum peer delinquency (or risk score) as opposed to the mean measure of peer delinquency or risk. Additionally, columns 4-5 show that opposite-gender peer delinquency and imprisonment risk does not raise a youth's own likelihood of entering prison before age 20. The findings are consistent with the primary peer effects estimates using same-gender mean peer delinquency and risk.

Imberman et al. (2012) show that peer effects estimates can be sensitivity to functional form, and we use a similar approach to estimate a nonlinear model of peer influence. Table 2.A.15 presents the results of five regressions of own imprisonment on the share of peers in each quartile of imprisonment risk. Column 1 uses a pooled sample of all youth, and columns 2-5 present regression estimates for youth in each risk quartile. We are unable to reject that any particular coefficient is significantly different from zero, and we do not observe any clear systematic nonlinear relationship between peer risk and a youth's own outcome in terms of coefficient signs. We conclude there is no evidence of nonlinear peer effects in this context.

### 2.5.2 Additional Outcomes: High School Dropout and Teen Parenthood

An absence of evidence for crime peer effects does not necessarily imply an absence of peer influence along other policy-relevant margins of child development and behavior. We consider two other important non-criminal behaviors that could be subject to peer influence: high school dropout and teen parenthood. Table 2.A.16 describes the effect of peer imprisonment risk on a youth's ever dropping out of public high school, for a sample of youth who were enrolled in public school immediately before placement and had not yet dropped out of school. Although there was no significant association between the share of delinquent peers and dropout (Table 2.7.4 panel A), find that a one standard deviation increase in peer imprisonment risk is associated with a marginally significant 0.94 percentage point (2.5 percent) increase in a youth's likelihood of dropout. The point estimates are positive across all placement types, although only statistically significant for temporary placement settings, and the full sample point estimate's statistical significance is sensitive to the inclusion additional youth-level controls. The association between peer risk and reduced educational attainment is robust to using high school equivalent completion as the outcome variable, which is only available beginning in 2009 (Table 2.A.16).[26]

Given a small but statistically significant association between peer risk and educational attainment, we consider mechanisms through which youth are impacted by congregate care peers. We find no clear heterogeneity in the peer effects coefficient across most of a youth's own placement experiences and peer demographic characteristics, respectively (Tables 2.A.18 and 2.A.19). However, we do find that a youth's average number of peers is associated with a stronger relationship between peer risk and dropout (2.A.18). Furthermore, Table 2.7.8 shows that Black youth and youth who had a parent imprisoned before age 10 are significantly more susceptible to peer influence as it pertains to educational attainment.

[26]High school completion differs slightly from dropout in that some students who drop out may re-enroll or later earn a GED.

This latter finding is consistent with the theory that youth with potentially unstable home environments are most susceptible to risky peers.

Second, we consider the effect of peer composition on teen parenthood. This outcome is estimated only for female youth in our sample due to measurement error for males. Table 2.A.20 presents estimates of the effect of various peer characteristics on teen parenthood. We do not find strong evidence of peer influence across the observed measures. Considering the risk measure with the largest t-statistic (average male peer imprisonment risk) more closely, the association between male peer imprisonment risk and teen parenthood is not significant different from zero for female youth across placement settings (Table 2.A.21) and appears relatively invariant to a number of potential risk factors (Table 2.A.22). These findings suggest that female youth's teen parenthood may not be significantly affected, on average, by a youth's composition of peers in congregate care along select observable dimensions.

### 2.5.3 Additional Discussion

Given a strong positive association between congregate care placement and criminality, our findings motivate further investigation about which non-peer aspects of congregate care *do* influence youth's outcomes. Figure 2.A.7 plots the distribution of provider fixed effect coefficients with respect to youth imprisonment, with and without demographic controls. The distribution of the gray histogram shows considerable variation in unconditional youth imprisonment by provider, which can be attributed to small sample variation in imprisonment risk or if certain providers cater to high-risk youth. The blue presents variation in provider association with teen prison after controlling for youth imprisonment risk and other demographic characteristics. Although the distribution of provider fixed effects narrows after controlling for observable youth risk factors, there remains considerable variation in youth outcomes by provider. This distribution of outcomes is suggestive of variable provider quality and serves to motivate ongoing and future work on the effects of congregate care placement, and provider amenities, on youth outcomes (Taylor, 2021).

Our study has several limitations that affect interpretation of the results. First, the results use complete data on state prisons but do not include jails, which account for most experiences of incarceration, or federal prisons (five percent of the incarcerated population in Wisconsin; Jones, 2018). Thus, imprisonment captures those who have committed comparatively severe crimes and are serving comparably long sentences. On the other hand, this restriction may be advantageous if state prison sentencing is less affected by justice system biases that can lead to incarceration for relatively minor crimes.[27] Second, we have no information on the underlying criminal offenses that led to imprisonment, which could provide further insight into the nature of, or heterogeneity in, peer effects. Third, peer effects in congregate care may manifest themselves in ways that are not captured by the outcomes in this paper. To the extent possible, however, we have examined a variety of short- and medium-term outcomes that measure extensive and intensive margins of antisocial behavior and potential criminality.

Lastly, our sample is comprised of youth who experienced a foster care placement relatively late in childhood. Relative to the foster care population as a whole, youth in our sample experienced high rates of congregate care placement and a substantial fraction entered foster care for non-maltreatment-related reasons. As such, our estimates of prison entry and other outcomes must be interpreted within these sample parameters and should not be generalized to the entire foster care population. The results, however, are highly relevant for those youth who are most at risk of incarceration. This group has long been viewed as a particularly important, vulnerable, and policy-relevant subset of the foster care population.

Lastly, we account for stable characteristics of individual placement facilities, but it is possible that providers alter the amount or supervision of peer interactions based on the composition of risk among currently placed youth to ensure safety and mitigate disruptive

[27] Although individual incarceration need not imply criminal activity or vice-versa, conditional on provider and time of entry we assert that prison sentences are a reasonable proxy for severe criminal activity on average.

behavior. If so, this complicates our conceptualization of "exposure" to high risk peers, whereby the level of exposure (contact, and especially less structured contact) is inversely related to peer risk composition. Thus, we cannot fully discern whether the absence of peer effects in this context is invariant to staffing levels and monitoring, or mitigated by risk-dependent reactions with facilities.

## 2.6 Conclusion

Foster care is an intrusive and intensive intervention for children who face extreme and complex risks in their familial environments. Children who experience foster care in late childhood and adolescence have substantial histories of trauma and adversity and thus are positioned for poor life outcomes in the absence of significant investments in compensatory support (Font and Gershoff, 2020). Yet, these same youth often have the least supportive experiences in foster care – they experience high levels of congregate care and placement instability, have limited opportunities for adoption, and are more likely to age out of care (U.S. Children's Bureau, 2015; Konijin et al., 2019; Snowden, Leon, and Sieracki, 2008). An array of policy changes has sought to improve outcomes for older youth, including unsuccessful efforts to increase adoption rates (Brehm, 2018) and potentially more successful efforts to support youth positioned to age out of care (Lee et al., 2014; Courtney and Hook, 2017). Most recently, reductions in the use of congregate care and changes in the nature of remaining congregate care placements are anticipated as states implement the Family First Prevention Services Act (FFPSA). Yet, for older youth in foster care, there remains little evidence about whether, how, and for whom foster care has adverse effects. Few studies have rigorously estimated the effect of child welfare amenities on child outcomes, and few have studied the role of children's residential peers, in foster care or otherwise, in shaping young adolescent behavior.

In this paper, we study the effect of deviant peers on outcomes for youth in congregate

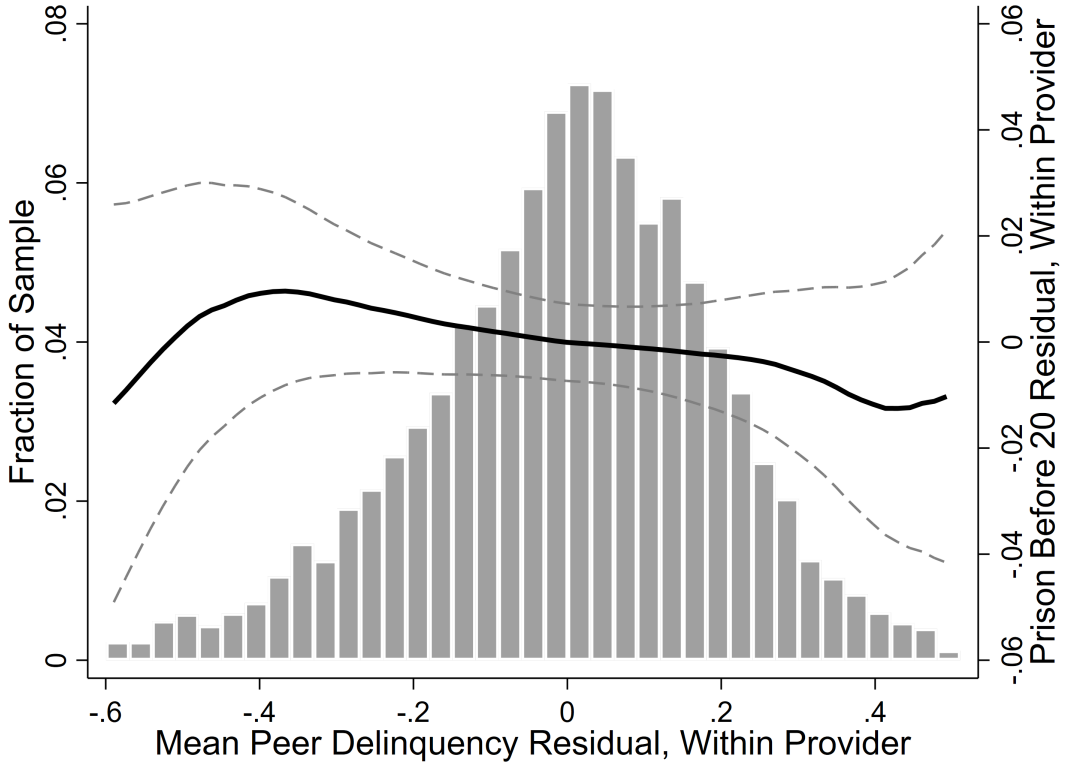
care using natural variation in the share of delinquent peers in foster care group-based settings. We supplement this peer effects measure with a highly predictive imprisonment risk score using multiple decades of matched family administrative data from Wisconsin. Our results reveal no evidence of negative peer effects in foster care with respect to short-term antisocial behaviors and future imprisonment, both for youth placed due to maltreatment and youth placed for delinquency-related reasons. These results are robust to placement characteristics, child demographics, and alternative models of peer influence. These findings may be considered somewhat surprising in light of evidence of peer influence on recidivism in juvenile jails (Bayer at al., 2009; Stevenson, 2017) and prisons (Drago and Galbiati, 2012). Given that youth often move between several residential settings, it is possible that the nature of interactions in congregate care are transient enough that, compared to other types of environments, peer composition does not have an appreciable influence on youth outcomes. We present supplemental evidence that peer imprisonment risk increases high school dropout, but the estimated magnitudes are small (2.5 percent increase in dropout per one standard deviation increase in peer risk), only marginally statistically significant, and not robust across all peer measures.

To the extent that our findings generalize to youth in family homes and to the extent that foster parents are adequately prepared, our results suggest that policymakers may not need to be as concerned about placing higher-risk youth in family-like settings with lower-risk youth under FFPSA reforms that transition away from the use of congregate care. Furthermore, our results suggest that policymakers need not be concerned that children who are still placed in congregate care for temporary stays will be adversely influenced by the higher-risk, more selected group of peers who still require congregate care after the reform. Although this paper does not directly estimate the impact of congregate care on youth outcomes compared to alternative options such as family homes, the results do imply that if congregate care were to negatively affect youth generally, it would be due to non-peer features of the institutional environment that perhaps account for significant variation in

outcomes by provider documented in Figure 2.A.7. Our findings therefore highlight the need for continued research on the impact of residential providers on children's development, a particularly urgent domain of inquiry given that approximately one in eight Black and Native American children in the United States are placed in foster care before age 18 (Wildeman and Emanuel, 2014). FFPSA-imposed changes to the child welfare system, including the transition from a more generic model of congregate care to the qualified residential treatment model, will provide new opportunities to assess the impact of foster care features on children's wellbeing.

## 2.7 Figures and Tables

Figure 2.7.1: Distribution of Peer Delinquency and Relationship with Own Imprisonment




Notes: This figure shows the distribution of delinquent peers and the relationship between peer delinquency and own teen imprisonment. The histogram shows the residualized distribution of a youth's share of delinquent peers for youth in the analysis sample after controlling for provider-by-gender fixed effects, year-of-entry fixed effects, and quarter-of-entry fixed effects, with the density on the left axis. The solid line is a local linear regression of residualized prison before age 20 on peer imprisonment risk, with the dashed lines representing the upper and lower bounds of a 95 percent confidence interval. Standard errors are two-way clustered by provider and youth.

Table 2.7.1: Characteristics of Foster Care Youth by Provider Type

| | (1) Group Home | (2) Residential Facility | (3) Temporary |
|---|---|---|---|
| *Outcomes* | | | |
| Prison Before Age 20 | 0.11 | 0.12 | 0.11 |
| Months in Prison Before Age 20 | 8.84 | 9.15 | 7.98 |
| Placement Exit Detained | 0.096 | 0.048 | 0.071 |
| Placement Exit Runaway | 0.13 | 0.027 | 0.040 |
| High School Dropout | 0.40 | 0.36 | 0.37 |
| Wellbeing Index | -0.16 | 0.15 | 0.094 |
| Parent Before Age 20 (Female) | 0.43 | 0.36 | 0.40 |
| *Key Regressors* | | | |
| Own Risk Score | 0.12 | 0.11 | 0.10 |
| Peer Risk Score | 0.11 | 0.11 | 0.099 |
| Standardized Peer Risk | 0.096 | 0.11 | -0.23 |
| Maltreatment (Non-Delinquent) Placement | 0.26 | 0.26 | 0.33 |
| Own Delinquency | 0.74 | 0.74 | 0.67 |
| Peer Delinquency | 0.67 | 0.63 | 0.60 |
| Standardized Peer Delinquency | 0.10 | -0.046 | -0.10 |
| *Demographics* | | | |
| Female | 0.31 | 0.31 | 0.35 |
| Black | 0.20 | 0.15 | 0.20 |
| Age at First FC Placement | 13.3 | 12.9 | 13.2 |
| Lifetime Placements | 7.65 | 7.48 | 7.71 |
| *Family Background* | | | |
| Mother Share Months in Prison Before 10 | 0.0023 | 0.0022 | 0.0034 |
| Father Share Months in Prison Before 10 | 0.015 | 0.012 | 0.015 |
| Marital Birth (if known) | 0.31 | 0.35 | 0.32 |
| Mother N Kids | 3.80 | 3.55 | 3.70 |
| Mother Age First Birth | 20.6 | 21.0 | 20.7 |
| SNAP Before Placement | 0.87 | 0.83 | 0.86 |
| SSI Before Placement | 0.17 | 0.23 | 0.18 |
| *Placement* | | | |
| Length of Stay (Days) | 161.7 | 114.6 | 45.5 |
| Placement Avg N of Peers | 4.46 | 13.9 | 4.70 |
| Avg Days with a Given Same-Gender Peer | 69.5 | 54.6 | 18.6 |
| Girls-Specific Home | 0.20 | 0.027 | 0 |
| Boys-Specific Home | 0.57 | 0.35 | 0 |
| Age at Placement Start | 15.2 | 14.6 | 14.8 |
| N Past Referrals | 1.82 | 1.85 | 1.96 |
| N Providers | 89 | 58 | 50 |
| N Unique Kids | 2,771 | 1,777 | 2,046 |
| N Observations | 3,658 | 2,271 | 2,603 |

Table 2.7.2: Association between Peer Imprisonment Risk and Own Imprisonment Risk

| | All | Group Home | Residential Facility | Temporary |
|---|---|---|---|---|
| | Own Risk Score | Own Risk Score | Own Risk Score | Own Risk Score |
| *Panel A: All Youth* | | | | |
| Standardized Peer Risk | -0.0012 | -0.0021* | -0.0021 | -0.0004 |
| | (0.0007) | (0.0011) | (0.0013) | (0.0013) |
| Standardized Peer Delinquency | 0.0010 | -0.0004 | 0.0025 | 0.0022** |
| | (0.0006) | (0.0009) | (0.0015) | (0.0008) |
| Placement Controls | × | × | × | × |
| Mean of Dep Var | 0.1132 | 0.1184 | 0.1146 | 0.1045 |
| Observations | 8,474 | 3,643 | 2,248 | 2,583 |
| *Panel B: Maltreated (Non-Delinquent) Youth* | | | | |
| Standardized Peer Risk | -0.0014 | -0.0010 | -0.0038 | -0.0020 |
| | (0.0010) | (0.0014) | (0.0025) | (0.0016) |
| Standardized Peer Delinquency | -0.0007 | -0.0012 | 0.0010 | -0.0004 |
| | (0.0008) | (0.0018) | (0.0011) | (0.0012) |
| Placement Controls | × | × | × | × |
| Mean of Dep Var | 0.0738 | 0.0733 | 0.0769 | 0.0722 |
| Observations | 2,332 | 920 | 578 | 834 |

Notes: Each column in this table presents the results of four separate regressions of own predicted imprisonment risk on a standardized measure of peer imprisonment risk or peer delinquency, for all youth and maltreated youth, respectively. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, and quarter-of-entry fixed effects. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.3: Association between a Youth's Own Outcomes and Own Delinquency or Risk

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Delinquency* | | | | | | | |
| Own Delinquency | -0.1466*** | 0.0424*** | 2.1902*** | 0.0219*** | 0.0117 | 0.0239 | 0.1207*** |
| | (0.0326) | (0.0110) | (0.6823) | (0.0074) | (0.0082) | (0.0212) | (0.0253) |
| *Panel B: Prison Risk Score* | | | | | | | |
| Own Risk Score | -3.1837*** | 0.9347*** | 75.8833*** | 0.2609*** | 0.3967*** | 0.6082*** | 3.6097** |
| | (0.2846) | (0.1358) | (9.6323) | (0.0707) | (0.0779) | (0.1941) | (1.7554) |
| Placement Controls | × | × | × | × | × | × | × |
| Mean of Dep Var | -0.0022 | 0.1134 | 8.6939 | 0.0761 | 0.0769 | 0.3795 | 0.3584 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 | 8,474 | 6,785 | 2,522 |

Notes: This table presents the relationship between own delinquency (panel A) or own imprisonment risk (panel B) and future outcomes. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.4: Delinquent Peers and Youth Outcomes

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | | |
| Standardized Peer Delinquency | 0.0105 | -0.0043 | 0.1177 | 0.0001 | 0.0005 | -0.0046 | 0.0017 |
| | (0.0151) | (0.0048) | (0.2453) | (0.0039) | (0.0038) | (0.0062) | (0.0118) |
| *Panel B: Peer Risk* | | | | | | | |
| Standardized Peer Risk | -0.0114 | 0.0006 | 0.1792 | 0.0021 | -0.0019 | 0.0094* | 0.0071 |
| | (0.0106) | (0.0041) | (0.2575) | (0.0034) | (0.0036) | (0.0057) | (0.0115) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | -0.0022 | 0.1134 | 8.6939 | 0.0761 | 0.0769 | 0.3795 | 0.3584 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 | 8,474 | 6,785 | 2,522 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and youth outcomes, where each peer effect coefficient is estimated using a pooled sample of youth placed in foster care due to maltreatment and youth placed in foster care due to own delinquent behavior. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.5: Delinquent Peers and Own Imprisonment
Heterogeneity by Placement Type

| | Group Home | Residential Facility | Temporary |
|---|---|---|---|
| | Prison Before Age 20 | Prison Before Age 20 | Prison Before Age 20 |
| *Panel A: Peer Delinquency* | | | |
| Standardized Peer Delinquency | -0.0081 | 0.0076 | -0.0053 |
| | (0.0069) | (0.0146) | (0.0060) |
| *Panel B: Peer Risk* | | | |
| Standardized Peer Risk | -0.0025 | 0.0101 | -0.0012 |
| | (0.0059) | (0.0111) | (0.0063) |
| Placement Controls | × | × | × |
| Own Risk Score | × | × | × |
| Mean of Dep Var | 0.1150 | 0.1192 | 0.1061 |
| Observations | 3,643 | 2,248 | 2,583 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and teen imprisonment across different types of foster care providers. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.6: Delinquent Peers and Own Imprisonment
Heterogeneity by Child Welfare Experiences

| | Prison Before Age 20 | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Peer Delinquency | -0.0036 | -0.0052 | -0.0002 | 0.0451 |
| | (0.0054) | (0.0062) | (0.0059) | (0.0310) |
| Length of Stay in Months × Standardized Peer Delinquency | -0.0002 | | | |
| | (0.0009) | | | |
| Placement Avg N of Peers, in Tens × Standardized Peer Delinquency | | 0.0026 | | |
| | | (0.0173) | | |
| N Months with Given Peer × Standardized Peer Delinquency | | | -0.0035 | |
| | | | (0.0028) | |
| Age at Placement Start × Standardized Peer Delinquency | | | | -0.0032 |
| | | | | (0.0020) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 |

Notes: This table presents heterogeneity in the estimated relationship between a youth's standardized share of delinquent peers and teen imprisonment by characteristics of the youth's provider and foster care experiences. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.7: Delinquent Peers and Own Imprisonment
Heterogeneity by Youth Background Characteristics

| | Prison Before Age 20 | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Peer Delinquency | -0.0086 | -0.0039 | -0.0050 | -0.0068 |
| | (0.0058) | (0.0065) | (0.0059) | (0.0052) |
| Black=1 × Standardized Peer Delinquency | 0.0204 | | | |
| | (0.0127) | | | |
| Female=1 × Standardized Peer Delinquency | | -0.0012 | | |
| | | (0.0069) | | |
| Own Risk Score × Standardized Peer Delinquency | | | 0.0056 | |
| | | | (0.0578) | |
| Parent Prison in Childhood=1 × Standardized Peer Delinquency | | | | 0.0146 |
| | | | | (0.0098) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 |

Notes: This table presents heterogeneity in the estimated relationship between a youth's standardized share of delinquent peers and teen imprisonment by characteristics of the youth's personal background prior to placement. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.7.8: Peer Imprisonment Risk and High School Dropout Heterogeneity by Youth Background Characteristics

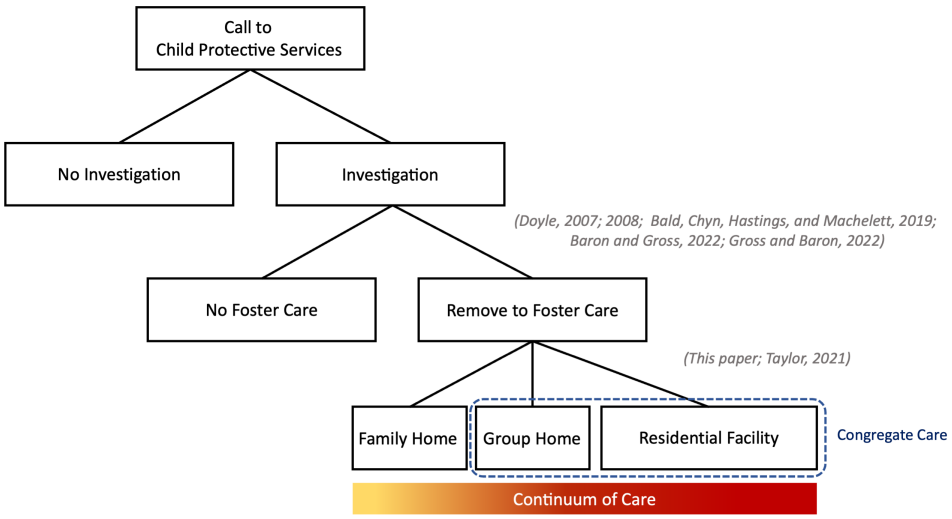
| | High School Dropout | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Peer Risk | 0.0018 | 0.0010 | 0.0174 | 0.0030 |
| | (0.0068) | (0.0086) | (0.0110) | (0.0068) |
| Black=1 × Standardized Peer Risk | 0.0303** | | | |
| | (0.0148) | | | |
| Female=1 × Standardized Peer Risk | | 0.0265 | | |
| | | (0.0168) | | |
| Own Risk Score × Standardized Peer Risk | | | -0.0702 | |
| | | | (0.0998) | |
| Parent Prison in Childhood=1 × Standardized Peer Risk | | | | 0.0334* |
| | | | | (0.0175) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.3795 | 0.3795 | 0.3795 | 0.3795 |
| Observations | 6,785 | 6,785 | 6,785 | 6,785 |

Notes: This table presents heterogeneity in the estimated relationship between standardized peer mean imprisonment risk and own high school equivalent completion by characteristics of the youth's personal background prior to placement. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

# Appendix

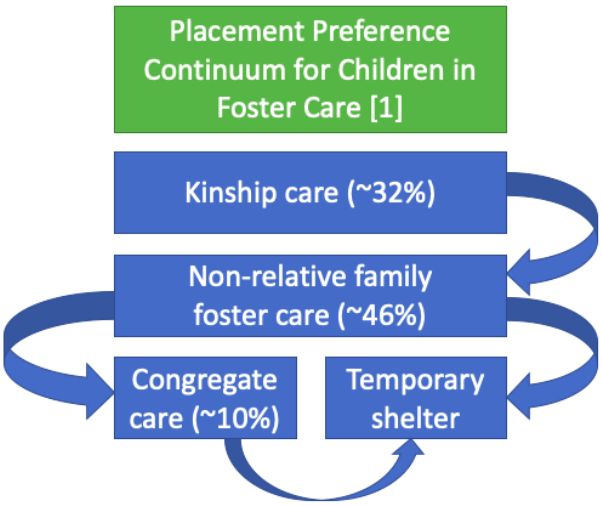
## 2.A Appendix Figures and Tables

Figure 2.A.1: Child Welfare Context and Relation to Literature



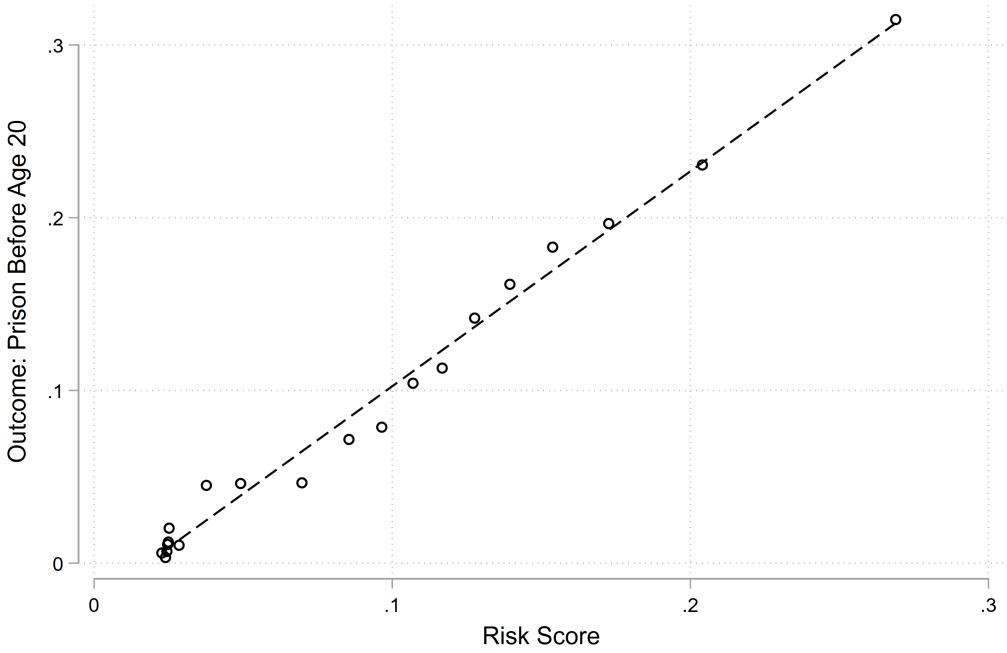

Notes: This diagram presents a simplified overview of key decision points in the child welfare system.

Figure 2.A.2: Continuum of Care



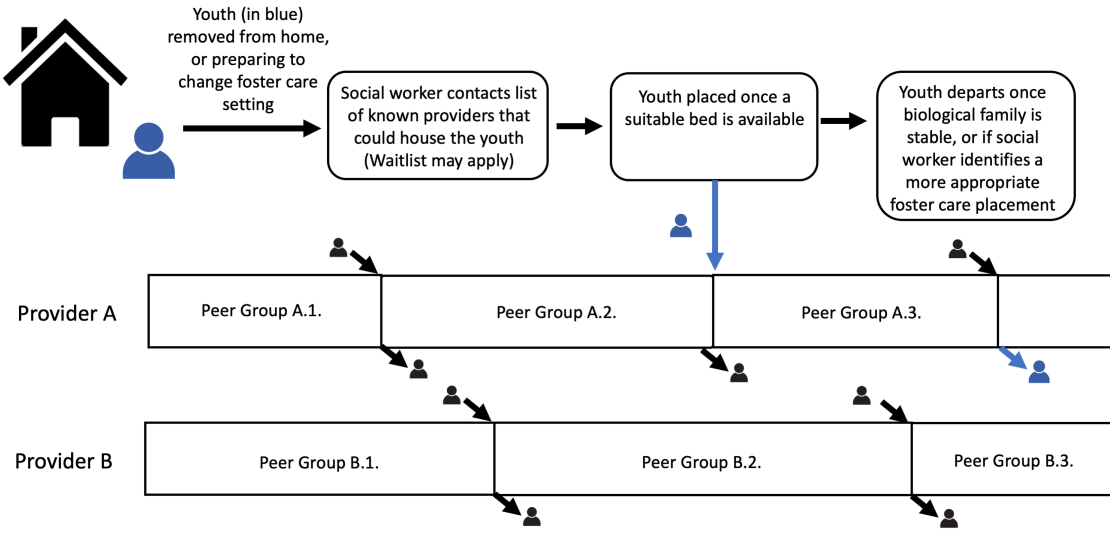

Notes: This figure presents an stylized illustration of the continuum of care in foster care. Children removed from their parents by child protective services are often first placed with a kin caregiver, if possible and appropriate. U.S. Department of Health and Human Services. (2019). *The AFCARS Report: Preliminary FY1 2018 Estimates as of August 22, 2019* (No. 26) . https://www.acf.hhs.gov/sites/default/files/cb/afcarsreport26.pdf; National Conference of State Legislatures. (2016). *The Child Welfare Placement Continuum: What's Best for Children?* http://www.ncsl.org/research/human-services/the-child-welfare-placement-continuum-what-s-best-for-children.aspx

Figure 2.A.3: Risk Score Accuracy, All Kids




Notes: This figure shows a binned scatter plot of the accuracy of the prison before age 20 risk score in predicting actual youth imprisonment before age 20. The risk score is estimated using a LASSO logit procedure on a sample of all youth placements in foster care.

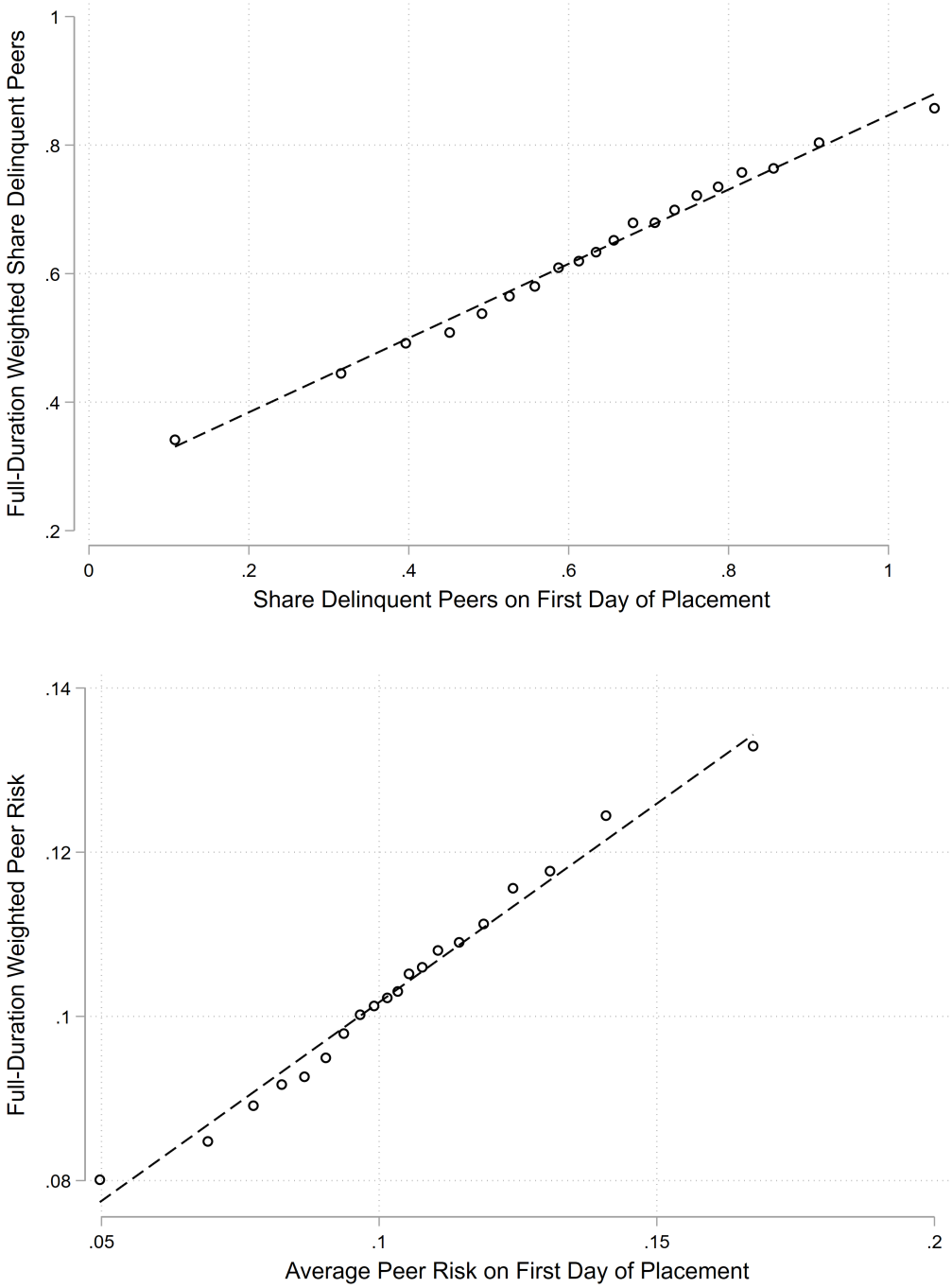
Figure 2.A.4: Assignment Process




Notes: This diagram presents a simplified overview of the assignment process of a child to a foster home. The

analysis in this paper relies on variation in peer characteristics within a provider over time, which change with each new placement and/or departure.

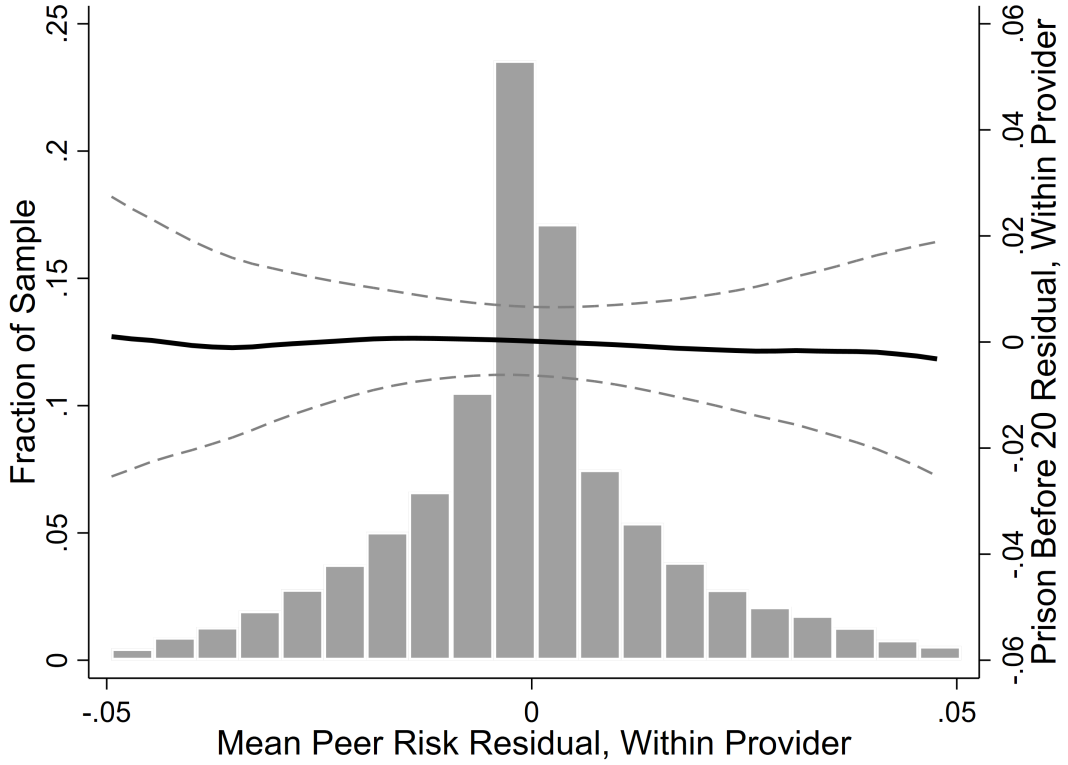
Figure 2.A.5: Peer Measures: Entire Placement vs. Beginning of Placement




Notes: This figure compares the similarity between a first-day-of-placement peer measure and an entire placement (duration-weighted) peer measure using a binned scatter plot. The top panel presents the rela-

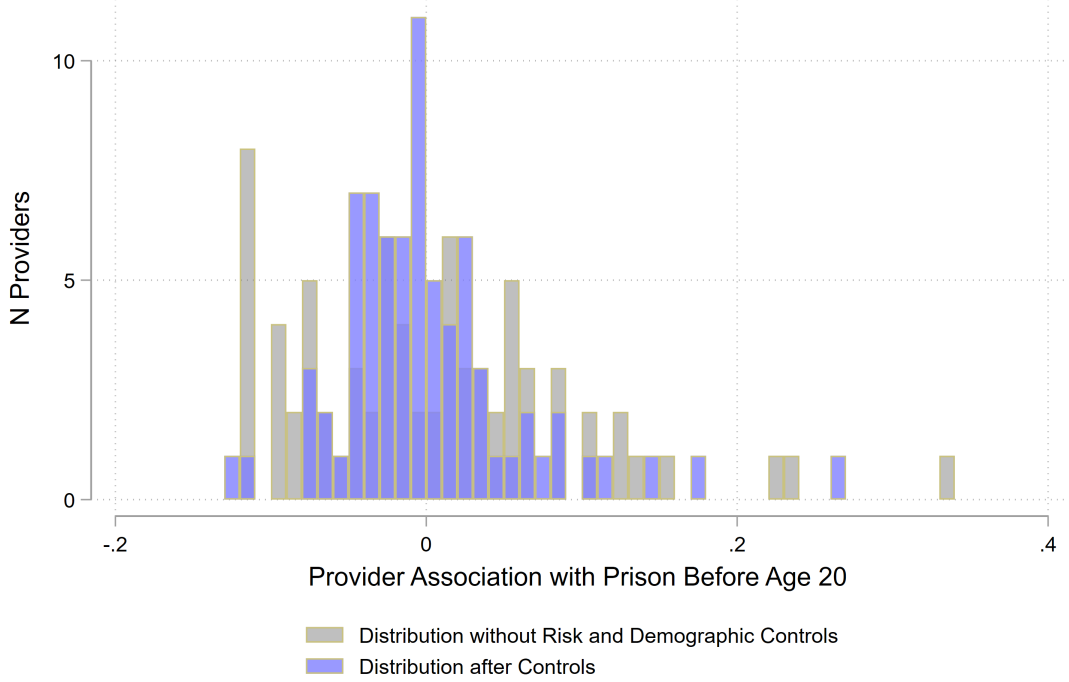
tionship for the share of delinquent peers. The bottom panel presents the relationship for the mean peer risk score. The measures are similar on average and we use the duration-weighted measure in the primary specifications in this paper.

Figure 2.A.6: Distribution of Peer Imprisonment Risk and Relationship with Own Imprisonment




Notes: This figure shows the distribution of peer imprisonment risk and the relationship between peer risk and own teen imprisonment. The histogram shows the residualized distribution of peer risk for youth in the analysis sample after controlling for provider-by-gender fixed effects, year-of-entry fixed effects, and quarter-of-entry fixed effects, with the density on the left axis. The solid line is a local linear regression of residualized prison before age 20 on peer imprisonment risk, with the dashed lines representing the upper and lower bounds of a 95 percent confidence interval. Standard errors are two-way clustered by provider and youth.

Figure 2.A.7: Distribution of Youth Imprisonment Outcomes by Provider




Notes: This figure presents the distribution of mean teen imprisonment rates for youth by congregate care provider relative to an omitted provider, with distributions before and after controlling for youth demographics and risk factors.

Table 2.A.1: LASSO Coefficients of Standardized Variables
Sorted by Penalty Importance

| Feature | LASSO Penalty Coefficient | Postselection Coefficient (Odds Ratio) |
|---|---|---|
| Constant Term | -2.632 | 0.059 |
| Female × No Prior Detention Facility Placement | -0.27 | 0.564 |
| Male × At Least One Prior Detention Facility Placement | 0.141 | 1.003 |
| Male × Mother Teen at First Birth | 0.139 | 1.27 |
| Male × Own Delinquency Placement | 0.129 | 1.333 |
| Female × No SSI Before Placement | -0.108 | 0.487 |
| Male × Parents Not Married at Birth | 0.096 | 1.234 |
| Male × Fraction of Months Mother Welfare Receipt Before Age 10 | 0.082 | 1.83 |
| Male × No Prior Allegations for "Other" Abuse Victim | 0.061 | 1.312 |
| Male × At Least One Prior Temporary Placement | 0.054 | 0.946 |
| At Least One Prior Residential Facility Placement× Fraction of Months Mother SNAP Receipt Before Age 10 | 0.043 | 1.40 |
| Male × Fraction of Months Mother SNAP Receipt Before Age 10 | 0.041 | 1.144 |
| Male × Fraction of Months Mother in Prison Before Age 10 | 0.032 | 2796.707 |
| No Prior Residential Facility Placement× Previous N Days in Foster Care | -0.025 | 1.00 |
| Male × Three or More Running Foster Care Placements in Lifetime | 0.025 | 1.238 |
| Male × At Least One Prior Residential Facility Placement | 0.023 | 0.964 |
| Male × Age at Start of Placement | 0.022 | 0.994 |
| No Prior Referrals to CPS× Five Fathers Associated with Mom | 0.019 | 2.863 |
| Male × SNAP Receipt Before Placement | 0.018 | 1.06 |
| Male × Not First Foster Care Placement | 0.018 | 1.162 |
| Male × No Prior Allegations for Emotional Damage Abuse Victim | 0.016 | 1.236 |
| Own Delinquency Placement × 5 Fathers Associated with Mom | 0.015 | 1.646 |
| No Prior Detention Facility Placement × Previous N Days in Foster Care | -0.013 | 1.00 |
| No Prior Group Home Placement× Fraction of Months Father Employed Before Age 10 | -0.01 | 0.689 |
| SSI Before Placement × 5 Fathers Associated with Mom | 0.007 | 1.877 |
| Mother Teen at Index Child's Birth× Mother N Kids | 0.005 | 1.042 |
| Maltreatment (Non-Delinquent) Placement × No Prior Residential Facility Placement | -0.004 | 0.92 |
| No Running Sex Abuse Allegations × Seven Fathers Associated with Mom | 0.003 | 3.709 |
| At Least Two Running Supervision Neglect Victim Allegations × Seven Fathers Associated with Mom | 0.001 | 9.941e+09 |
| At Least Two Running Physical Abuse Victim Allegations × Three Fathers Associated with Mom | 0.001 | 1.727 |
| At Least One Prior Detention Facility Placement × At Least One Prior Temporary Placement | 0 | 1.389 |
| Maltreatment (Non-Delinquent) Placement × No Prior Temporary Placement | 0 | 0.747 |

Notes: Table includes all regressors selected by LASSO logit model for predicting prison before age 20. Observations used for prediction are at the youth-placement level.

Table 2.A.2: Delinquent Peers and Youth Outcomes: Heterogeneity by Own Delinquency

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | | |
| Standardized Peer Delinquency | 0.0081 | 0.0008 | 0.4799 | 0.0031 | -0.0050 | -0.0078 | -0.0073 |
| | (0.0205) | (0.0062) | (0.3630) | (0.0055) | (0.0055) | (0.0122) | (0.0154) |
| Own Delinquency=1 × Standardized Peer Delinquency | 0.0055 | -0.0081 | -0.5482 | -0.0049 | 0.0080 | 0.0044 | 0.0145 |
| | (0.0255) | (0.0075) | (0.3915) | (0.0069) | (0.0079) | (0.0154) | (0.0178) |
| *Panel B: Peer Risk* | | | | | | | |
| Standardized Peer Risk | -0.0001 | 0.0002 | 0.5894** | 0.0031 | -0.0095* | 0.0076 | 0.0060 |
| | (0.0161) | (0.0050) | (0.2925) | (0.0053) | (0.0053) | (0.0111) | (0.0148) |
| Own Delinquency=1 × Standardized Peer Risk | -0.0158 | 0.0005 | -0.5972 | -0.0015 | 0.0109* | 0.0025 | 0.0032 |
| | (0.0216) | (0.0061) | (0.4158) | (0.0060) | (0.0058) | (0.0142) | (0.0190) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | -0.0022 | 0.1134 | 8.6939 | 0.0761 | 0.0769 | 0.3795 | 0.3584 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 | 8,474 | 6,785 | 2,522 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and youth outcomes. For each regression, the base term is the peer effect coefficient for youth placed in foster care due to maltreatment, and the interaction term is the difference in the peer effect coefficient for youth placed in foster care due to own delinquent behavior. Own delinquency is controlled for but not displayed due to correlation with the risk score control. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.3: Delinquent Peers and Own Imprisonment, Robustness to Controls

| | Prison Before Age 20 | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Standardized Peer Delinquency | -0.0034 | -0.0043 | -0.0040 | -0.0044 | -0.0018 |
| | (0.0049) | (0.0048) | (0.0047) | (0.0047) | (0.0060) |
| Placement Controls | × | × | × | × | × |
| Own Risk Score | | × | × | × | × |
| Demographic Controls | | | × | × | |
| Child History Controls | | | × | × | |
| Family Controls | | | × | × | |
| Peer Demographic Controls | | | | × | |
| Provider-by-Gender-Year Fixed Effects | | | | | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.1134 | 0.1143 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 | 8,294 |

Notes: This table presents the robustness of estimates of the relationship between a youth's standardized share of delinquent peers and teen imprisonment to different sets of controls. Column 1 includes controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, and average number of non-sibling peers during a placement. Column 2 additionally controls for a youth's own imprisonment risk score. Column 3 includes additional controls for youth demographic characteristics (race-ethnicity, gender, age, birth year, and SSI and SNAP receipt before placement), child welfare history (previous number of days in foster care, prior number of referrals to child protective services, prior number of foster care placements, and age at first placement in foster care), and family background (non-marital birth, teen birth, mother's first birth a teen birth, number of fathers associated with the youth's mother, number of children ever born to the youth's mother, and mother's mean time in prison before the youth turned 10). Column 4 includes additional controls for the share of female and Black peers, respectively, and the mean age of peers at the start of their placement. Finally, column 5 controls for provider-by-gender-by-year fixed effects for additional granularity. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.4: Delinquent Peers and Youth Outcomes, Males Only

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout |
|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | |
| Standardized Peer Delinquency | -0.0021 | -0.0038 | 0.4470 | 0.0057 | -0.0010 | -0.0006 |
| | (0.0174) | (0.0066) | (0.3462) | (0.0037) | (0.0039) | (0.0082) |
| *Panel B: Peer Risk* | | | | | | |
| Standardized Peer Risk | -0.0086 | -0.0016 | 0.1265 | 0.0049 | -0.0002 | 0.0014 |
| | (0.0139) | (0.0056) | (0.3543) | (0.0037) | (0.0039) | (0.0085) |
| Placement Controls | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × |
| Mean of Dep Var | -0.0710 | 0.1589 | 11.9019 | 0.0815 | 0.0712 | 0.3789 |
| Observations | 5,770 | 5,770 | 5,770 | 5,770 | 5,770 | 4,650 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and youth outcomes for a subsample of males only. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.5: Delinquent Peers and Youth Outcomes, First Congregate Care Placement Only

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | | |
| Standardized Peer Delinquency | -0.0017 | -0.0042 | -0.1916 | -0.0041 | 0.0101** | -0.0022 | -0.0205 |
| | (0.0230) | (0.0077) | (0.3700) | (0.0060) | (0.0050) | (0.0131) | (0.0239) |
| *Panel B: Peer Risk* | | | | | | | |
| Standardized Peer Risk | 0.0368* | -0.0036 | -0.2170 | 0.0006 | -0.0136** | -0.0075 | 0.0302 |
| | (0.0187) | (0.0082) | (0.5183) | (0.0049) | (0.0054) | (0.0118) | (0.0188) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | 0.1031 | 0.0961 | 7.2759 | 0.0685 | 0.0517 | 0.3596 | 0.3520 |
| Observations | 2,863 | 2,863 | 2,863 | 2,863 | 2,863 | 2,294 | 804 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and youth outcomes for a subsample of youth who are in their first congregate care placement. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.6: Delinquent Peers and Youth Outcomes, Counting Peers of All Genders

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | | |
| Standardized Peer Delinquency | 0.0230 | -0.0069 | -0.0734 | -0.0009 | 0.0017 | -0.0153* | -0.0001 |
| | (0.0171) | (0.0051) | (0.2793) | (0.0050) | (0.0047) | (0.0078) | (0.0154) |
| *Panel B: Peer Risk* | | | | | | | |
| Standardized Peer Risk | -0.0086 | -0.0014 | -0.2419 | 0.0023 | 0.0055 | -0.0046 | 0.0348 |
| | (0.0223) | (0.0084) | (0.5146) | (0.0059) | (0.0061) | (0.0124) | (0.0251) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | -0.0022 | 0.1134 | 8.6939 | 0.0761 | 0.0769 | 0.3795 | 0.3584 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 | 8,474 | 6,785 | 2,522 |

Notes: This table presents the relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and youth outcomes, where the peer measure is calculated using peers of all genders instead of only peers of the same gender. The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.7: Delinquent Peers and Youth Outcomes, Omitting Counties Near Milwaukee

| | (1) Wellbeing Index | (2) Prison Before Age 20 | (3) Months in Prison Before Age 20 | (4) Placement Exit Detained | (5) Placement Exit Runaway | (6) High School Dropout | (7) Parent Before Age 20 |
|---|---|---|---|---|---|---|---|
| *Panel A: Peer Delinquency* | | | | | | | |
| Standardized Peer Delinquency | 0.0091 | -0.0041 | 0.0597 | -0.0007 | 0.0024 | -0.0055 | 0.0147 |
| | (0.0171) | (0.0055) | (0.2612) | (0.0043) | (0.0042) | (0.0071) | (0.0110) |
| *Panel B: Peer Risk* | | | | | | | |
| Standardized Peer Risk | -0.0137 | 0.0010 | 0.1340 | 0.0026 | -0.0023 | 0.0117* | 0.0067 |
| | (0.0120) | (0.0045) | (0.2760) | (0.0035) | (0.0041) | (0.0064) | (0.0122) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | 0.0070 | 0.1064 | 8.1689 | 0.0726 | 0.0791 | 0.3840 | 0.3585 |
| Observations | 7,384 | 7,384 | 7,384 | 7,384 | 7,384 | 5,917 | 2,304 |

Notes: This table provides a robustness check for potential missing delinquency-related placements under the custody of Milwaukee County by omitting four counties bordering Milwaukee County plus the next closest county (Kenosha). The wellbeing index is a standardized summation of standardized components: prison before age 20, placement ending in detention, placement ending in runaway, and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. Teen parenthood is calculated for female youth only and not included in the aggregate index. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score.

Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.8: Heterogeneity by Provider Median Placement Duration

| | (1) Prison Before Age 20 All Youth | (2) Prison Before Age 20 Maltreated Youth |
|---|---|---|
| *Panel A: Peer Delinquency* | | |
| Standardized Peer Delinquency | 0.0009 | -0.0011 |
| | (0.0056) | (0.0065) |
| Provider Above Median Placement Duration=1 × Standardized Peer Delinquency | -0.0148 | -0.0230 |
| | (0.0095) | (0.0144) |
| *Panel B: Peer Risk* | | |
| Standardized Peer Risk | 0.0025 | 0.0018 |
| | (0.0051) | (0.0050) |
| Provider Above Median Placement Duration=1 × Standardized Peer Risk | -0.0041 | -0.0094 |
| | (0.0085) | (0.0100) |
| Placement Controls | × | × |
| Own Risk Score | × | × |
| Mean of Dep Var | 0.1134 | 0.0540 |
| Observations | 8,474 | 2,332 |

Notes: This table presents the heterogeneity in the estimated relationship between a youth's standardized share of delinquent peers (panel A) or standardized mean peer imprisonment risk (panel B) and teen imprisonment, by the median number of days for a placement with the given provider. For each regression, the base term is the association between the characteristic and teen imprisonment within a provider over time, and the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.9: Peer Risk and Own Imprisonment
Heterogeneity by Child Welfare Experiences

| | Prison Before Age 20 | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Peer Risk | 0.0019 | -0.0034 | 0.0013 | 0.0175 |
| | (0.0049) | (0.0081) | (0.0053) | (0.0436) |
| Length of Stay in Months × Standardized Peer Risk | -0.0005 | | | |
| | (0.0010) | | | |
| Placement Avg N of Peers, in Tens × Standardized Peer Risk | | 0.0121 | | |
| | | (0.0228) | | |
| N Months with Given Peer × Standardized Peer Risk | | | -0.0009 | |
| | | | (0.0025) | |
| Age at Placement Start × Standardized Peer Risk | | | | -0.0011 |
| | | | | (0.0027) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 |

Notes: This table presents the heterogeneity in the estimated relationship between a youth's standardized mean peer imprisonment risk and teen imprisonment by characteristics of the youth's provider and foster care experiences. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.10: Peer Risk and Own Imprisonment
Heterogeneity by Background Characteristics

| | Prison Before Age 20 | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Peer Risk | -0.0010 | -0.0014 | 0.0027 | 0.0002 |
| | (0.0048) | (0.0055) | (0.0054) | (0.0043) |
| Black=1 × Standardized Peer Risk | 0.0045 | | | |
| | (0.0097) | | | |
| Female=1 × Standardized Peer Risk | | 0.0064 | | |
| | | (0.0063) | | |
| Own Risk Score × Standardized Peer Risk | | | -0.0180 | |
| | | | (0.0504) | |
| Parent Prison in Childhood=1 × Standardized Peer Risk | | | | 0.0026 |
| | | | | (0.0117) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 | 8,474 |

Notes: This table presents the heterogeneity in the estimated relationship between a youth's standardized mean peer imprisonment risk and teen imprisonment by characteristics of the youth's personal background prior to placement. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.11: Delinquent Peers and Youth Outcomes, Disaggregated by Gender

| | Prison Before Age 20 | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Standardized Female Peer Delinquency | -0.0048 | | | |
| | (0.0058) | | | |
| Female=1 × Standardized Female Peer Delinquency | 0.0004 | | | |
| | (0.0070) | | | |
| Standardized Male Peer Delinquency | | -0.0039 | | |
| | | (0.0065) | | |
| Female=1 × Standardized Male Peer Delinquency | | 0.0117 | | |
| | | (0.0104) | | |
| Standardized Female Peer Risk | | | 0.0018 | |
| | | | (0.0051) | |
| Female=1 × Standardized Female Peer Risk | | | 0.0040 | |
| | | | (0.0060) | |
| Standardized Male Peer Risk | | | | -0.0015 |
| | | | | (0.0055) |
| Female=1 × Standardized Male Peer Risk | | | | 0.0042 |
| | | | | (0.0070) |
| Placement Controls | × | × | × | × |
| Own Risk Score | × | × | × | × |
| Mean of Dep Var | 0.0826 | 0.1243 | 0.0826 | 0.1243 |
| Observations | 5,717 | 7,652 | 5,717 | 7,652 |

Notes: This table presents the heterogeneity in the estimated relationship between a youth's peer characteristics and own teen imprisonment by own gender and peer gender. For each regression, the base term is the association between the characteristic and teen imprisonment for males, and the interaction term is the difference in the peer effect coefficient for females. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.12: Peer Demographics and Own Imprisonment

| | Prison Before Age 20 | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Share Black Peers | 0.0112 | | |
| | (0.0224) | | |
| Share Female Peers | | -0.0026 | |
| | | (0.0242) | |
| Average Peer Age | | | 0.0048 |
| | | | (0.0044) |
| Placement Controls | × | × | × |
| Own Risk Score | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 |

Notes: This table presents the relationship between peer demographic characteristics and own teen imprisonment. Own delinquency is controlled for but not displayed due to correlation with the risk score control. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.13: Peer Demographics and Own Imprisonment, by Gender

| | Prison Before Age 20 | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Share Black Peers | 0.0116 | | |
| | (0.0295) | | |
| Female=1 × Share Black Peers | -0.0012 | | |
| | (0.0360) | | |
| Share Female Peers | | -0.0055 | |
| | | (0.0401) | |
| Female=1 × Share Female Peers | | 0.0081 | |
| | | (0.0487) | |
| Average Peer Age | | | 0.0074 |
| | | | (0.0083) |
| Female=1 × Average Peer Age | | | -0.0053 |
| | | | (0.0088) |
| Placement Controls | × | × | × |
| Own Risk Score | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 |
| Observations | 8,474 | 8,474 | 8,474 |

Notes: This table presents the heterogeneity in the estimated relationship between peer demographic characteristics and own teen imprisonment by gender. For each regression, the base term is the association between the characteristic and teen imprisonment for males, and the interaction term is the difference in the peer effect coefficient for females. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.14: Alternative Models of Peer Risk and Own Imprisonment

| | Prison Before Age 20 | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Standardized Peer Delinquency | -0.0037 | | | -0.0004 | |
| | (0.0048) | | | (0.0057) | |
| Any Delinquent Peer | -0.0066 | | | | |
| | (0.0065) | | | | |
| Standardized Peer Risk | | 0.0014 | 0.0009 | | 0.0070 |
| | | (0.0046) | (0.0041) | | (0.0052) |
| Max Peer Score | | -0.0342 | | | |
| | | (0.0675) | | | |
| Min Peer Score | | | -0.0697 | | |
| | | | (0.3609) | | |
| Standardized Opp Gend Peer Delinquency | | | | -0.0009 | |
| | | | | (0.0047) | |
| Standardized Opp Gend Peer Risk | | | | | 0.0026 |
| | | | | | (0.0043) |
| Placement Controls | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.1134 | 0.1134 | 0.0944 | 0.0944 |
| Observations | 8,474 | 8,474 | 8,474 | 4,895 | 4,895 |

Notes: This table presents alternative models for the relationship between peer delinquency or risk and own teen imprisonment. Column 1 includes a binary indicator for having at least one delinquent peer during a placement spell. Columns 2 and 3 include the maximum and minimum imprisonment risk of a peer during a placement spell, respectively. Columns 4 and 5 include coefficients for opposite gender peer delinquency and risk, respectively, for placements shared with at least one youth of the opposite gender. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.15: Nonlinear Model of Peer Risk and Own Imprisonment

| | (1) All | (2) Bottom Risk Quartile | (3) Second Quartile | (4) Third Quartile | (5) Top Quartile |
|---|---|---|---|---|---|
| *Outcome: Prison Before 20* | | | | | |
| Share of Peers in Risk Quartile 2 | -0.0029 | 0.0289 | 0.0447 | 0.0341 | -0.0175 |
| | (0.0225) | (0.0314) | (0.0510) | (0.0401) | (0.0590) |
| Share of Peers in Risk Quartile 3 | 0.0189 | -0.0331 | 0.0873 | 0.0803 | -0.0214 |
| | (0.0267) | (0.0432) | (0.0547) | (0.0587) | (0.0558) |
| Share of Peers in Risk Quartile 4 | -0.0134 | -0.0481 | 0.0305 | -0.0527 | -0.0191 |
| | (0.0268) | (0.0441) | (0.0533) | (0.0428) | (0.0694) |
| Placement Controls | × | × | × | × | × |
| Own Risk Score | × | × | × | × | × |
| Mean of Dep Var | 0.1134 | 0.0525 | 0.1039 | 0.1181 | 0.1698 |
| Observations | 8,474 | 1,885 | 1,973 | 2,270 | 2,226 |

Notes: This table presents estimates of the nonlinear relationship between a youth's peer imprisonment risk and teen imprisonment. Each regression includes a youth's share of peers in the second, third, and fourth quantiles of peer imprisonment risk, with the first quantile as an omitted category, where a higher quantile indicates greater risk. Columns 2-5 present results disaggregated by a youth's own risk quantile. Risk quantiles are calculate using the full set of congregate care placements in Wisconsin from 2004-2019 before any sample restrictions. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.16: Association between Peer Imprisonment Risk and High School Dropout Heterogeneity by Placement Type

| | All | Group Home | Residential Facility | Temporary |
|---|---|---|---|---|
| | High School Dropout | High School Dropout | High School Dropout | High School Dropout |
| Standardized Peer Risk | 0.0094* | 0.0030 | 0.0174 | 0.0147* |
| | (0.0057) | (0.0093) | (0.0119) | (0.0084) |
| Placement Controls | × | × | × | × |
| Own Incar Risk Score | × | × | × | × |
| Mean of Dep Var | 0.3795 | 0.3974 | 0.3583 | 0.3723 |
| Observations | 6,785 | 2,952 | 1,778 | 2,055 |

Notes: This table presents the relationship between standardized peer mean imprisonment risk and own high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, and quarter-of-entry fixed effects. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.17: Association between Peer Imprisonment Risk and High School Completion Heterogeneity by Placement Type

| | All | Group Home | Residential Facility | Temporary |
|---|---|---|---|---|
| | High School Eqv Completion | High School Eqv Completion | High School Eqv Completion | High School Eqv Completion |
| Standardized Peer Risk | -0.0172** | -0.0027 | -0.0249 | -0.0297*** |
| | (0.0083) | (0.0139) | (0.0265) | (0.0093) |
| Placement Controls | × | × | × | × |
| Own Incar Risk Score | × | × | × | × |
| Mean of Dep Var | 0.5906 | 0.5688 | 0.6125 | 0.6020 |
| Observations | 4,353 | 1,860 | 1,169 | 1,324 |

Notes: This table presents the relationship between standardized peer mean imprisonment risk and own high school equivalent completion. High school completion is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.18: Peer Imprisonment Risk and High School Dropout Heterogeneity by Child Welfare Experiences

| | High School Dropout | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Age at Placement Start × Standardized Peer Risk | -0.0012 | | | | | | |
| | (0.0047) | | | | | | |
| Own Delinquency=1 × Standardized Peer Risk | | 0.0025 | | | | | |
| | | (0.0142) | | | | | |
| Predicted Teen Parent Risk × Standardized Peer Risk | | | -0.1301 | | | | |
| | | | (0.1244) | | | | |
| SSI Before Placement=1 × Standardized Peer Risk | | | | 0.0012 | | | |
| | | | | (0.0145) | | | |
| Placement Avg N of Peers × Standardized Peer Risk | | | | | 0.0030*** | | |
| | | | | | (0.0009) | | |
| Length of Stay in Months × Standardized Peer Risk | | | | | | -0.0024 | |
| | | | | | | (0.0017) | |
| N Months with Given Peer × Standardized Peer Risk | | | | | | | -0.0054 |
| | | | | | | | (0.0048) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Incar Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | 0.3795 | 0.3795 | 0.3808 | 0.3795 | 0.3795 | 0.3795 | 0.3795 |
| Observations | 6,785 | 6,785 | 2,135 | 6,785 | 6,785 | 6,785 | 6,785 |

Notes: This table presents heterogeneity in the estimated relationship between a youth's standardized peer delinquency and high school dropout by characteristics of the youth's provider and foster care experiences. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.19: High School Dropout: Heterogeneity by Peer Characteristics

| | High School Dropout | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Standardized Female Peer Incar Risk | 0.0026 | | | | |
| | (0.0074) | | | | |
| Standardized Male Peer Incar Risk | | 0.0022 | | | |
| | | (0.0068) | | | |
| Standardized Female Peer Delinquency | | | -0.0107 | | |
| | | | (0.0076) | | |
| Standardized Male Peer Delinquency | | | | -0.0028 | |
| | | | | (0.0085) | |
| Share Female Peers | | | | | 0.0171 |
| | | | | | (0.0497) |
| Placement Controls | × | × | × | × | × |
| Own Incar Risk Score | × | × | × | × | × |
| Mean of Dep Var | 0.3641 | 0.3790 | 0.3641 | 0.3790 | 0.3795 |
| Observations | 4,534 | 6,122 | 4,534 | 6,122 | 6,785 |

Notes: This table presents estimates of the relationship between peer characteristics and high school dropout. High school dropout is calculated only for youth still enrolled at the time of placement and who did not disenroll due to moving out of state. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, and own imprisonment risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.20: Peers and Teen Parenthood

| | Parent Before Age 20 (Females Only) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Standardized Female Peer Incar Risk | 0.0004 | | | | | |
| | (0.0113) | | | | | |
| Standardized Male Peer Incar Risk | | 0.0132 | | | | |
| | | (0.0102) | | | | |
| Standardized Female Peer Delinquency | | | -0.0027 | | | |
| | | | (0.0127) | | | |
| Standardized Male Peer Delinquency | | | | -0.0054 | | |
| | | | | (0.0129) | | |
| Share Female Peers | | | | | -0.0441 | |
| | | | | | (0.0796) | |
| Standardized Peer Teen Parent Risk | | | | | | 0.0031 |
| | | | | | | (0.0101) |
| Placement Controls | × | × | × | × | × | × |
| Own Incar Risk Score | × | × | × | | × | × |
| Teen Parent Risk Score | × | × | × | × | × | × |
| Mean of Dep Var | 0.3581 | 0.3493 | 0.3581 | 0.3493 | 0.3581 | 0.3581 |
| Observations | 2,365 | 1,632 | 2,365 | 1,632 | 2,365 | 2,365 |

Notes: This table presents estimates of the relationship between peer characteristics and teen parenthood for females 13 years and up. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, own imprisonment risk score, and own teen parenthood risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.21: Peers and Teen Parenthood
Heterogeneity by Placement Type

| | All | Group Home | Residential Facility | Temporary |
|---|---|---|---|---|
| | Parent Before Age 20 (Female) | Parent Before Age 20 (Female) | Parent Before Age 20 (Female) | Parent Before Age 20 (Female) |
| Standardized Male Peer Incar Risk | 0.0132 | 0.0193 | 0.0171 | 0.0082 |
| | (0.0102) | (0.0193) | (0.0197) | (0.0151) |
| Placement Controls | × | × | × | × |
| Own Incar Risk Score | × | × | × | × |
| Teen Parent Risk Score | × | × | × | × |
| Mean of Dep Var | 0.3493 | 0.3506 | 0.3316 | 0.3622 |
| Observations | 1,632 | 328 | 564 | 740 |

Notes: This table presents heterogeneity in the estimated relationship between standardized male peer mean imprisonment risk and teen parenthood for females 13 years and up by type of provider. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, own imprisonment risk score, and own teen parenthood risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.A.22: Peers and Teen Parenthood
Heterogeneity by Risk Factors

| | Parent Before Age 20 (Females Only) | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Age at Placement Start × Standardized Male Peer Incar Risk | -0.0089 | | | | | | |
| | (0.0099) | | | | | | |
| Own Delinquency=1 × Standardized Male Peer Incar Risk | | -0.0180 | | | | | |
| | | (0.0337) | | | | | |
| Predicted Teen Parent Risk × Standardized Male Peer Incar Risk | | | -0.2091 | | | | |
| | | | (0.1729) | | | | |
| SSI Before Placement=1 × Standardized Male Peer Incar Risk | | | | 0.0316 | | | |
| | | | | (0.0350) | | | |
| Placement Avg N of Peers × Standardized Male Peer Incar Risk | | | | | 0.0008 | | |
| | | | | | (0.0022) | | |
| Length of Stay in Months × Standardized Male Peer Incar Risk | | | | | | 0.0032 | |
| | | | | | | (0.0038) | |
| N Months with Given Peer × Standardized Male Peer Incar Risk | | | | | | | 0.0051 |
| | | | | | | | (0.0100) |
| Placement Controls | × | × | × | × | × | × | × |
| Own Incar Risk Score | × | × | × | × | × | × | × |
| Mean of Dep Var | 0.3493 | 0.3493 | 0.3493 | 0.3493 | 0.3493 | 0.3493 | 0.3493 |
| Observations | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 |

Notes: This table presents heterogeneity in the estimated relationship between standardized male peer mean imprisonment risk and teen parenthood for females 13 years and up by placement experiences and youth characteristics prior to placement. For each regression, the interaction term is the change in the peer effect coefficient per one unit increase in the characteristic described. The base terms of the interaction are included in each regression but are not displayed for presentational clarity. All regressions include controls for provider-by-gender fixed effects, year-of-entry fixed effects, quarter-of-entry fixed effects, number of non-sibling peers, own imprisonment risk score, and own teen parenthood risk score. Standard errors are provided in parentheses and are two-way clustered by provider and youth. Significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

# Chapter 3

# Gatherings, Giving, and Life Trajectories: Evidence from Student Missions Conferences[1]

[1]I thank Leah Boustan, Janet Currie, Dan Hungerman, seminar participants at ASREC Europe, and members of the Princeton Industrial Relations Section for their valuable comments and suggestions. Rachel Conner and InterVarsity Christian Fellowship/USA provided the data and invaluable institutional context necessary to undertake this project. The author received IRB approval for the analysis of the data from his institution. The author has no additional disclosures.

## 3.1 Introduction

Large social gatherings, such as religious retreats, sporting events, and political rallies, are frequently cited as highly formative life experiences. On the other hand, such events are considered by others to be transient episodes that are primarily symbolic or leisure. In what ways do these "manufactured" gatherings affect participant social and economic behavior? Recent studies have found mixed evidence of the effectiveness of social, political, and religious gatherings in producing short-run or lasting change (Snyder Jr and Yousaf 2020; Lowe 2020). Understanding when and for whom gatherings impact long-run, cause-specific investment is important for advancing our knowledge of how gatherings affect society, their utility for constrained philanthropic and mission-driven organizations (Khan 2020), cost-effective nudges (Carrell and Sacerdote 2017; Porter and Serra 2020), and welfare losses associated with potentially inefficient, ill-timed social investments (Montero and Yang 2021; Suzuki 2021).

This paper studies the impact of a particularly salient contemporary religious gathering, Urbana, the largest North American student missions conference, on support for the sponsor organization and student life trajectories. I link together multiple sources of novel administrative and survey data collected by the event sponsor, to measure participation effects. Using variation in the triennial timing of the five-day conference from 2000-2017, I find that a conference produces a 26 percent increase in following-year student donations to the sponsor organization and a 41 percent increase in short-run staff worker employment for the organization. Labor supply results are revealed by a unique institutional feature that removes the demand constraint: workers must partially support-raise for themselves, so the organization employs nearly every applicant who meets requisite qualifications.[2] Short-run findings are robust to a separate identification strategy that combines rotational variation with variation in college club propensity to attend.

The impacts of attendance persist long after leaving college. I estimate the long-run

[2]Giving results are largely independent of staff support-raising.

impacts of gathering attendance by exploiting an important administrative data feature (being able to observe registrants who cancelled or did not show) and geographic variation in weather-related cancellations in flights to the conference, which are uncorrelated with pre-conference religiosity and organizational participation. These estimates, together with supplemental regression specifications using other sources of variation, confirm long-run impacts of conference attendance on future giving and full-time employment with the sponsor organization: a 10 percent increase in weather-related flight cancellations from a student's university's state is associated with a $216 decrease in long-run giving through Spring 2020. Under a bounding exercise using variation in attendance by graduation cohort, I find that each additional senior student attendee is associated with a $159-$599 increase in giving within three years of graduation and an additional 0.005 to 0.02 employees, about one additional employee per every 100 attendees. Overall, attendance appears to cause a 22 percent increase in long-run giving and a 36 percent increase in long-run staff employment.

Attendance appears to affect other margins of behavior as well. I find suggestive evidence that conference attendance increases student attendance of short-term mission trips to other countries. Supplemental survey data from graduating seniors shows that attendance induced certain students to find post-college employment with religious organizations, even beyond the sponsor organization, and campus-level survey data suggests that the gathering may have affected the religious affiliation of some attendees.[3]

Finally, I consider which groups are most affected by conference attendance. In contrast to prior literature that documents stronger effects of social movements on younger participants (Hungerman and Moorthy 2020), the effects of conference attendance in this context are strongest in absolute and percentage terms for undergraduate seniors, who are closer to graduation. This finding is consistent with the hypothesis that certain gatherings can be most salient for attendees at life transition points. Additionally, I find that gatherings are

[3]Forthcoming editions of this paper will consider complementary mechanisms of investments in identity capital versus increasing the salience of religious careers (Bénabou and Tirole 2011), and incorporate long-run socioeconomic outcomes from consumer insights data sources: I find preliminary evidence that attendance may have increased long-run marriage rates for men.

most influential for those students who are already actively involved within a club affiliate of the sponsor organization. Relatedly, I consider the role of targeted messaging on underrepresented students using a difference-in-differences design studying the 2015 conference's public support of Black Lives Matter. I find that BLM messaging did not affect Black student donations or sponsor-related employment.

This paper provides new insights to a growing literature on the impacts of social gatherings on the economy and society. While past work has considered the impacts of religious revivals (Lowe 2020), this paper examines a planned gathering that can – and indeed has – been replicated globally.[4] In this sense, this paper is adjacent to studies on political campaigns, protests, and participation (Madestam et al. 2013; Fujiwara, Meng and Vogl 2016; Snyder Jr and Yousaf 2020; Hungerman and Moorthy 2020; Larreboure and González 2021; Castro and Retamal 2022), but with an extended set of outcomes from participant-level administrative data. In one of the few studies to identify individual participants, Clingingsmith, Khwaja and Kremer (2009), show that Hajj participation increased Islamic religious practices and inclusive attitudes for at least 5-8 months. It is unclear, however, to what extent these results generalize to other types of gatherings due to the extensive and physically-taxing nature of Hajj travel and participation.[5]

Second, this paper considers the relationship between the timing of social investments, prosocial behavior, and economic mobility. While past studies have documented negative consequences of ill-timed festivals in developing countries, religious or otherwise (Montero and Yang 2021; Suzuki 2021), on household savings and economic growth, this paper presents contemporary evidence of how gatherings close to graduation create life-altering changes in career choice with potential implications for long-run living arrangements (Carrell and Sacerdote 2017; Porter and Serra 2020).[6] This work relatedly deepens our understanding of

[4]The type of gathering studied in this paper is also of independent interest: Urbana has been a model for similar student gatherings in the United States, Africa, Latin America, Southeast Asia, and Europe. https://intervarsity.org/news/the-international-urbana-movement

[5]The authors find a reduction in self-reported emotional and physical wellbeing among participants, especially for women.

[6]Latter result is forthcoming with the integration of Infutor data.

the determinants of charitable giving (Paxton, Velasco and Ressler 2020) by showing how temporary, targeted experiences can yield further giving through habit formation and inertia for students who are near the beginning of their professional careers.

Finally, my work provides insights for how organizations and social movements recruit prospective donors and staff for cause areas. Whereas Khan (2020) shows how mission-minded organizations can use motivational videos to improve worker performance, this paper provides evidence of how showcasing organizational vision at a gathering increases worker supply along the extensive margin, as well as donations along both extensive and intensive margins. These findings suggest that planned gatherings can be an effective tool for public and nonprofit organizations to recruit and compensate employees, especially those motivated by the mission of the organization (c.f., Ashraf et al. 2015).

The paper proceeds as follows. Section 2 provides institutional context about the Urbana student missions conference gathering. Section 3 describes data sources. Section 4 presents the empirical strategy and findings for short-run effects of gathering attendance. Section 5 presents medium and long-run results from a cohort-specific design. Section 6 provides further estimates of long-run results using a cancellation-based design. Section 7 considers mechanisms and heterogeneity. Section 8 presents the results of group-specific messaging on student behavior. Finally, Section 9 concludes.

## 3.2 The Urbana Missions Conference

Much of the world's population participates in religious services, retreats, and gatherings each year. The Urbana missions conference is the largest student missions conference in North America, and at one time the largest in the world. The sponsor organization, InterVarsity Christian Fellowship, began the conference in 1946 with the goal of equipping and nudging Canadian and U.S. college students to participate in Christian international missions.[7] Since

[7]Missions involve attempts to spread Christian beliefs to new areas through preaching and conversations about the Christian faith, sometimes accompanied by poverty alleviation work. Thomas (2016) reviews

the 1970s, the conference has also publicized to students global and domestic challenges such as HIV/AIDS and U.S. racial oppression.[8]

The Urbana missions conference consists of a combination of Christian religious programming, workshops, and a job fair for domestic and international religious employers. For most of its history, the Urbana missions conference has occurred once every three years at the University of Illinois at Urbana-Champaign (1948-2003) and St. Louis, MO (2006-2018). During the study period of 2000-2015, the conference took place during five days from December 27-31 and attendance ranged from approximately 10,000-20,000 participants per gathering. Conference content during these years included worship with religious and international music, talks from missionaries and other figures involved with religious and service work, Bible study and prayer, leadership development workshops, and in recent years special tracks dedicated to content such as hackathons, domestic and international development, entrepreneurship, pastoral ministry, and medicine. An additional component of the Urbana missions conference is a job fair where students can solicit representatives from domestic and overseas Christian organizations about employment and volunteer opportunities. The conference's vocation and lifestyle-oriented content is reflected in recent Urbana slogans: *"What life will you live?"*, *"Find your life,"* and *"Experience a sacred space to interact with missions organizations and experts to discern where your passions fit in God's global work."* The Urbana missions conference has been a model for similar large student missions conferences in other continents and for other large North American Protestant and Catholic student conferences.[9]

Many conference attendees are undergraduate participants of an affiliated sponsor organization club (InterVarsity fellowship chapter) on their college campus. However, attendees also include students not affiliated with the sponsor organization, including some students

some of the language and history of the earliest Urbana missions conferences during the postwar era of North American evangelicalism.

[8]https://mem.intervarsity.org/resources/tom-skinners-urbana-70-address-racism-and-world-evangelism

[9]The conference has offshoots in Africa, Latin America, Southeast Asia, and Europe.https://intervarsity.org/news/the-international-urbana-movement

of other or no particular religious affiliation. The conference has a registration fee of approximately $300-$600 including meals and accommodations, with subsidies available based on financial need.

A final noteworthy aspect of the Urbana missions conference is the "Urbana offering," a voluntary collection of donations for sponsor organization global ministry partners taken on the final day of the missions conference. Although student attendees frequently donated in cash during the study period of interest, any personally-identifiable donations such as checks or digital giving will contribute to a near-mechanical association of conference attendance with short-run student giving.

## 3.3 Data Sources

An important contribution of this paper relative to other studies on large gatherings is the integration of conference registration records with individual survey and administrative data to study the effects of attendance on students over time. Several sources of individual- and aggregate-level data, provided and linked by the department of Development Research and Data Analytics at InterVarsity, are used to estimate the impacts of gatherings. These include conference registration records, donation records, staff employment records, event registrations, departing student form survey data, and annual field report campus-level survey data.

### 3.3.1 Conference Registration Data

All students who register for the Urbana missions conference are given a digital registration record, regardless of whether the student attends. This record includes information about the student's self-reported graduation year, school, demographics, and whether the student is a member of a college club affiliated with the sponsor organization. Additional information is compiled by the sponsor organization, including the before- and after-subsidy registration fee paid by a student, a registration cancellation date if applicable, and an indicator for

whether the student attended (arrived and received an entry badge for) the conference.

### 3.3.2 Flight Cancellation Records

One identification strategy in this paper uses weather-related flight cancellations during conference commuting days to show that bad weather reduces event attendance and affects long-run student outcomes. Flight cancellation data for major airlines are publicly available from the Bureau of Transportation Statistics (BTS) of the United States Department of Transportation (DOT).[10] The data contain information about flight date, airport of scheduled departure, airport of scheduled arrival, delays until departure, air time, flight cancellation, and reason for cancellation. Weather conditions account for the majority of flight cancellations. I define a weather-related flight cancellation as a flight that is cancelled due to adverse weather conditions (2003-present), or a flight that is cancelled for any reason in years where a reason is not available (2000).

### 3.3.3 Administrative Microdata

The sponsor organization maintains donor records for every donation with identifiable donor information. These records include the date and dollar amount of each individual donation from November 2, 1998 through April 30, 2020. Married couples receive a joint household identifier for donor records; for the small subset of donations that are not individually identified I match an individual to their household ID and divide the donation amount by two.

Each donation record includes an anonymous person-level identifier that can be used to link donation records to other data sources from the sponsor organization. The sponsor organization matches individuals across different record sources using a fuzzy matching process involving first and last name, date of birth, home address, email address, gender, and race/ethnicity if available. Analysts from the sponsor organization estimate that, for

[10]Extracts were downloaded from https://www.transtats.bts.gov/ on August 23, 2021. For the closest Web Archive version of the page, see https://web.archive.org/web/20210620090047/https://transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ

registrants who donate for the first time within three years of the conference, match rates from a non-anonymous donation to a conference registration record are greater than 80 percent. They estimate that matches are near 100 percent from registration records to staff employment records. This paper employs a variety of robustness checks to confirm that the estimated impacts of conference attendance are not due to systematic differences in record linkages between attendees and non-attendees.

Anonymized staff employment records provided by the sponsor organization include employment histories for all full-time and part-time field staff employees, who oversee affiliated college student organizations, counsel students and provide religious and leadership training, and frequently serve as chaplains in college religious life offices. Employees who work in non-student-facing administrative roles are not included in the dataset. Field staff records are available from 1979 through March 31st, 2021 and include start and end dates for each college campus staffed, staff demographic information, own undergraduate institution, and graduation year when known.

An additional source of administrative data are event registrations from a select number of other events organized by the sponsor organization. Of these, the most common are Global Program mission trips (short-term international programs) and gatherings for students involved with Greek-letter organizations (fraternities and sororities).

### 3.3.4 Survey Data

All graduating and otherwise disenrolling students are invited to complete a departing student form. These forms are often filled out by students themselves, but are sometimes filled out by a student's staff worker. The departing student form includes information about a student's undergraduate institution, graduation year, self-reported involvement in club leadership roles, non-binding indicators about whether the student will support the sponsor organization after graduation (e.g., pray, give, volunteer), and optional fields for the student's job plans after college.

A more comprehensive but aggregated source of survey data is the campus-level Annual Field Report, completed by each affiliated student organization during each Spring semester.[11] The Annual Field Report campus survey is submitted by staff often in consultation with student leadership. The survey includes self-reported estimates of regular club members and visitors, member demographics, attendance at various types of club events. The survey also includes several metrics for student beliefs and behavior that are relevant for this study: the number of student members who regularly attend church, the number of departing seniors who became Christians in college, the number of members who identified as not Christian at some point during the academic year, and the number of students who plan to talk about Christianity with an international student friend.

## 3.4 Short-Run Effects of Gatherings

### 3.4.1 Short-Run Empirical Strategy

The triennial timing of conference gatherings provides an intuitive identification strategy for estimating the impact of the conference on short-run student outcomes. I estimate the following regression specification:

$$Y_{jt} = \beta_0 + \beta_1 ConfYear_t + \lambda_{j \times cycle(t)} + \epsilon_{jt}, \tag{3.1}$$

where $Y_{jt}$ is the outcome of interest for college $j$ during academic year $t$ and $ConfYear_t$ is an binary indicator for whether the Urbana conference took place in academic year $t$. $\beta_1$ is the primary coefficient of interest, representing the change in $Y_{jt}$ associated with a conference. $\lambda_{j \times cycle(t)}$ are college-by-conference-cycle fixed effects, isolating variation to within a given college and the years before, during, and after a particular conference year. For example, the 2014-2015, 2015-2016, and 2016-2017 academic years for the University of Michigan receive

[11] A less detailed Fall Field Report is also completed annually, but has been omitted from the analysis due to inconsistent records at some schools. Field reports are technically submitted at the club level, but in this paper schools with multiple affiliated clubs have been aggregated to the campus level.

the same fixed effects coefficient. $\epsilon_{jt}$ is a random error term for unexplained variation in the outcome. Standard errors are clustered at the college level.

The primary assumption used in model 3.1 is that the timing of conferences is uncorrelated with systematic differences in the outcome variable. This assumption would be violated, for example, if economic recessions occurred more or less often during conference years and affected giving and staff employment. In conversations with sponsor organization representatives, I have been told there are no systematic organizational differences in staff onboarding and recruitment efforts during conference years. In some cases, smaller events are scheduled to years where the conference is not held, such as the triennial staff conference which is held 13 months following the Urbana missions conference. To the extent that these smaller events increase giving and related outcomes, the estimates relying solely on variation in conference timing may be underestimates of the true effect of the conference.

Equation 3.1 is agnostic about whether the effect of conferences is through student attendance or some other channel, such as staff worker attendance. To estimate the impact of student attendance more directly and weaken the assumptions needed about year-to-year variation in confounders, I estimate an additional specification relying on the interaction of conference years with a college's propensity to attend:

$$\begin{aligned} Y_{jt} = \beta_0 + \beta_1 AttendPropensity_{j\times cycle(t)} + \beta_2 ConfYear_t \times AttendPropensity_{j\times cycle(t)} \\ + \beta_3 Size_{j\times cycle(t)} + \beta_4 Size_{j\times cycle(t)} \times AttendPropensity_{j\times cycle(t)} + \gamma_t + \epsilon_{jt}, \end{aligned} \tag{3.2}$$

where again $Y_{jt}$ is the outcome of interest for college $j$ during academic year $t$, $ConfYear_t$ is an binary indicator for whether the Urbana conference took place in academic year $t$, and $\epsilon_{jt}$ is a random error term for unexplained variation in the outcome. In this specification, year fixed effects $\gamma_t$ control for annual trends in the outcome of interest shared across all colleges. Variation used to estimate the effect of attendance, therefore, relies on an additional dimension: college-level conference attendance propensity. $AttendPropensity_{j\times cycle(t)}$ is the total number attendees for a school within a cycle, applied also to the years before and after the conference year. $\beta_2$ is the coefficient of interest, the change in outcome associated with an

additional attendee in a conference year. I also control for multiple measures of college club size (number of members, number of discussion group leaders, number of student leaders) interacted by conference year to control for year-to-year trends that systematically impact larger or smaller college clubs.

The advantage of equation 3.2 is that it allows for a more direct estimate of student attendance on short-run outcomes. The propensity measure is agnostic as to the reasons for high attendance propensity relative to club size. For example, the University of Illinois Urbana-Champaign (UIUC) has a relatively greater conference year attendance shock than Cornell University because UIUC is located close to the conference site. One modest but important assumption for a causal interpretation of $\beta_2$ is that schools with greater attendance propensity are not systematically exposed to different shocks to outcome $Y_{jt}$ during conference years. A potential shortcoming of this specification is that it requires effects of conference attendance to be short-lasting and without spillovers to other years. To the extent that this assumption is violated, estimates of $\beta_2$ will be underestimates of the true effects of conference attendance.

### 3.4.2 Short-Run Sample and Descriptive Statistics

The sample used to estimate short-run effects consists of a balanced panel of college-by-year observations running from 2003-2017.[12] Two conditions are imposed in the construction of the sample. First, colleges must have filled out the Annual Field Report Spring Survey for each Spring from 2003 to 2017. Second, each college must have had a total of at least 10 undergraduate conference attendees during the sample period. The remaining sample consists of 279 unique colleges and 4,185 college-year observations.

Conference years are defined as observations in the calendar year following a conference. For example, outcomes collected in 2004 are considered to be a conference year because the conference took place in late December 2003. Conference attendance during December 2003

[12]Some colleges include multiple clubs affiliated with the sponsor organization. These clubs are aggregated to the college level throughout.

is applied to the 2004 calendar year observation.

Table 3.10.1 presents summary statistics for the sample used to estimate short-run and cohort-level effects of conference attendance. Column 1 describes key variables that apply to conference years ($n = 1,395$), and Column 2 non-conference years ($n = 2,790$). The average number of undergraduate attendees in a conference year is 17.8 per college, and the number of attendees is on average increasing by grade. By construction, there are zero attendees during non-conference years. Additionally, we observe that undergraduate student donations in a conference year (i.e., the year immediately following a conference) appear to exceed those during other years. The same pattern appears to apply for new staff employment from current or former undergraduates from a college. The "Campus Survey" and "Survey Demographics" panels compare student behavior outcomes and campus demographics, respectively, taken from the Annual Field Report Spring Survey. Notably, student demographics are not noticeably different during conference years, consistent with the assumption that the two samples are comparable on underlying characteristics.

### 3.4.3 Short-Run Results

Table 3.10.2 presents estimates of the effect of conference gatherings on student outcomes. Panel A presents estimates of equation 3.1, which compares conference years versus non-conference years. Panel B presents estimates of equation 3.2, which compares conference years versus non-conference years for schools with high attendance propensities relative to schools with low attendance propensities. Panel A therefore estimates the impact of the conference as a whole whereas the "Campus Attendance in Cycle × Conference Year" interaction term in Panel B estimates the impact of one additional undergraduate attendee.

Panel A, column 1 shows that conference attendance is 17.8 per college during conference years, compared to zero attendance during non-conference years. Panel B, column 1 shows that college attendance is mechanically one student greater per each student attendee.

Table 3.10.2 columns 2 presents short-run estimates of the impact of conferences on

student giving. Undergraduate student giving per is $416 more per college during conference years compared to non-conference years (Panel A) and approximately $18 more per student attendee (Panel B), with a mean of $1,600 per college-year over the full sample. This result suggests that the conference contributes to a modest 26 percent short-run increase in student giving, or a $116,000 increase across all schools in sample in a given year.

An important qualifier when considering the external validity of the student giving result is that a call for donations is taken on the last day of the conference (the "Urbana offering"), which could contribute to a mechanical increase in student giving. Column 3 presents estimates of conference impacts on giving excluding the donations from December 25th-January 4th, days including and surrounding the conference window. Although conference years are still associated with an increase in student giving per college ($98 relative to a mean of $1,407), this result is greatly attenuated compared to column 2 and no longer significantly different from zero on a per-attendee basis. This findings suggests that much of the short-run effect of the conference on student giving is due to the collection taken at the conference itself.

Finally, column 4 presents of conference impacts on staff employment; the number of students who attended a college who become employed by the sponsor organization between April and November, the main staff hiring period. Panel A shows that a conference year is associated with 0.1 additional new staff hires per college compared to a mean of 0.26, approximately a 41 percent increase in new staff hires and a total of 29 additional new staff across all schools in the sample. Similarly, Panel B shows that each additional undergraduate attendee is associated with 0.008 new staff hires in the next hiring period, about 1 new hire per 100 student attendees. Note that undergraduate attendees in the denominator include students who are not graduating, so the per-student coefficient may be an underestimate if the effect is limited to graduating seniors.

On the whole, Table 3.10.2 shows that the Urbana missions conference increases short-run student giving and staff employment by economically significant amounts. The approx-

imate magnitudes and significance of these estimates are robust to substituting a college's within-cycle attendance with a simulated attendance measure using average attendance in the preceeding and proceeding conference cycles (Table 3.A.1). The estimates are also robust to restricting the outcomes only to students who at any point filled out a departing student form (Table 3.A.2). Students who fill out a departing student form are extremely likely to be successfully matched across administrative records, and the number of students who fill out a departing student form does not increase in conference years (Table 3.A.4), so the estimates in Table 3.A.2 show that the presented short-run findings are not due to data linkage differences for conference attendees.

Beyond giving and employment for the sponsor organization, Table 3.A.3 estimates the effect of conferences on student preferences and job plans. These estimates use questions from departing student forms, which are completed by students and occasionally sponsor organization staff workers around the time of student graduation. Although only a fraction of students complete these forms, column 1 shows that attending the conference does not increase the number of submitted forms, meaning that one can compare responses for conference versus non-conference years with fewer selection concerns. Columns 2-4 indicate that nonbinding pledges taken by the students for praying and volunteering do not increase significantly with conference attendance, with the exception of a marginally significant one additional form with a financial giving pledge per 100 attendees.

A subset of students record job plans following graduation. Columns 5 shows that conference attendance is associate with an increase in the number of students indicating submitting a form with a job plan, and column 6 shows that this increase is primarily concentrated among employment with religious organizations. Table 3.A.4 presents job responses in more detail. Job plans for religious organizations beyond the sponsor organization increase significantly during a conference year, whereas job plans for tech jobs and other large firms fall or remain unchanged despite an increase in the overall number of students with a job plan. These estimates strongly suggest that the Urbana missions conference affects the initial career tra-

jectories of students by reallocating some students into jobs at religious organizations, even beyond the sponsor organization itself. Overall, the findings of Table 3.A.4 provide qualitative support for the findings of Table 3.10.2, and suggest that labor market decisions may be particularly sensitive to nudges during job search periods. The reallocation of students to jobs with religious organizations could be welfare-improving or welfare-decreasing depending on the counterfactual job or non-employment choice of the student, which is less clear from the data. A second important extension for future research is the persistence of careers with religious organizations over time.

Finally, Table 3.A.5 estimates the effect of conferences on student beliefs and behaviors included in the club-level Annual Field Report survey. I find that conference years are associated with a small increase in reported total club members, number of non-Christians students involved in at least half of club events, the number of students who plan to talk about Christianity with an international student friend, and the number of club members who regularly attend church (panel A). However, most outcomes do not appear to be associated with event attendance with the exception of the number of seniors reported to have become Christian during college, which increases by 0.015 (eight percent) per attendee.[13] One explanation for the difference in panels A and B is that conference content affects staff *perceptions* of their club, independent of club size and number of student conference attendees. An important feature of the dual design is that panel B isolates variation in student attendance whereas panel A simply compares conference years with non-conference years. Therefore, panel A includes level shifts in staff perceptions when filling out the annual survey form. The results suggest that the Urbana missions conference may increase the number of students who identify as Christian five months after the event, but that other changes in student behavior from the campus survey may be due to changes in reporting behavior rather than ground truth.

[13]Urbana organizers report a sizeable number of students who become professing Christians at the conference. A fraction of students do not identify as Christian prior to the conference but are invited to attend by friends, family, and organization staff.

## 3.5 Cohort-Level Effects

### 3.5.1 Cohort-Level Empirical Strategy

Having studied the short-run effects of conference attendance, I now consider the longer-run effects of conference attendance on graduation year cohorts. For this analysis, I again exploit variation in exposure to the conference by graduation year and college attendance propensity:

$$Y_{jt} = \beta_0 + \beta_1 ConfYearJunior_t + \beta_2 ConfYearSenior_t + \lambda_{j\times cycle(t)} + \epsilon_{jt}, \tag{3.3}$$

and

$$\begin{aligned} Y_{jt} = \beta_0 &+ \beta_1 AttendPropensity_{j\times cycle(t)} + \beta_2 ConfYearJunior_t \times AttendPropensity_{j\times cycle(t)} \\ &+ \beta_3 ConfYearSenior_t \times AttendPropensity_{j\times cycle(t)} + \beta_4 Size_{j\times cycle(t)} \\ &+ \beta_5 Size_{j\times cycle(t)} \times AttendPropensity_{j\times cycle(t)} + \gamma_t + \epsilon_{jt}, \end{aligned} \tag{3.4}$$

where $Y_{jt}$ is the outcome interest for the cohort attending college $j$ and graduating in year $t$. $ConfYearJunior_t$ and $ConfYearSenior_t$ are binary indicators for whether students were juniors or seniors during a conference year, respectively. In equation 3.3, $\beta_1$ is the difference between students who were juniors during a conference year (regardless of personal attendance) and students who were sophomores during a conference year. $\beta_2$ is the difference between conference-year seniors and conference-year sophomores. As with the short-run estimates, $\lambda_{j\times cycle(t)}$ are college-by-conference-cycle fixed effects, $\epsilon_{jt}$ is the error term, and standard errors are clustered at the college level.

While equation 3.3 compares cohorts who graduated during conference years versus those who graduated during other years, equation 3.4 interacts graduating timing with college attendance during the conference cycle, $AttendPropensity_{j\times cycle(t)}$. $\beta_3$ is therefore interpreted as the change in outcome for conference-year seniors relative to conference-year sophomores for each additional college attendee of any grade.

A significant challenge in identifying long-run cohort-level effects is underidentification: differences in outcomes across cohorts could be due to (1) the number of attendees per

cohort, and (2) the *timing* of attendance during one's undergraduate career. For example, a hypothetical greater level of long-run giving for conference-year seniors relative to conference-year sophomores could be due to systematically higher attendance rates for seniors, or a greater impact of the conference on students close to graduation, or both. To address this challenge, I implement a bounding exercise to estimate the effect of attendance on seniors: (i) assuming a homogeneous impact of attendance across all grades (upper bound) and (ii) assuming attendance only has an impact on seniors (lower bound). This choice of assumptions is motivated by the grade heterogeneity estimates presented in the cancellation design section of this paper, which suggest that the impact of attendance is monotonically increasing by grade.

The lower and upper bound estimates of conference impacts on senior attendees for equation 3.3 are calculated as follows:

$$\underline{\theta} = \frac{\hat{\beta}_2}{\overline{Mean_Seniors_Attend_{jt}}} \tag{3.5}$$

and

$$\overline{\theta} = \frac{\hat{\beta}_2}{(\overline{Mean_Seniors_Attend_{jt}} + \overline{Mean_Fresh_Attend_{jt}}) - \overline{Mean_Soph_Attend_{jt}}} \tag{3.6}$$

where $\hat{\beta}_2$ is estimated from equation 3.3. $\underline{\theta}$ is calculated by attributing cohort differences entirely to the level of senior attendance. $\overline{\theta}$ is calculated by treating all attendance equally and attibuting differences in outcomes to the difference between senior and sophomore cohort attendance. (Note that senior cohort attendance also includes freshman attendance, because the conference is held every three years and seniors were eligible to attend as sophomores.)

Likewise, for equation 3.4 the bounds are calculated as:

$$\underline{\theta'} = \frac{\hat{\beta}_3 * \overline{Mean_Total_Attend_{jt}}}{\overline{Mean_Seniors_Attend_{jt}}} \tag{3.7}$$

and

$$\underline{\theta'} = \frac{\hat{\beta}_3 * \overline{Mean_Total_Attend_{jt}}}{(\overline{Mean_Seniors_Attend_{jt}} + \overline{Mean_Fresh_Attend_{jt}}) - \overline{Mean_Soph_Attend_{jt}}} \tag{3.8}$$

where $\hat{\beta}_3$ is estimated from equation 3.4.

### 3.5.2 Cohort-Level Sample and Descriptive Statistics

For comparability, the cohort-level analysis uses the same sample as the short-run analysis: 279 large schools with graduating cohorts from 2003-2017. The "Cohort Level (Graduation Year Level) Outcomes" panel of Table 3.10.1 presents select cohort-level outcomes. College-graduation-year cohorts that graduated during a conference year have higher levels of giving in the three years following graduation than those that did not, as well as higher levels of lifetime staff employment.[14]

Figure 3.10.1 provides visual intuition for the cohort-level design. Panel A presents cyclical trends in future staff employment by cohort for all students with an available graduation year and panel B presents trends in future giving, respectively, where each point represents the outcome for a certain graduating year on the x-axis. The figure compares sophomore and senior graduating cohorts for visual simplicity. With minimal exceptions, the number of student-conference-year attendees oscillates such that conference-year seniors (who were able to attend either senior or freshman year) consistently have greater attendance than adjacent conference-year sophomores. Similarly, giving and staff employment outcomes follow an oscillating sawtooth pattern as well, typically with greater values for conference-year senior cohorts. These trends are suggestive that either frequency of attendance, timing of attendance relative to graduation, or some combination of both affect student decisions. I formalize the intuition of Figure 3.10.1 by estimating the cohort-level empirical specifications in the section to follow.

### 3.5.3 Cohort-Level Results

Table 3.10.3 presents estimates of conference effects on cohort-level outcomes. Similar to short-run estimates, panel A estimates differences in outcomes for cohorts by graduation year timing relative to conference year, whereas panel B estimates differences in outcomes

[14] I focus on lifetime staff employment because it is a more rare outcome and most initial staff employment still occurs in the years following graduation.

using graduation timing interacted with college attendance propensity.

Columns 1 and 2 of Table 3.10.3 panel A show that giving in the three years following graduation (June of graduation year $t$ through May of year $t+3$) is significantly greater for conference-year seniors relative to conference-year sophomores, approximately \$801 per college-cohort relative to a mean of \$5,992. Panel B shows this estimate in per-attendee terms, which is robust to the inclusion size-by-conference year controls. I note two important observations: First, estimates of three-year giving do not include the offering taken at the conference itself and are truly long-run impacts. Second, effects of conference attendance on giving appear to increase monotonically by grade, but the junior cohort is not significantly different from the sophomore cohort.

Column 3 of Table 3.10.3 panel B shows that staff employment increases for conference-year seniors relative to conference-year sophomores per each additional college conference attendee, but that this estimate loses precision (albeit maintains its magnitude) after including club size controls. The estimates are suggestive that the conference increases lifetime staff participation but are not statistically significantly from zero.

Applying the bounding exercise to estimate the impact per senior attendee yields the following estimates for giving:

$$[\underline{\theta}_g, \overline{\theta}_g] = [\$137, \$213]$$

$$[\underline{\theta'}_g, \overline{\theta'}_g] = [\$153, \$239]$$

and for staff employment:

$$[\underline{\theta}_s, \overline{\theta}_s] = [0.0054, 0.0084]$$

$$[\underline{\theta'}_s, \overline{\theta'}_s] = [0.0087, 0.0151]$$

A rough interpretation of these estimates is that every 100 additional senior attendees is associated with approximately \$17,500 in additional giving in the three years following graduation and approximately one additional staff worker hired by the sponsor organization.

## 3.6 Long-Run Effects

There are at least three barriers to estimating the impact of gathering attendance on person-level outcomes. First, in most settings microdata of attendees are not available or difficult to identify. Second, there is not always a clear comparison group of non-attending individuals. Third, even in the presence of a comparison group, the association between attendance and future outcomes may be confounded by unobserved characteristics such as interest in the theme of the gathering.

I overcome each of these challenges using institutional features from the Urbana missions conference. First, linked individual identifiers on all attendees are recorded by the sponsor organization, allowing one to measure outcomes of attendees over time that would be infeasible to estimate using population-level data. Second, the Urbana missions conference creates individual identifiers for not just attendees but any student who *registers* but does not attend the conference. This group of non-attending registrants constitutes a reasonable comparison group. Finally, I observe (1) a rich set of pre-determined observable characteristics associated with religiosity and organizational commitment, and (2) weather-related flight cancellations during conference commuting days, which allow for two separate strategies of estimating the long run effect of conference attendance on student life trajectories.

### 3.6.1 Long-Run Empirical Strategy

The first estimation strategy compares outcomes for attending and non-attending registrants, controlling for a rich set of observable pre-conference characteristics:

$$Y_{ist} = \beta_0 + \beta_1 Attended_{ist} + \beta_2 X_{ist} + \gamma_t + \epsilon_{ist}, \tag{3.9}$$

where $Y_{ist}$ is the outcome for pre-registered student $i$ from college state $s$ in year $t$, $Attended_{ist}$ is a binary indicator for whether the student attended the conference or not, $X_{ist}$ is a rich set of individual-level controls for past involvement involvement with the sponsor organization,[15] $\gamma_t$ are conference year fixed effects, and $\epsilon_{ist}$ is the random error term.

[15] In the baseline specification, these controls include: past amount of giving, past event and mission trip

Standard errors are clustered at the college level. If attendance is uncorrelated with (unobserved) religiosity or other unobserved determinants of $Y_{ist}$ after conditioning on observed demographics and behavior $X_{ist}$, then $\beta_1$ can be interpreted as the causal effect of attendance on student outcomes. In the results section, I show and discuss why this assumption is plausible.

The second estimation strategy compares outcomes for registered students who faced differing levels of bad weather during expected planning and commuting days to the conference:

$$Y_{ist} = \beta_0 + \beta_1 ShareCancelled_{st} + \theta^d PlaceboCancelled^d_{st} + \beta_2 X_{ist} + \gamma_s + \gamma_t + \epsilon_{ist}, \quad (3.10)$$

where $Y_{ist}$ is the outcome for pre-registered student $i$ from college state $s$ in year $t$, $X_{ist}$ is a set of basic demographic controls. Standard errors are clustered by state of enrolled college.

$ShareCancelled_{st}$ is a continuous variable measuring the share of flights from December 25-27th (the primary planning and commuting days for the conference) that were cancelled due to weather. Cancelled flights are defined as the share of all flights departing from airports in a student's college's state and arriving at St. Louis or Chicago (the two primary hubs for the conference) that are cancelled due to weather.[16] $\beta_1$ therefore estimates the effect of flight cancellations on student outcomes. State fixed effects ($\gamma_s$) and conference year fixed effects ($\gamma_t$) control for persistent geographic differences in weather and year-to-year weather shocks, respectively, each of which may be correlated with outcomes through a confounding variable, for example, if bad weather in a particular year happens to coincide with an increase in disposable income used for giving.

$\beta_1$ can be interpreted as being mediated through conference attendance and/or experience if one assumes that weather-related flight cancellations do not affect student outcomes through any other mechanism. One way to examine this assumption is through placebo

attendance, indicators for gender and race/ethnicity, college fixed effects, detailed student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. These indicators are derived primarily from the conference registration form and partially from the graduating student form.

[16] In practice, Delaware college students are matched to Pennsylvania airports and Washington, D.C. college students are matched to Virginia airports.

variables for flight cancellations during other days of December, $PlaceboCancelled_{st}^{d}$. Under a hypothesis that flight cancellations do not affect outcomes independent of the conference, we expect estimates of these daily indicators, $\hat{\theta}^{d}$, to not be significantly different from zero.

### 3.6.2 Long-Run Sample and Descriptive Statistics

The sample consists of undergraduate students at U.S. universities who registered for any of the six conferences from 2000 to 2015.[17] I omit remaining registrants who were likely miscoded as students: registrants listed as a member of a graduate student chapter, registrants with past staff employment, and registrants who previously attended a faculty conference. Students are excluded who cancelled their enrollment five or more days prior to the conference, or who registered within two days of the conference start. Finally, I exclude a handful of students with missing gender because they are less likely to be matched to other administrative records.

Table 3.10.4 presents characteristics of the sample of undergraduate conference registrants. From 2000-2015, approximately 51,313 students registered for the conference at least three days in advance, of whom 48,596 attended (column 1) and 2,717 did not attend (column 2).[18] Representatives of the sponsor organization consider the attendance indicator to be an accurate indicator of student participation in the conference, as students must sign in on site to receive a badge needed to enter the conference. Although reasons for cancellation or non-attendance ("no-show") are not specified in administrative data, event organizers cite transportation and illness as two reasons for non-attendance.

Comparing outcomes for attendees and non-attendees in Table 3.10.4, attendees show markedly greater future participation with the sponsor organization. 40 percent of attendees give at least one donation during the conference period or afterward, compared to

[17] Approximately 1,082 registrants in 2015 (12 percent of the cleaned sample for the year) did not enter a university or state identifier and are dropped from the analysis.

[18] Observations are technically defined at the student-conference-year level, so some students may appear twice if they attended as a freshman and as a senior. For the purposes of the analysis, I consider these as distinct observations.

only 19 percent of non-attendees. Attendees give on average $2,230 in the future compared $724 for non-attendees. Attending students are approximately four times more likely to join organizational staff in the future (2.7 percent of attendees become employed as staff) compared to non-attendees (0.7 percent). Finally, attendees are also more likely to participate in an overseas short-term mission trip organized by the sponsor organization. The difference in outcomes shows that, even conditional on registration, conference attendance is highly predictive of future involvement with the sponsor organization. However, it is not clear if attendance increases participation or if the difference between the two groups is entirely explained by pre-existing underlying differences between students who choose to attend and students who choose not to attend.

The "Registration and Conditions" panel provides further context for conference details. Among attendees, 2.2 percent of college-state flights are cancelled during typical commuting days, compared to 2.9 percent for non-attendees. This difference suggests that at least a fraction non-attendance may be due to disruptive weather. Attendees pay on average $379 dollars to attend the conference, which is typically computed as the list price minus any subsidies provided by the sponsor organization. This amount, independent of travel costs, shows that conference participation is a non-trivial investment for students. The registration fee is not considered a donation for the purposes of analysis. Non-attendees pay on average $113, which may reflect refunds. However, payment data are not recorded consistently with regard to list price and price paid and may be mechanically set to zero for some non-attendees, even if the student did not receive a refund. Organizational representatives note that blanket refunds are not a typical policy.

Finally, the "Student Involvement" and "Demographics" panels compare characteristics of attendees and non-attendees. Attendees are more likely to have a student leadership role [19] and have similar levels of any past giving, previous conference attendance (during

[19]Student leadership information is technically collected after the conference around the time of graduation, but, given the timing of club leadership elections, would be pre-determined for all seniors and a large fraction of juniors in the sample.

high school or freshman year of college), and sponsor-organization mission trip experience as non-attendees. While both groups are majority female, non-attendees are more likely to be Non-Hispanic Black or Hispanic, and less likely to be Non-Hispanic white or Asian. This difference suggests that attendees and non-attendees may differ along important demographic dimensions that could affect future participation. Both groups are similar in terms of the distribution of college grade level, with the number of registrants increasing monotonically by grade.

### 3.6.3 Long-Run Results

Table 3.A.6 provides details about the balance of the OLS with controls design in model 3.9 and the flight cancellation design in model 3.10. Column 1 presents the coefficients and joint F-statistic from a regression of attendance on indicators for past involvement (past attendance, giving, and mission trip participation), conditioning on demographic characteristics, student leadership roles, and days registered in advance of the conference. The $p$-value of the F-statistic is 0.29, suggesting that attendance is orthogonal to past indicators of involvement conditional on salient observable characteristics. This result lends credibility to a causal interpretation of the coefficient on attendance in equation 3.9.

Table 3.A.6 columns 2-5 provide more information about flight cancellation variation. The goal of the weather-related flight cancellation design is to provide variation in attendance that is independent of underlying religiosity and other unobserved interest in the sponsor organization. Column 2 presents a balance test by regressing the share of flight cancellations on observable student registrant characteristics. As expected, I find no systematic association between past involvement and flight cancellations, and the joint F-statistic on observable characteristics is 0.52 with a $p$-value of 0.88. This exercises supports the assumption that flight cancellations provide a source of variation independent of unobserved characteristics as well.

Column 3 presents the effect of flight cancellations on conference attendance. I find

that a 10 percentage point increase in weather-related flight cancellations departing from a student's college's state is associated with a 2.2 percentage point reduction in the odds of attendance, with an F-statistic of 17.82 ($p < 0.001$). Furthermore, we find no systematic association between placebo date flight cancellations and conference attendance.

It is reasonable to consider estimating an instrumental variables design with flight cancellations as an instrument for attendance, given that flight cancellations appear exogenous (column 2) and exhibit a strong first stage (column 3). Column 4, however, presents a possible violation of the exclusion restriction, which requires that flight cancellations only affect outcomes through attendance. In particular, column 4 shows that an increase in flight cancellations is associated with a reduction in giving *even when excluding the sample to attendees only.* This finding is consistent with one or more hypotheses: (i) Bad commuting weather affects the dosage of treatment by delaying a student's travel and reducing the number of days a student participates at the conference, (ii) bad commuting weather negatively affects a student's conference experience, and (iii) the attendance variable is coded with error. Given strict protocol around recorded attendance to enter the conference, hypotheses (i) and (ii) are most plausible. Table 3.A.6 Column 5 shows that a traditional instrumental variables strategy estimates extremely large effects of event attendance that are an order of magnitude greater than unconditional comparisons of means in Table 3.10.5 column 1. For the remainder of the paper, therefore, reduced form regressions of outcomes on flight cancellations are presented. Table 3.A.6 provides evidence that this effect of flight cancellations is mediated through a student's experience of the missions conference, but that there are dimensions of participation that may not be adequately interpretable using an instrumental variables strategy with a binary attendance indicator. This observation is also relevant and potentially cautionary to other empirical contexts that use weather as an explanatory variable for event attendance (e.g., Hungerman and Moorthy 2020).

Figure 3.10.2 presents the visual intuition for the flight cancellations design. Panel A presents first stage regression estimates of attendance on flight cancellations, also shown in

Table 3.A.6 column 3.[20] The x-axis presents flight cancellations by day of December, and the y-axis presents the regression coefficient and standard error. One can see that registrant attendance falls for flights cancelled during days when preparing for or commuting to the conference.[21] However, cancellations on placebo days are generally not statistically significant and not systematically different from zero. Similarly, panel B presents the reduced form estimates of future giving on flight cancellations, similar to that in Table 3.10.5 column 3. Flight cancellations during conference commuting days are again associated with a statistically significant reduction in future giving.

Table 3.10.5 estimates the effect of conference attendance on future giving and staff employment using models 3.9 and 3.10. Columns 1 and 4 present estimates from OLS regressions with basic demographic controls and state and year fixed effects, columns 2 and 5 present estimates controlling for a detailed set of observed characteristics defined in the empirical strategy section, and columns 3 and 6 present reduced form estimates of the effect of flight cancellations on the dependent variable.

Column 1 shows that student attendance is associated with a $1,328 increase in future giving relative to a mean of $2,150. After controlling for detailed student characteristics such as leadership roles, past giving, and demographics, the point estimate falls to a still significant $482 (column 2). This set of results shows that a simple comparison of means (column 1) is confounded by underlying differences between attendees and non-attendees. However, given the richness of the observed covariates in column 2 and the balance in attendance on select predetermined characteristics in Table 3.A.6, the point estimate in column 2 has a more plausible causal interpretation. The estimates of equation 3.9 with full controls suggest that attendance is associated with a $482 (22 percent) increase in long-run giving (column 2) and a 0.9 percentage point (36 percent) increase in the likelihood of future staff employment (column 5). These percentage changes are of similar magnitudes to the short-run effects

[20] Figure 3.A.1 provides a binned scatter visualization of the first stage without placebos.

[21] I pool December 25-27 because some students commute by car or bus over multiple days and weather a few days prior to the conference may affect travel plans.

presented in Table 3.10.2.

Columns 3 and 6 provide supporting evidence of conference attendance effects by regressing outcomes on flight (non-)cancellations. Note that the regressor is defined here as share of flights *not* cancelled, so that supporting point estimates would be of similar sign to other columns. Column 3 shows that a ten percentage point reduction in flight cancellations from a student's college's state is associated with a $216 increase in long-run giving, a strong result considering that many students attend college out of state. Column 6 suggests that a ten percentage point reduction in flight cancellations is associated with a 0.5 percentage point increase in the probability of joining staff. However, this latter result is imprecisely estimated and not statistically significantly different from zero, a drawback of the flight cancellations design.

Taken together, the rich controls and flight cancellation designs of Table 3.10.5 provide supporting evidence that conference attendance affect not just short-run student outcomes but longer-run trajectories as well. These findings still hold when excluding students who did not have a first-time donation as the conference itself (Table 3.A.7), providing supporting evidence that the association between attendance and future giving is not caused by unforeseen differences in data linkage due to giving at the conference. Conference attendance is also associated with increases in other dimensions of giving and staff affiliation as well (Table 3.A.8), albeit the flight cancellation results suffer for statistical imprecision. Finally, conference attendance is also associated with an increase in future organization-sponsored mission trip participation and future conference attendance (Table 3.A.9), showing that participation in one event may create a feedback loop for participation in future events.

### 3.6.4 Dynamic Trends

Having estimated long-run effects of conference attendance, I now briefly visualize dynamic trends in giving and staff employment. Figures 3.10.3 and 3.A.2 provide a descriptive comparison of long-run outcomes for registrants who attended compared to registrants attendees

who did not attend. Rates of giving diverge during the quarter of the conference, and this is due to the differential giving at the conference itself. From there, giving again diverges two quarters later (which is the a quarter of graduation for seniors) and remains persistently higher over the following 30 quarters. A similar pattern emerges for post-graduation staff employment.

For a more plausibly causal visualization of dynamic trends following attendance, the following specification is estimated:

$$Y_{it} = \beta_0 + \sum_{k=-9}^{19} \mathbb{1}\{K_{st} = k\}(\gamma_k Attend_i + \rho_k Score_i) + \mathbb{1}\{K_{st} \leq -4\}(\gamma_{-10} Attend_i + \rho_{-10} Score_i) + \mathbb{1}\{K_{st} \geq 3\}(\gamma_{20} Attend_i + \rho_{20} Score_i) + \beta_1 X_i + \lambda_t + \epsilon_{it}, \tag{3.11}$$

where $Y_{it}$ is the outcome for student-by-conference-year attendee $i$ in quarter $t$, $X_{ist}$ is a vector of student school and grade characteristics, $\lambda_t$ are quarter-by-year fixed effects to control for shared calendar time trends in outcomes, and $\epsilon_{it}$ is the error term with standard errors clustered at the student level. The coefficients of interest are $\gamma_k$, which are event time coefficients for attendee outcomes relative to non-attendees relative to the quarter of the conference. To parsimoniously control for underlying differences between attendees and non-attendees, I estimate a predicted level of future giving using pre-conference characteristics, created using a separate regression on the full sample of registrants. $\rho_k$ are event time coefficients scaled by predicted future giving and used as control variables in the model.

Figure 3.A.3 visually presents the estimation of the event study equation 3.11 of attendance on giving over time. One observes approximately parallel trends prior to the conference, followed by a ten dollar increase in giving per attending student during the conference quarter, and a persistently greater level of giving thereafter with the gap in giving increasing over time.

The dynamic findings lend support to the conclusion that conference attendance has long-run impacts on student behavior. It is important to consider several institutional details, however, for the sake of external validity to other large social gatherings. First, for at least

a fraction of years, conference attendees are given marketing promotions for future Urbana conferences that non-attendees are not. These follow-up promotions are effectively part of the attendance treatment that go beyond the conference itself and could help sustain differences between the treatment and control groups. Second, some donations are recurring, which could contribute to persistent differences in giving over time where the conference itself acts as a nudge and catalyst for short-run behavior (Thaler and Benartzi 2004).

## 3.7 Mechanisms and Heterogeneity

We have thus far shown that the Urbana missions conference has economically significant impacts on student behavior during and beyond graduation. We now turn to consider *for whom* the conference has an impact and what it can tell us about the gathering's mechanisms of influence.

Who is impacted most by gatherings? One theory regarding the role of large gatherings is that they influence the behavior of younger attendees whose preferences are more malleable (Hungerman and Moorthy 2020). On the other hand, effects of gatherings may be stronger for people at life transition points where messaging may be particularly salient, in this case older students close to graduation. I test these potentially competing hypotheses by estimating regression specifications by grade:

$$\begin{aligned} Y_{ist} = \beta_0 &+ \beta_1(Grade1_{ist} \times Attend_{ist}) + \beta_2(Grade2_{ist} \times Attend_{ist}) \\ &+ \beta_3(Grade3_{ist} \times Attend_{ist}) + \beta_4(Grade4_{ist} \times Attend_{ist}) + \beta_5 X_{ist} + \gamma_t + \epsilon_{ist}, \end{aligned} \tag{3.12}$$

and a flight cancellation specification:

$$\begin{aligned} Y_{ist} = \beta_0 &+ \beta_1(Grade1_{ist} \times ShareCancelled_{st}) \\ &+ \beta_2(Grade2_{ist} \times ShareCancelled_{st}) + \beta_3(Grade3_{ist} \times ShareCancelled_{st}) \\ &+ \beta_4(Grade4_{ist} \times ShareCancelled_{st}) + \beta_5 X_{ist} + \gamma_s + \gamma_t + \epsilon_{ist}, \end{aligned} \tag{3.13}$$

which are identical to equations 3.9 and 3.10, respectively, but now with the regressor of interest split by each undergraduate grade level: freshman, sophomore, junior, and senior.

Figure 3.10.4 presents first stage (panel A) and reduced form (panel B) estimates of the impact of flight cancellations on attendance and giving, respectively. Panel A shows that the first stage effect on attendance is of comparable magnitude and at least marginally statistically significant for all grade levels. However, panel B shows that the reduced form point estimates of the (negative) impact of weather on giving are monotonically increasing by grade level; near zero for freshman attendees and highly significant for senior attendees.

Table 3.10.6 provides further point estimates of grade heterogeneity in impacts on giving and staff employment. Columns 1 and 4 presents estimates of equation 3.12. Column 1 shows that the association between attendance and giving is largest for senior attendees, even after controlling for level of club involvement and other predictors of future giving. These estimates are consistent with estimates of the flight cancellation design in equation 3.13 (column 2, visualized in Figure 3.10.4, and column 3 with additional controls for past involvement). Similarly, column 4 presents an increasing relationship between attendance and staff employment by grade. The flight cancellation design is underpowered to detect changes in staff employment (columns 4 and 5), yet point estimates still monotonically increase by grade with the exception of senior attendees.

The presented estimates suggest that conference attendance is most impactful for older students, particularly undergraduate seniors, even independent of level of future involvement in the sponsor organization. These findings support the hypothesis that strategic investments at certain margins of life change can dominate age-specific experiences that are particularly influential for younger individuals.

The composition of conference attendees includes individuals who are self-affiliated with the sponsor organization ("in-group") and a significant fraction of students who are not ("out-group").[22] Past work on the study of religion in particular has found that effects of identity shocks are strongest along the intensive margins, that is, increasing the intensity of affiliation for current in-group members rather than changing the affiliation of out-group

[22]These other students are usually members of other Christian organizations or friends and family invited by attendees.

members along the extensive margin (Sinding Bentzen 2019). Does the effect of conference attendance differ by level of involvement in the sponsor organization?

Table 3.A.10 presents estimates of attendance impact heterogeneity by several margins of sponsor organization involvement: (i) self-reported membership on the conference registration form, (ii) students who filled out a departing (graduating) student contact form for the sponsor organization, and (iii) students ever recorded to be a student leader for a club affiliated with the sponsor organization. Columns 1-3 and 7-9 present estimates from OLS regressions controlling for a detailed set of observed characteristics, where the attendance indicator is interacted with an involvement indicator. Columns 4-6 and 10-12 present reduced form estimates of the effect of flight cancellations interacted with involvement indicator on the dependent variable.

Table 3.A.10 shows that the association of attendance (or flight cancellations) on outcomes is strongest for students already involved with the sponsor organization, along all three dimensions of involvement. Columns 3 and 9 show that the coefficient on attendance is particularly strong for student leaders: Attendance among student leader registrants is associated with a $4,260 increase in future giving compared to $187 among non-leaders and a 6.6 percentage point increase in future staff employment compared to a 0.8 percentage point increase for non-leaders. These findings are consistent with the flight cancellation results in columns 6 and 12, which lend support to a causal interpretation. Due to the relatively small number of student leader non-attendees in the sample ($n = 392$), student leaders are defined as a relatively broad category constituting 25 percent of the sample, consisting of students with any responsibilities beyond that of an ordinary member. The large point estimates despite this broad classification suggests that attendance may be even more impactful for the most involved leaders.

An overarching conclusion from these results is that mission conference attendance increases commitment intensity moreso than bringing in new supporters, from the perspective of the sponsor organization. However, it is entirely possible that non-affiliated students

increased giving and labor supply for *their own*, non-sponsor organizations such as other religious clubs and churches.[23] I cannot observe such outcomes in the available data. The relationship between age, commitment, and intervention impact suggest a tantalizing trade-off: it may be privately efficient for individual mission-driven organizations to invest in already-committed members, but potentially socially inefficient in contexts where investments in younger (and less committed) people yield greater commitment to a broad cause area. Additionally, the results presented here leave upon the possibility that conferences could provide positive or negative externalities for giving to similar types of organizations that are not the sponsor organization. These potential tradeoffs and coordination problems between non-profit organizations are a ripe area for future research.

## 3.8 The Effect of Messaging on Representation

Mission-minded organizations may seek to improve the representation of people from unrepresented backgrounds among stakeholders, employees, and leaders. Does group-salient messaging increase minority student giving and staff employment relative to other student attendees?

I consider the effect of targeted messaging on student outcomes by leveraging the Urbana mission conference's surprise 2015 promotion of Black Lives Matter. The St. Louis location of the 2015 conference was located 13 miles from Ferguson, MO, where Michael Brown was killed by police in August 2014. Beginning on the first full day of the conference, members of the Urbana worship team wore Black Lives Matter shirts, played gospel music, and Reverend Michelle Higgins, a Black woman and St. Louis church worship leader who founded the Faith for Justice coalition of Christian activists, delivered a plenary message on race and justice.[24] The conference's Black Lives Matter messaging generated considerable news coverage and

[23] For example, the effect of weather on attendance is stronger for non-affiliated students.

[24] "Black Lives Matter is not a mission of hate. It is not a mission to bring about incredible anti-Christian values and reforms to the world," Higgins said. "Black Lives Matter is a movement on mission in the truth of God." https://sojo.net/articles/intervarsity-s-unabashed-support-blacklivesmatter-may-be-its-boldest-move-yet

polarized responses among sponsor organization donors.[25]

The effects of messaging are estimated using difference-in-difference (DD) using the sample of all registrants:

$$Y_{ist} = \beta_0 + \beta_1 Black_{ist} + \beta_2 2015_t + \beta_3 (Black_{ist} \times 2015_t) + \beta_4 X_{ist} + \epsilon_{ist}, \tag{3.14}$$

where $Black_{ist}$ is an indicator for whether student $i$ from college $s$ in year $t$ indicated a race-ethnicity of Black or African American on the conference registration form and $2015_t$ is a year indicator for the messaging period. $X_{ist}$ is a vector of individual controls as included in previous specifications. The coefficient on the interaction of the two variables, $\beta_3$, is the coefficient of interest: The difference in outcomes for Black attendees relative to white attendees in 2015, relative to the difference between the two groups in 2000-2012. The relatively strong assumptions required for a causal interpretation of $\beta_3$ in relation to the question of interest are: (i) in the absence of changes at the 2015 conference, race-ethnic gaps would have remained proportional, and (ii) changes in race-ethnic differences in 2015 are entirely attributable to BLM messaging and not other conference features.

I further estimate a triple difference-in-difference (DDD) specification which compares Black attendees versus registrant non-attendees in 2015 to non-Black differences in the same year, relative to group-attendee differences in prior years:

$$\begin{aligned} Y_{ist} = \beta_0 + \beta_1 Black_{ist} + \beta_2 2015_t + \beta_3 Attend_{ist} + \beta_4 (Black_{ist} \times 2015_t) + \beta_5 (Black_{ist} \times Attend_{ist}) \\ + \beta_6 (2015_t \times Attend_{ist}) + \beta_6 (Black_{ist} \times 2015_t \times Attend_{ist}) + \beta_7 X_{ist} + \epsilon_{ist}, \end{aligned} \tag{3.15}$$

where $\beta_6$ is the triple-difference coefficient of interest.

I estimate specifications 3.14 and 3.15 as well as provide descriptive evidence in changes in Black and non-Black registrant giving over time to explore potential trends and messaging

[25] c.f. https://www.nytimes.com/2016/01/23/us/some-evangelicals-struggle-with-black-lives-matter-movement.html. Opposition to Black Lives Matter messaging included disagreement from some conservative Christians about goals of the Black Lives Matter movement, including advocacy for transgender and gay communities, perceived hostility toward law enforcement, and other political alliances that are associated with Democratic political goals. Further opposition related to a religious organization endorsing a perceived faith-independent political position, questioning the message's relevance to a global missions conference, and misinterpreted statements about other social issues.

backlash. For this section of the analysis giving outcomes have been restricted to the three years following graduation due to potential long-run effects of messaging on alumni donations.

Figure 3.A.5 presents descriptive trends of the mean three-year post-graduation giving amounts for Black and non-Black conference registrants, by conference year. There is a generally increasing trend in recorded giving for both groups of students, with a slight decline for both groups in 2015. Triple-difference and difference-in-difference estimates (presented in Table 3.A.11 and visually by group in Figure 3.A.6) show no significant change in Black student giving relative to non-Black student giving in 2015, nor a within-group difference for attendees relative to non-attendees. Similarly, Figures 3.A.7 and 3.A.8 show no unusual break in trend for staff employment outcomes.

The findings suggest that the salience of group-specific messaging surprisingly had no discernible effect on Black student outcomes relative to other students. The results showcase the limits of group-specific messaging in improving the representation of individuals from underrepresented backgrounds and/or with other background characteristics: the supply of such messaging is not sufficient to overcome other structural barriers to giving and employment. Future work could consider the effect of highly salient endorsements on alumni donor behavior, and the effect of mentorship on diversity in mission-driven organizations.

## 3.9 Conclusion

Large gatherings are ubiquitous in most societies, yet their effects on participants are still relatively unknown. This paper uses novel linked administrative data to show that the Urbana missions conference has had both short-run and enduring impacts on college student giving, career choices, and other salient outcomes. These impacts are largest for students who are already most involved with the organization and close to graduation, which suggests that the *timing* and *context* of social investments are as important as the content of the investment itself. These findings are of particular relevance to public, nonprofit, and other

mission-driven organizations that are concerned with recruiting employees and donors, or changing public behavior.

Despite the broad impacts of conference attendance, there is no evidence of impact of group-specific messaging, particularly Black Lives Matter messaging, on students from underrepresented backgrounds. This finding suggests that advancing opportunities and buy-in for prospective stakeholders from underrepresented groups may require more deliberate and concerted investments, such as mentorship programs. The role of messaging on *alumni* giving and the impact of mentors in religious contexts remain relatively unexplored areas of inquiry.

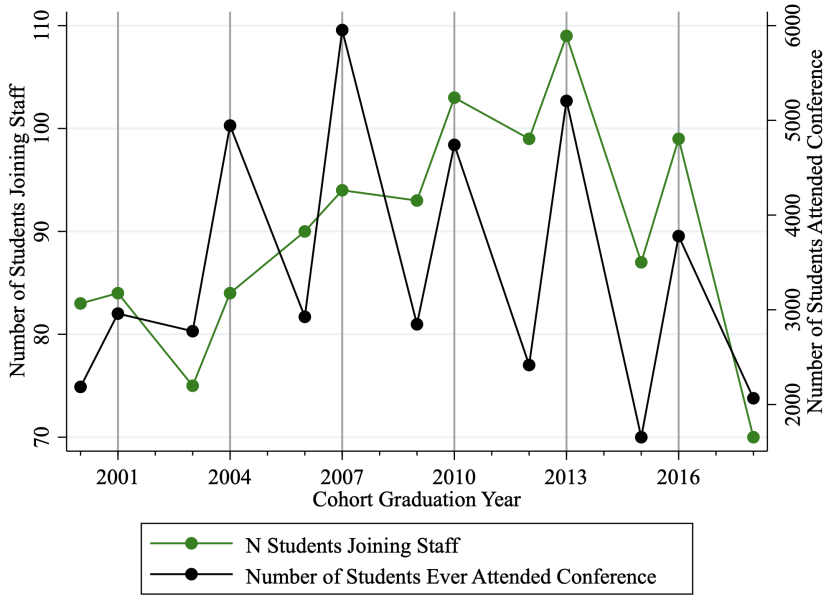
In the years following the first wave of the COVID-19 pandemic, large gatherings have begun to re-emerge together with virtual programming.[26] A promising area for future work is to understand the impact of virtual alternatives to in-person events, and whether improved flexibility provided by virtual options outweighs potential negative externalities with respect to in-person engagement. An additional area for future research is to consider contexts where gatherings may be an important input in career and lifestyle decisions, such as graduate student participation in the AEA annual conference and various lifestyle movements.

[26]Urbana itself has had limited virtual programming since 2015.

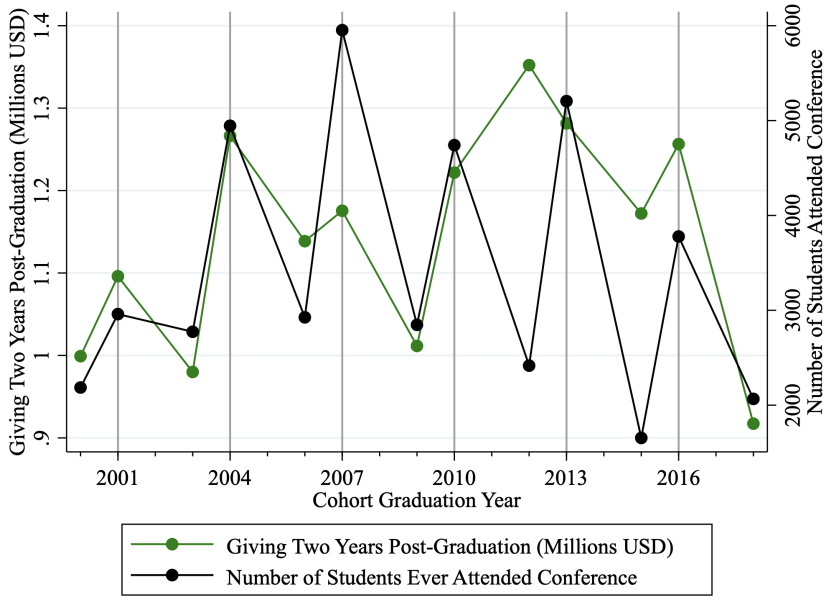
## 3.10 Figures and Tables

Figure 3.10.1: Descriptive Cyclical Patterns of Conference Attendance and Outcomes by Graduation Cohort

(1) Panel A: Attendance and Future Staff Employment by Cohort




(2) Panel B: Attendance and Future Giving by Cohort

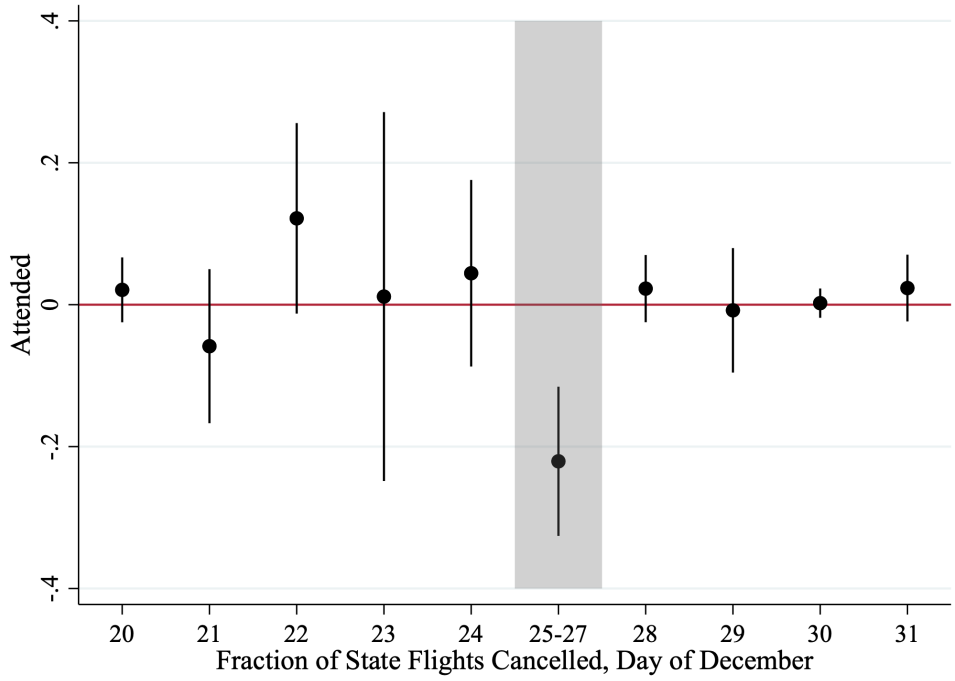



Notes: This figure illustrates the descriptive relationship between a cohort's total conference attendance prior to graduation, the number of students in a cohort joining the organization's staff, and the total amount of giving from a cohort giving within two years of graduation. Each tick on the x-axis represents a cohort's graduation year, and each solid vertical line represents a cohort of students who were seniors

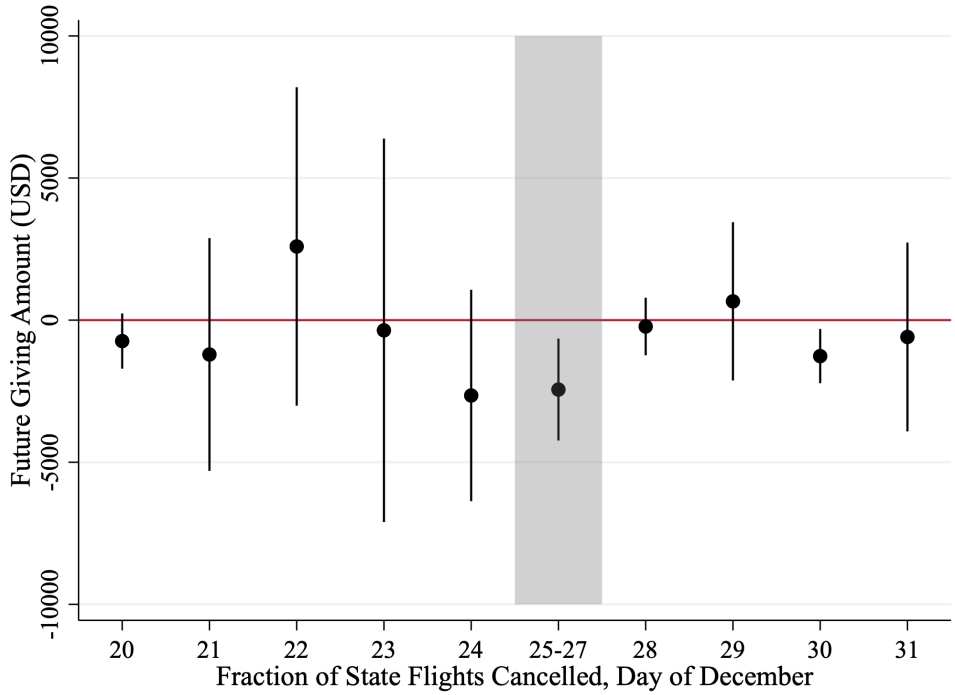
during a conference year. The cyclical pattern of attendance across cohorts is the result of the rotational timing of the conference. Because the conference occurs every three years, conference-year seniors have two opportunities to attend the conference before graduation: as seniors and as freshmen. The attendance shown in this figure includes double-counting students who attended the conference twice during college. Cohorts in this figure have been restricted to students who were seniors or sophomores (not juniors) during a conference year, for clarity of presentation.

Figure 3.10.2: Flight Cancellations, Conference Attendance, and Outcomes

(1) Panel A: Conference Attendance (First Stage with Placebos)



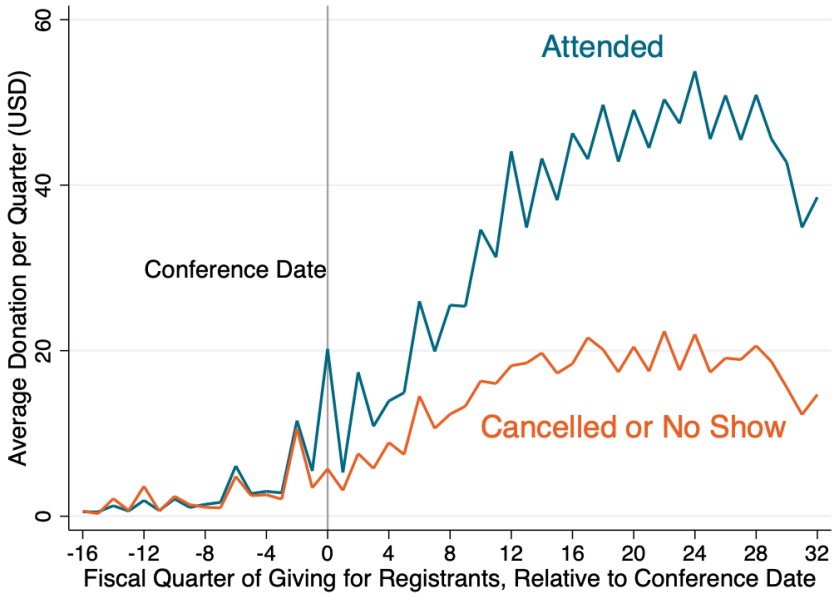

(2) Panel B: Future Giving (Reduced Form with Placebos)




Notes: This figure shows the coefficients and 95 percent confidence intervals of regressions of a student's conference attendance and future giving, respectively, on the share of daily weather-related flight cancellations for each date in late December. Flight cancellations are calculated for major airlines departing a state's

airports and arriving in St. Louis or Chicago. A student's state is defined as the state of the student's college. December 25-27 are typical planning and commuting days for travel to the conference, which occurs December 27-31. Standard errors are clustered at the state level.

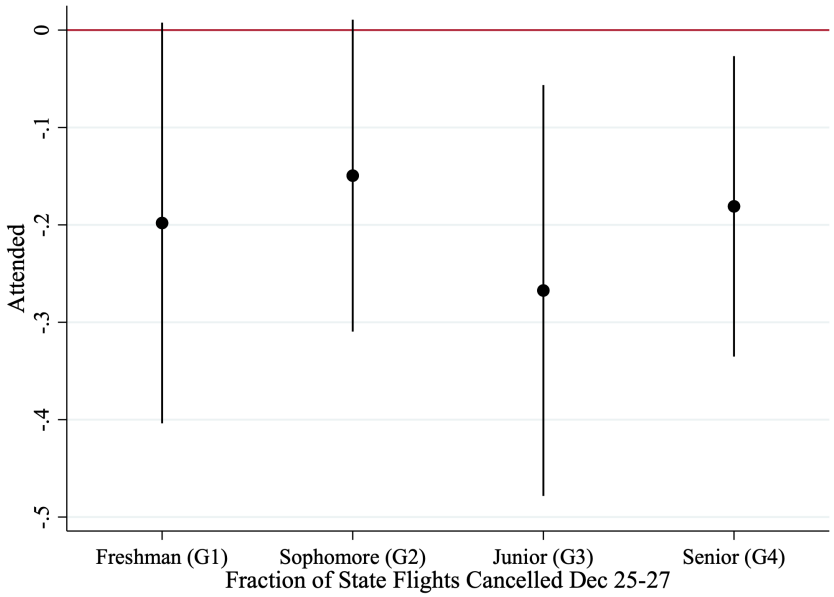
Figure 3.10.3: Descriptive Long-Run Giving Trajectories of Attending vs. Non-Attending Registrants



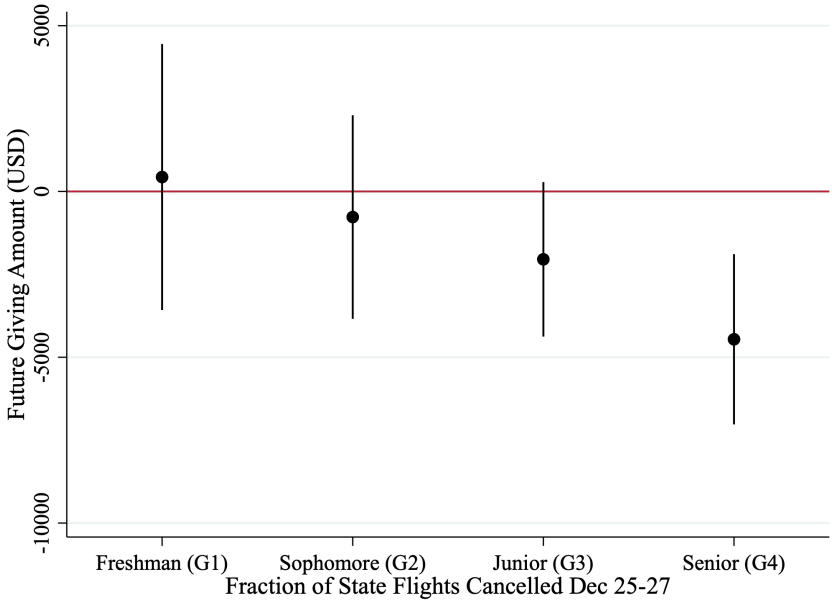

Notes: This figure shows the dynamic trend of quarterly donations from undergraduate students who registered and attended the conference compared to undergraduate students and registered but did not attend. Sample restricted to student-registrant observations from 2000-2012.

Figure 3.10.4: Flight Cancellations and Outcomes: Heterogeneity by Grade

(1) Panel A: First Stage Effect of Weather on Attendance, by Grade




(2) Panel B: Reduced Form Effect of Weather on Giving, by Grade




Notes: This figure presents the coefficients and 95 percent confidence intervals of regressions of a student's conference attendance and future giving, respectively, on the share of daily weather-related flight cancellations from December 25-27 interacted with an indicator for student grade. Flight cancellations are calculated for major airlines departing a state's airports and arriving in St. Louis or Chicago. A student's state is defined as the state of the student's college. December 25-27 are typical planning and commuting days for travel to the conference, which occurs December 27-31. Standard errors are clustered at the state level.

Table 3.10.1: Descriptive Statistics by College and Year, 2003-2017

| | Conference and Year After | Other Years |
|---|---|---|
| *Conference Attendance* | | |
| Campus Attendance | 17.8 | 0 |
| Freshman Attendance | 2.20 | 0 |
| Sophomore Attendance | 4.30 | 0 |
| Junior Attendance | 5.42 | 0 |
| Senior Attendance | 5.86 | 0 |
| *Short-Run Giving and Employment* | | |
| Total Student Donations in Year | 1877.0 | 1460.9 |
| New Staff Hired April-Nov | 0.32 | 0.22 |
| *Campus Survey* | | |
| Total Members | 85.4 | 84.3 |
| N Discussion Group Participants | 55.4 | 55.0 |
| N Student Leaders | 19.1 | 19.0 |
| N Cross-Cultural Friendships | 13.6 | 12.9 |
| N Became Christian | 7.16 | 7.36 |
| N Seniors Became Christian | 1.91 | 1.74 |
| N Regularly Attend Church | 52.3 | 51.4 |
| N Volunteers | 24.5 | 23.9 |
| *Survey Demographics* | | |
| Share Women | 0.60 | 0.60 |
| Share Black (Non-Hispanic) | 0.10 | 0.11 |
| Share White (Non-Hispanic) | 0.57 | 0.57 |
| Share Hispanic | 0.050 | 0.050 |
| Share Asian (Non-Hispanic) | 0.14 | 0.13 |
| *Cohort Level (Graduation Year Level) Outcomes* | | |
| Giving 3 Years Post-Graduation (USD) | 6432.1 | 5771.5 |
| Ever Joined Staff | 0.30 | 0.27 |
| N Unique Schools | 279 | 279 |
| School-Year Observations | 1,395 | 2,790 |

Notes: This table describes the sample of college-by-year observations used to measure the short-run impact of missions conferences using a rotational timing and attendance propensity design. The panel titled "Cohort Level" describes longer-run outcomes by college and graduation year that are used for the cohort-level analysis.

Table 3.10.2: Conference Short-Run Effects
Donors and Staff

| | (1)<br>Campus Attendance | (2)<br>Total Student Donations in Year | (3)<br>Student Donations, Excl Conference Window | (4)<br>New Staff Hired April-Nov |
|---|---|---|---|---|
| *Panel A: Conference Year* | | | | |
| Conference Year | 17.7742*** | 416.1357*** | 97.5365** | 0.1047*** |
| | (1.2227) | (53.8612) | (40.5411) | (0.0207) |
| School × Conference Cycle FEs | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | |
| Campus Attendance in Cycle | 0.0000 | 58.1608*** | 55.0339*** | 0.0044*** |
| | (0.0000) | (8.4759) | (7.9674) | (0.0012) |
| Campus Attendance in Cycle × Conference Year | 1.0000*** | 18.2331*** | 4.5719 | 0.0077*** |
| | (0.0000) | (4.9836) | (4.1769) | (0.0022) |
| Year Fixed Effects | × | × | × | × |
| Size Controls | × | × | × | × |
| Size × Conference Year Controls | × | × | × | × |
| Mean of Dep Var (Levels) | 5.9247 | 1,599.6262 | 1,407.2963 | 0.2550 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 |

Notes: This table presents estimates of the effect of conference timing and attendance on student donations and staff employment in the following year. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with college attendance in that conference cycle. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.10.3: Conference Cohort-Level Effects

| | Giving 3 Years Post-Graduation (USD) | | Ever Joined Staff | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Panel A: Cohort Graduation Relative to Conference Year* | | | | |
| Junior During Conference Year | 279.9752 | 279.9752 | -0.0022 | -0.0022 |
| | (260.6896) | (260.6896) | (0.0188) | (0.0188) |
| Senior During Conference Year | 800.5952*** | 800.5952*** | 0.0315 | 0.0315 |
| | (233.4141) | (233.4141) | (0.0198) | (0.0198) |
| School × Conference Cycle FEs | | × | | × |
| *Panel B: Cohort Graduation Relative to Conference Year × Attendance* | | | | |
| Campus Attendance in Cycle | 319.2389*** | 234.5304*** | 0.0107*** | 0.0056*** |
| | (32.1771) | (47.9252) | (0.0013) | (0.0019) |
| Campus Attendance in Cycle × Junior During Conference Year | 32.5961 | 31.7063 | -0.0010 | -0.0017 |
| | (32.3576) | (42.5373) | (0.0019) | (0.0024) |
| Campus Attendance in Cycle × Senior During Conference Year | 67.1397*** | 50.5932* | 0.0027** | 0.0032 |
| | (24.1811) | (28.0865) | (0.0013) | (0.0024) |
| Year Fixed Effects | × | × | × | × |
| Size Controls | | × | | × |
| Size × Conference Year Controls | | × | | × |
| Mean of Dep Var (Levels) | 5,991.7076 | 5,991.7076 | 0.2757 | 0.2757 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 |

Notes: This table presents estimates of the effect of conference timing and attendance on long-run donations and staff employment by graduation year. Panel A shows the difference in the dependent variable at the college-graduation-year level for graduation cohorts that had a conference during junior or senior year, respectively, compared to cohorts that had a conference during sophomore year. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on grade during conference year interacted with college attendance in that conference cycle. The regression estimates in Panels A and B must be re-scaled using a bounding exercise to estimate the effect of an additional senior attendee on long-run outcomes. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The results in Panel A are unchanged by the inclusion of college-by-conference-cycle fixed effects due to balance in the sample of college observations over time. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.10.4: Characteristics of Undergraduate Conference Registrants, 2000-2015

| | Attended | Cancelled or No-Show |
|---|---|---|
| *Outcomes and Key Regressors* | | |
| Attended Conference | 1 | 0 |
| Ever Future Giving | 0.40 | 0.19 |
| N Future Gifts | 28.2 | 9.88 |
| Amount Future Giving | 2229.8 | 723.6 |
| Ever Joined Staff | 0.027 | 0.0066 |
| Days on Staff | 55.6 | 14.6 |
| Future Mission Trip | 0.037 | 0.014 |
| *Registration and Conditions* | | |
| Share State Flights Cancelled Dec 25-27 | 0.022 | 0.029 |
| Final Payment Amount for Conference | 378.7 | 113.1 |
| *Student Involvement* | | |
| Ever Student Leader | 0.25 | 0.14 |
| Any Past Giving | 0.14 | 0.11 |
| Previously Attended Conference | 0.059 | 0.069 |
| Past Mission Trip | 0.022 | 0.021 |
| *Demographics* | | |
| Female | 0.60 | 0.63 |
| Black (Non-Hispanic) | 0.065 | 0.13 |
| White (Non-Hispanic) | 0.50 | 0.44 |
| Hispanic | 0.048 | 0.079 |
| Asian/PI (Non-Hispanic) | 0.30 | 0.23 |
| Race/Ethnicity Missing | 0.071 | 0.089 |
| Grade 1 (Freshman) | 0.14 | 0.15 |
| Grade 2 (Sophomore) | 0.24 | 0.25 |
| Grade 3 (Junior) | 0.30 | 0.29 |
| Grade 4 (Senior) | 0.31 | 0.32 |
| Registrant Observations | 48,596 | 2,717 |

Notes: This table describes the sample of undergraduate registrant observations used to measure the long-run (and short-run) impact of missions.

Table 3.10.5: Estimates of Conference Attendance on Giving and Staff Employment

| | Amount Future Giving | | | Ever Joined Staff | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Attended Conference | 1328.0*** | 482.4*** | | 0.0231*** | 0.00927*** | |
| | (111.6) | (82.25) | | (0.00189) | (0.00189) | |
| Share State Flights Not Cancelled Dec 25-27 | | | 2162.2** | | | 0.00518 |
| | | | (919.3) | | | (0.0221) |
| State, Year, and Grade FEs | × | × | × | × | × | × |
| Detailed Controls | | × | | | × | |
| Mean of Dep Var | 2,150 | 2,157 | 2,150 | .026 | .0261 | .026 |
| Observations | 51,311 | 51,110 | 51,311 | 51,311 | 51,110 | 51,311 |

Notes: This table presents estimates of the effect of conference attendance on future donations and staff employment. Columns 1 and 4 present estimates from OLS regressions with basic controls. Columns 2 and 5 present estimates controlling for a detailed set of observed characteristics. Columns 3 and 6 present reduced form estimates of the effect of flight cancellations on the dependent variable. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.10.6: Estimates of Conference Attendance on Giving and Staff Employment, by Grade

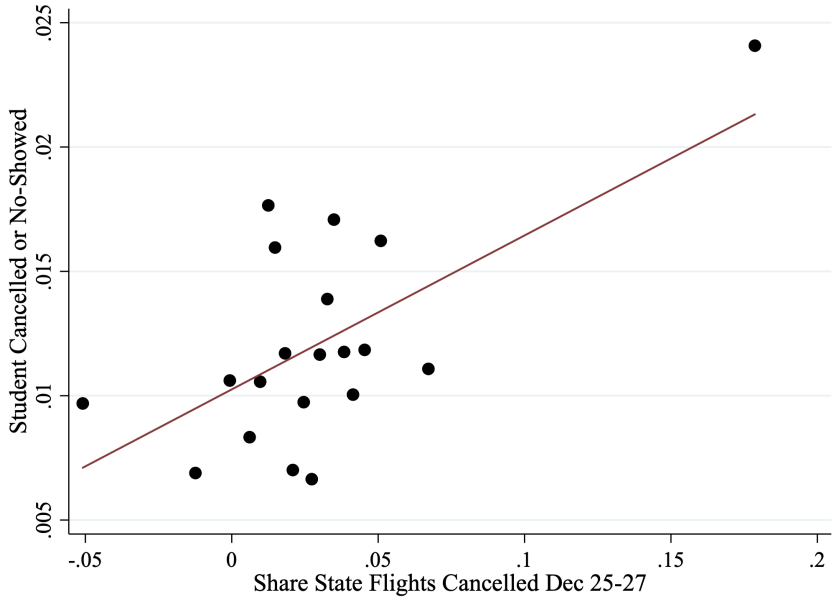
| | Amount Future Giving | | | Ever Joined Staff | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Attended × Freshman (G1) | 252.0* | | | 0.00364 | | |
| | (152.2) | | | (0.00334) | | |
| Attended × Sophomore (G2) | 265.5* | | | 0.00335 | | |
| | (142.0) | | | (0.00408) | | |
| Attended × Junior (G3) | 245.6 | | | 0.00831*** | | |
| | (173.5) | | | (0.00290) | | |
| Attended × Senior (G4) | 924.2*** | | | 0.0180*** | | |
| | (176.4) | | | (0.00377) | | |
| Share Flights Not Cancelled × Freshman (G1) | | -406.8 | 628.2 | | -0.0396 | -0.0311 |
| | | (1956.1) | (1579.9) | | (0.0407) | (0.0373) |
| Share Flights Not Cancelled × Sophomore (G2) | | 888.3 | 1364.0 | | 0.0135 | 0.00948 |
| | | (1542.6) | (1351.2) | | (0.0356) | (0.0328) |
| Share Flights Not Cancelled × Junior (G3) | | 1906.7* | 1741.5 | | 0.0159 | 0.0196 |
| | | (1122.8) | (1067.3) | | (0.0316) | (0.0340) |
| Share Flights Not Cancelled × Senior (G4) | | 4357.8*** | 2675.6*** | | 0.00917 | -0.00374 |
| | | (1226.3) | (869.5) | | (0.0325) | (0.0257) |
| State, Year, and Grade FEs | × | × | × | × | × | × |
| Detailed Controls | × | | × | × | | × |
| Mean of Dep Var | 2,157 | 2,150 | 2,150 | .0261 | .026 | .026 |
| Observations | 51,110 | 51,311 | 51,309 | 51,110 | 51,311 | 51,309 |

Notes: This table presents heterogeneity in the estimates of the effect of conference attendance on future donations and staff employment by grade. Columns 1 and 3 present estimates from OLS regressions controlling for a detailed set of observed characteristics, where the attendance indicator is interacted with grade. Columns 2 and 4 present reduced form estimates of the effect of flight cancellations interacted with grade on the dependent variable. Columns 3 and 6 present flight cancellations results with additional controls. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

# Appendix

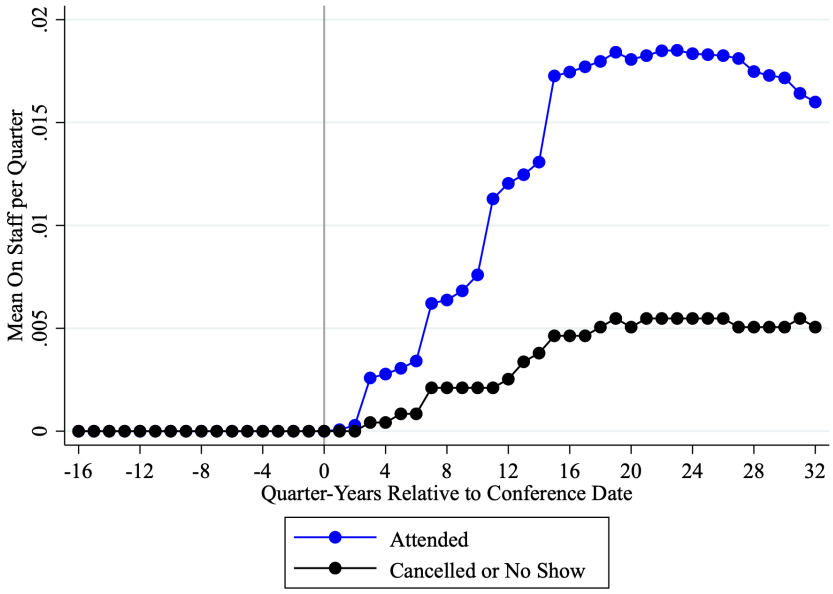
## 3.A Appendix Figures and Tables

Figure 3.A.1: Relationship between Flight Cancellations and Conference Attendance



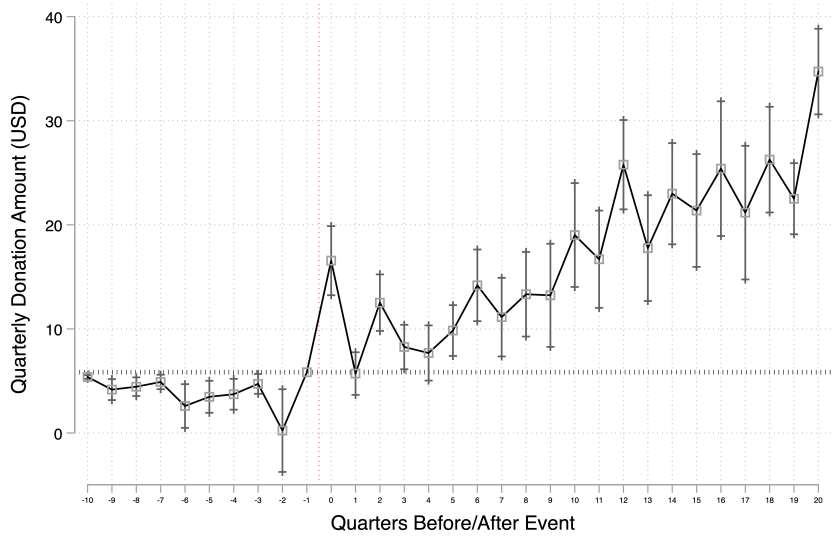

Notes: This binned scatter plot illustrates the first stage relationship between the a state's share of weather-related cancelled flights and non-attendance (cancelled or no-showed) for students attending colleges in the state, controlling for state and conference year fixed effects. Flight cancellations are calculated using flights from major airlines departing from a state's airports and with destinations in either St. Louis or Chicago.

Figure 3.A.2: Descriptive Long-Run Staff Employment Trajectories of Attending vs. Non-Attending Registrants




Notes: This figure shows the dynamic trend of staff employment of undergraduate students who registered and attended the conference compared to undergraduate students and registered but did not attend. Sample restricted to student-registrant observations from 2000-2012.

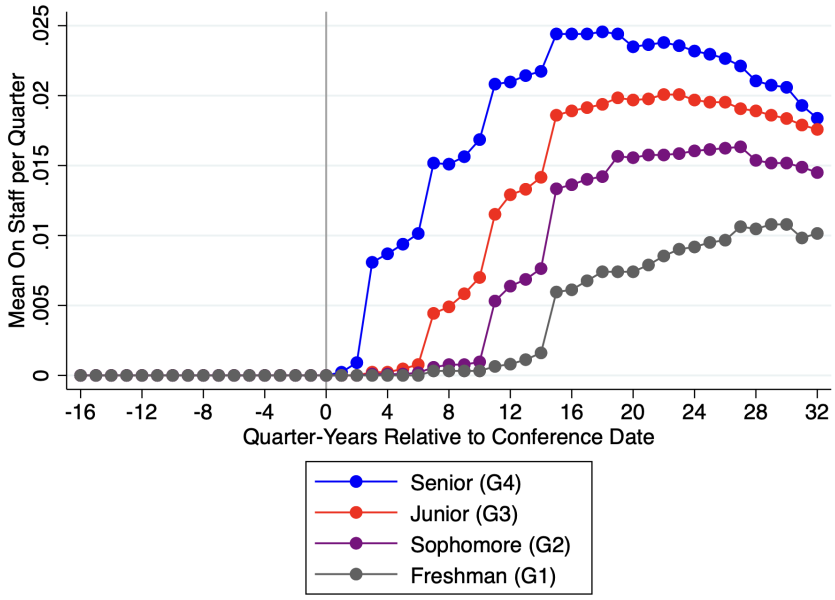
Figure 3.A.3: Event Study Estimates of Attendance on Giving Over Time



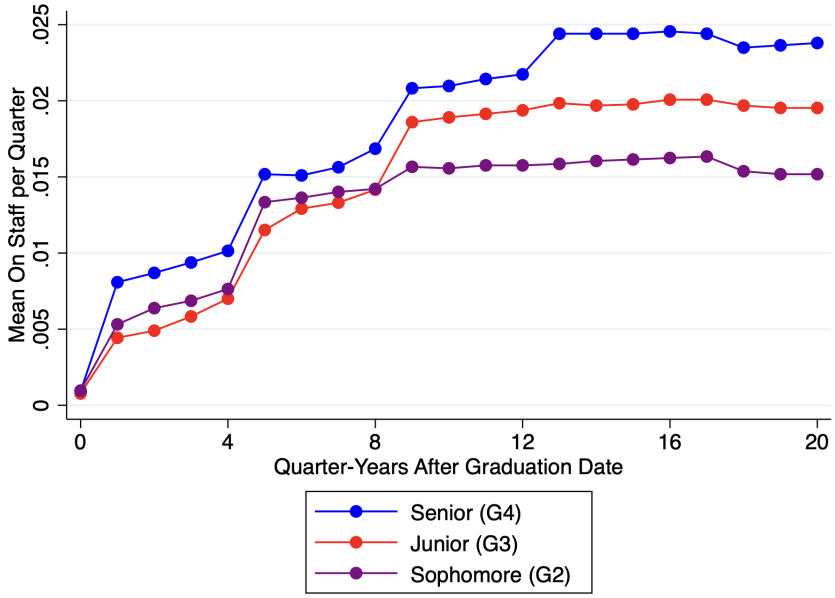

Notes: This figure presents event study estimates of quarterly donations from undergraduate students who registered and attended the conference relative to undergraduate students and registered but did not attend. Estimates control for predicted donations using pre-event characteristics, interacted with event time indicators. Standard errors are clustered at the student level.

Figure 3.A.4: On Staff per Quarter, Among Conference Attendees

(1) Panel A: Relative to Conference Date

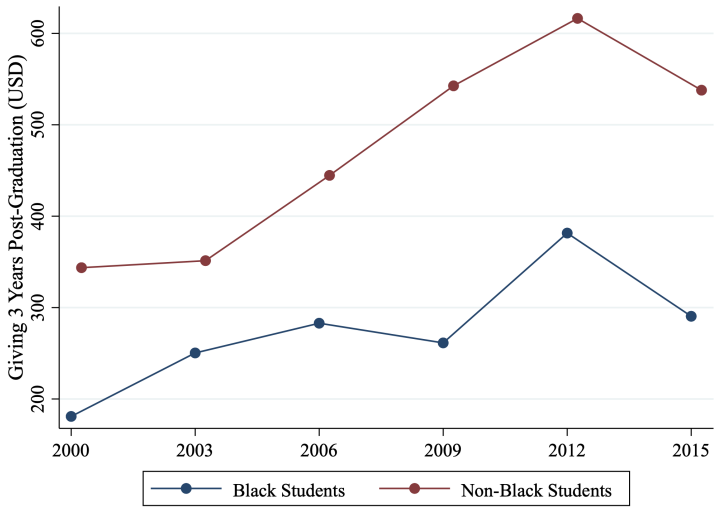



(2) Panel B: Relative to Graduation Date



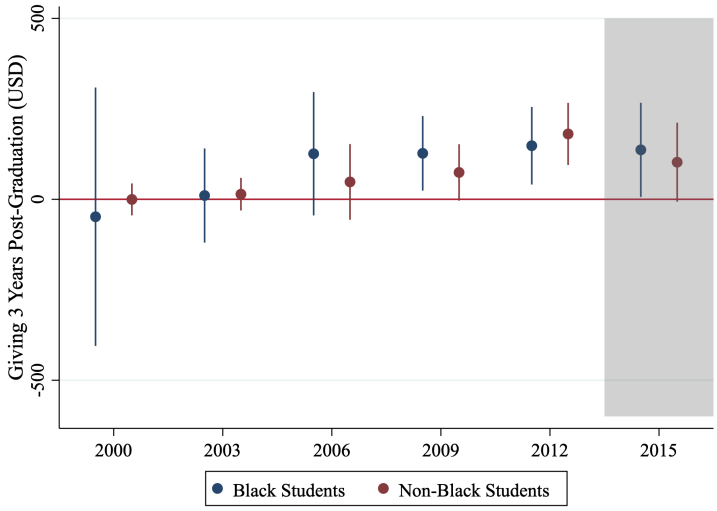

Notes: This figure shows the dynamic trend of staff employment of undergraduate students who registered and attended the conference by grade. In panel, the x-axis is measured in quarters relative to the conference date. In panel B, x-axis is measured in quarters relative to graduation date for comparability.

Figure 3.A.5: Giving Among All Registrants, by Race and Conference Year



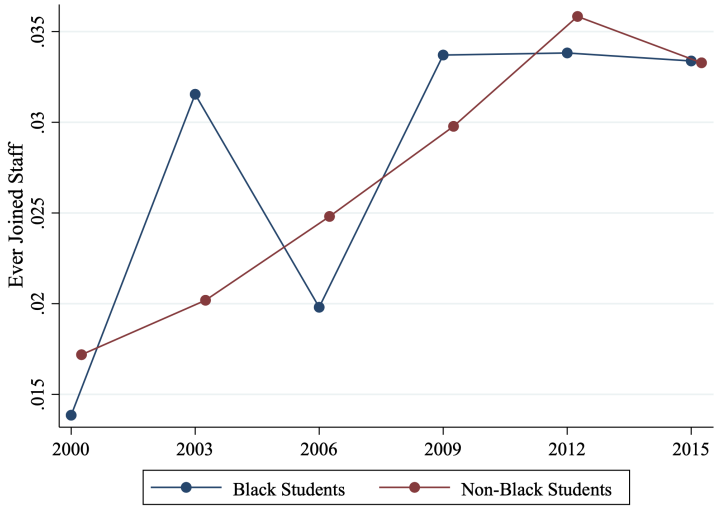

Notes: This figure presents the mean donation amount within three years of graduation for conference attendees, by race (Non-Hispanic Black compared to all other) and conference year.

Figure 3.A.6: Difference in Giving for Attendees vs. Non-Attendees, by Race and Conference Year



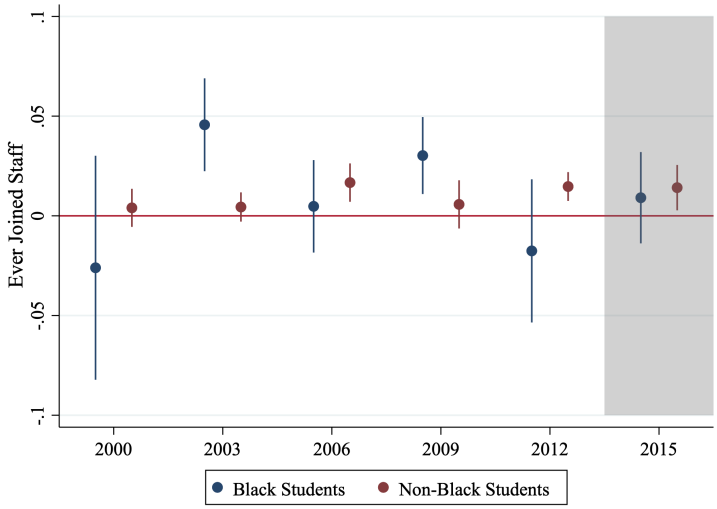

Notes: This figure presents point estimates and 95 percent confidence intervals from an OLS regression of total donation amount within three years of graduation on attendance interacted with race and year indicators. The regression includes controls for gender, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The shaded region represents the year of prominent Black Lives Matter messaging during the conference (2015). Standard errors are clustered at the student level.

Figure 3.A.7: Future Staff Employment Among All Registrants, by Race and Conference Year




Notes: This figure presents the share of conference attendees who became staff for the sponsor organization, by race (Non-Hispanic Black compared to all other) and conference year.

Figure 3.A.8: Difference in Future Staff Employment for Attendees vs. Non-Attendees by Race and Conference Year




Notes: This figure presents point estimates and 95 percent confidence intervals from an OLS regression of ever being employed as staff on attendance interacted with race and year indicators. The regression includes controls for gender, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The shaded region represents the year of prominent Black Lives Matter messaging during the conference (2015). Standard errors are clustered at the student level.

Table 3.A.1: Conference Short-Run Effects on Giving and Staff Employment
Robustness to Simulated Attendance

| | (1)<br>Campus Attendance | (2)<br>Total Student Donations in Year | (3)<br>Student Donations, Excl Conference Window | (4)<br>New Staff Hired April-Nov |
|---|---|---|---|---|
| *Panel A: Conference Year* | | | | |
| Conference Year | 17.7742*** | 416.1357*** | 97.5365** | 0.1047*** |
| | (1.2227) | (53.8612) | (40.5411) | (0.0207) |
| School × Conference Cycle FEs | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | |
| att_looclose_avg | -0.0000 | 61.7096*** | 58.2846*** | 0.0049*** |
| | (.) | (8.8652) | (8.3909) | (0.0012) |
| att_looclose_avg × Conference Year | 0.9474 | 12.4368*** | -0.4776 | 0.0068*** |
| | (.) | (4.7862) | (4.0759) | (0.0022) |
| Year Fixed Effects | × | × | × | × |
| Size Controls | × | × | × | × |
| Size × Conference Year Controls | × | × | × | × |
| Mean of Dep Var (Levels) | 5.9247 | 1,599.6262 | 1,407.2963 | 0.2550 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 |

Notes: This table presents estimates of the effect of conference timing and attendance on student donations and staff employment in the following year. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with mean college attendance in adjacent cycles. The difference between this table and Table 3.10.2 is that the Table 3.10.2 panel B interacts conference year with in-cycle attendance, whereas this table interacts conference year with a measure of attendance computed using other conference cycles. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.2: Conference Short-Run Effects on Giving and Staff Employment Restricted to Students who Filled Out Departing Student Form

| | (1) Campus Attendance | (2) Total Student Donations in Year, Filled Form | (3) Student Donations with Form, Excl Conference Window | (4) New Staff Hired April-Nov, Filled Form |
|---|---|---|---|---|
| *Panel A: Conference Year* | | | | |
| Conference Year | 17.7742*** | 317.7248*** | 77.5803** | 0.0749*** |
| | (1.2227) | (46.5054) | (36.3402) | (0.0159) |
| School × Conference Cycle FEs | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | |
| Campus Attendance in Cycle | 0.0000 | 47.4341*** | 44.8913*** | 0.0032*** |
| | (0.0000) | (6.5472) | (6.1391) | (0.0011) |
| Campus Attendance in Cycle × Conference Year | 1.0000*** | 14.8621*** | 4.0693 | 0.0047** |
| | (0.0000) | (4.6577) | (3.9517) | (0.0019) |
| Year Fixed Effects | × | × | × | × |
| Size Controls | × | × | × | × |
| Size × Conference Year Controls | × | × | × | × |
| Mean of Dep Var (Levels) | 5.9247 | 1,303.9279 | 1,156.4538 | 0.1730 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 |

Notes: This table presents estimates of the effect of conference timing and attendance on student donations and staff employment in the following year among students who completed a departing student form. This table shows that short-run findings are robust to restricting the sample to a set of students for whom data linkage is most likely. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with mean college attendance in that conference cycle. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.3: Conference Short-Run Effects
Departing Student Form Responses

| | (1) N Departing Student Forms | (2) Will Pray | (3) Will Give | (4) Will Volunteer | (5) Indicated a Job Plan | (6) Planned Religious Job | (7) Planned Religious Job for Host |
|---|---|---|---|---|---|---|---|
| *Panel A: Conference Year* | | | | | | | |
| Conference Year | -0.0674 | 0.0376 | 0.0265 | -0.0247 | 0.0720* | 0.0294* | 0.0061 |
| | (0.1299) | (0.1195) | (0.0658) | (0.0514) | (0.0392) | (0.0152) | (0.0132) |
| School × Conference Cycle FEs | × | × | × | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | | | | |
| Campus Attendance in Cycle | 0.0838*** | 0.0726*** | 0.0614*** | 0.0353*** | 0.0166*** | 0.0008 | 0.0012* |
| | (0.0100) | (0.0104) | (0.0108) | (0.0111) | (0.0028) | (0.0007) | (0.0006) |
| Campus Attendance in Cycle × Conference Year | 0.0064 | -0.0033 | 0.0114* | 0.0047 | 0.0102*** | 0.0056*** | 0.0040 |
| | (0.0134) | (0.0111) | (0.0059) | (0.0070) | (0.0036) | (0.0020) | (0.0024) |
| Year Fixed Effects | × | × | × | × | × | × | × |
| Size Controls | × | × | × | × | × | × | × |
| Size × Conference Year Controls | × | × | × | × | × | × | × |
| Mean of Dep Var (Levels) | 7.1360 | 5.6208 | 2.4769 | 1.6810 | 0.9577 | 0.1532 | 0.1027 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 | 4,185 | 4,185 | 4,185 |

Notes: This table presents estimates of the effect of conference timing and attendance on survey responses completed by graduating and disenrolling students in the following year. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with college attendance in that conference cycle. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

## Table 3.A.4: Short-Run Effect of Conference on Reported Post-Graduation Job

| | (1) N Departing Student Forms | (2) Indicated a Job Plan | (3) Planned Religious Job | (4) Planned Religious Job for Host | (5) Planned Religious Job Not Host | (6) Planned Tech Job | (7) Planned Job for Large Firm | (8) Planned Non-Religious Service-Oriented Job |
|---|---|---|---|---|---|---|---|---|
| *Panel A: Conference Year* | | | | | | | | |
| Conference Year | -0.0674 | 0.0731* | 0.0215 | 0.0086 | 0.0129** | -0.0093* | -0.0085 | 0.0068 |
| | (0.1299) | (0.0391) | (0.0156) | (0.0147) | (0.0063) | (0.0053) | (0.0148) | (0.0072) |
| School × Conference Cycle FEs | × | × | × | × | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | | | | | |
| Campus Attendance in Cycle | 0.0838*** | 0.0165*** | 0.0008 | 0.0015** | -0.0007** | 0.0013*** | 0.0040*** | 0.0004 |
| | (0.0100) | (0.0028) | (0.0008) | (0.0007) | (0.0004) | (0.0004) | (0.0007) | (0.0005) |
| Campus Attendance in Cycle × Conference Year | 0.0064 | 0.0101*** | 0.0053** | 0.0044* | 0.0009 | -0.0008 | -0.0008 | 0.0003 |
| | (0.0134) | (0.0036) | (0.0024) | (0.0026) | (0.0006) | (0.0005) | (0.0010) | (0.0007) |
| Year Fixed Effects | × | × | × | × | × | × | × | × |
| Size Controls | × | × | × | × | × | × | × | × |
| Size × Conference Year Controls | × | × | × | × | × | × | × | × |
| Mean of Dep Var (Levels) | 7.1360 | 0.9606 | 0.1627 | 0.1305 | 0.0323 | 0.0284 | 0.1333 | 0.0478 |
| Mean Attendance in Conf Year | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 17.7742 | 22.3006 | 17.7742 |
| Observations | 4,185 | 4,185 | 4,185 | 4,185 | 4,185 | 4,185 | 2,655 | 4,185 |

Notes: Regressions include campus fixed effects, cohort-by-region fixed effects, and additional campus-cohort controls. Notes: This table presents estimates of the effect of conference timing and attendance on survey responses completed by graduating and disenrolling students in the following year. Students were asked to record post-graduation job plans, which have been aggregated by sector. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with college attendance in that conference cycle. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

## Table 3.A.5: Conference Short-Run Effects Select Campus Survey Responses

| | (1) Total Members | (2) N Seniors | (3) N Seniors Became Christian | (4) N Non-Christians Involved | (5) N Cross-Cultural Friendships | (6) N Regularly Attend Church |
|---|---|---|---|---|---|---|
| *Panel A: Conference Year* | | | | | | |
| Conference Year | 1.0968** | 0.2006 | 0.1211 | 2.6698*** | 0.7004** | 0.9329** |
| | (0.4975) | (0.1543) | (0.0757) | (0.9930) | (0.3214) | (0.4174) |
| School × Conference Cycle FEs | × | × | × | × | × | × |
| *Panel B: Conference Year × Attendance* | | | | | | |
| Campus Attendance in Cycle | 0.1390** | 0.0750*** | -0.0005 | 0.2037 | 0.1730** | 0.2474*** |
| | (0.0619) | (0.0173) | (0.0080) | (0.1493) | (0.0701) | (0.0773) |
| Campus Attendance in Cycle × Conference Year | -0.0946 | 0.0086 | 0.0150** | -0.1070 | 0.0322 | -0.0397 |
| | (0.0664) | (0.0147) | (0.0060) | (0.1115) | (0.0312) | (0.0407) |
| Year Fixed Effects | × | × | × | × | × | × |
| Size Controls | | × | × | × | × | × |
| Size × Conference Year Controls | | × | × | × | × | × |
| Mean of Dep Var (Levels) | 84.6294 | 10.9324 | 1.8418 | 23.8246 | 13.0865 | 51.7174 |
| Mean Attendance in Conf Year | 17.7742 | 17.9221 | 18.0987 | 17.7904 | 17.7742 | 17.7841 |
| Observations | 4,185 | 4,083 | 3,678 | 4,179 | 4,185 | 4,182 |

Notes: This table presents estimates of the effect of conference timing and attendance on college-level survey estimates completed by staff in the following year. Panel A shows the difference in the dependent variable at the college-year level for years with a conference compared to years without a conference. Panel B presents estimates of the effect of an additional attendee by regressing the dependent variable on conference year interacted with college attendance in that conference cycle. A conference cycle is a three-year period defined as the year before, proceeding, and an additional year after a given conference. The size controls in Panel B include the mean number of club members, discussion group members, and student leaders, all averaged at the college-by-conference-cycle level. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

## Table 3.A.6: Flight Cancellation Instrument Details

| | Balance Test Attendance with Controls | Balance Test Flight Cancellations | First Stage Instrument Strength | Exclusion Restriction Test: Attendees Only | Example IV Estimate |
|---|---|---|---|---|---|
| | Attended Conference | Share Flights Cancelled Dec 25-27 | Attended Conference | Amount Future Giving | Amount Future Giving |
| Share Flights Cancelled Dec 25-27 | | | -0.221*** | -2154.8** | |
| | | | (0.0523) | (1035.3) | |
| Attended Conference | | | | | 10490.4** |
| | | | | | (5205.9) |
| Previously Attended Conference | -0.00952* | 0.000774 | | | |
| | (0.00506) | (0.000838) | | | |
| Any Past Giving | 0.00367 | -0.000388 | | | |
| | (0.00355) | (0.000551) | | | |
| Amount Past Giving | -0.00000143 | -0.000000124 | | | |
| | (0.00000223) | (0.000000212) | | | |
| Past Mission Trip | -0.00468 | -0.000799 | | | |
| | (0.00678) | (0.00127) | | | |
| N Days Registered in Advance | | -0.00000138 | | | |
| | | (0.00000288) | | | |
| Ever Student Leader | | -0.000142 | | | |
| | | (0.000345) | | | |
| Black (Non-Hispanic) | | 0.000406 | | | |
| | | (0.000554) | | | |
| White (Non-Hispanic) | | 0.000291 | | | |
| | | (0.000957) | | | |
| Hispanic | | 0.000591 | | | |
| | | (0.00134) | | | |
| Asian/PI (Non-Hispanic) | | 0.000970 | | | |
| | | (0.000641) | | | |
| Female | | 0.000227 | | | |
| | | (0.000327) | | | |
| Share Flights Cancelled Dec 20 | | | 0.0209 | | |
| | | | (0.0227) | | |
| Share Flights Cancelled Dec 21 | | | -0.0585 | | |
| | | | (0.0540) | | |
| Share Flights Cancelled Dec 22 | | | 0.122* | | |
| | | | (0.0668) | | |
| Share Flights Cancelled Dec 23 | | | 0.0114 | | |
| | | | (0.129) | | |
| Share Flights Cancelled Dec 24 | | | 0.0443 | | |
| | | | (0.0654) | | |
| Share Flights Cancelled Dec 28 | | | 0.0227 | | |
| | | | (0.0235) | | |
| Share Flights Cancelled Dec 29 | | | -0.00796 | | |
| | | | (0.0436) | | |
| Share Flights Cancelled Dec 30 | | | 0.00218 | | |
| | | | (0.0103) | | |
| Share Flights Cancelled Dec 31 | | | 0.0234 | | |
| | | | (0.0234) | | |
| State, Year, and Grade FEs | × | × | × | × | × |
| Detailed Controls | × | | | | |
| F-statistic | 1.24 | 0.52 | 17.82 | | |
| p-value of F-test | 0.2906 | 0.8830 | 0.0001 | | |
| Mean of Dep Var | .947 | 1,921 | .947 | 2,230 | 2,150 |
| Observations | 51,110 | 51,110 | 51,311 | 48,594 | 51,311 |

Notes: Column 1 shows that attendance is uncorrelated with a number of salient characteristics that could predict future giving (F-test p-value $> 0.10$), after conditioning on a set of background controls (school, grade, gender, age, sponsor organization membership, leadership role, and registration timing). Column 2 shows that the flight cancellation instrument is uncorrelated with a number of salient characteristics (F-test p-value $> 0.10$). Column 3 shows that the share of school state-specific weather-related flight cancellations during potential travel days Dec 25-27 is significantly correlated with attendance (F-statistic $> 10$) but is not for non-travel placebo dates, as expected. Finally, column 4 provides suggestive evidence that bad weather affects student outcomes independent of a student's own attendance by regressing future donations on flight cancellations for the subsample of students who attended the conference. This possible exclusion restriction violation could explain why the magnitude of IV point estimates of the effect of conference attendance are quite large (column 5). Detailed controls include indicators for gender and race/ethnicity, school fixed

effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by school with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.7: Estimates of Conference Attendance
Sample of Students who did not have First-Time Donation at the Conference

| | Amount Future Giving | | | Ever Joined Staff | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Attended Conference | 1183.3*** | 400.5*** | | 0.0212*** | 0.00867*** | |
| | (108.4) | (80.35) | | (0.00191) | (0.00188) | |
| Share State Flights Not Cancelled Dec 25-27 | | | 2133.8** | | | 0.00724 |
| | | | (983.0) | | | (0.0245) |
| State, Year, and Grade FEs | × | × | × | × | × | × |
| Detailed Controls | | × | | | × | |
| Mean of Dep Var | 2,019 | 2,027 | 2,019 | .024 | .0241 | .024 |
| Observations | 45,705 | 45,503 | 45,705 | 45,705 | 45,503 | 45,705 |

Notes: This table presents estimates of the association of conference attendance with future donations and staff employment for a subsample of students who did have their first recorded donation at the conference. In doing so, this table shows that the association between attendance and future giving is robust to any unforeseen mechanical difference in donor linkage due to giving at the conference. Columns 1 and 4 present estimates from OLS regressions with basic controls. Columns 2 and 5 present estimates controlling for a detailed set of observed characteristics. Columns 3 and 6 present reduced form estimates of the effect of flight cancellations on the dependent variable. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.8: Estimates of Conference Attendance on Giving and Staff Employment: Expanded Outcomes

| | Any Future Giving | | Amount Future Giving | | N Future Gifts | | Ever Joined Staff | | Days on Staff | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Attended Conference | 0.135*** | | 482.4*** | | 5.270*** | | 0.00927*** | | 16.16*** | |
| | (0.00815) | | (82.25) | | (0.947) | | (0.00189) | | (4.793) | |
| Share State Flights Not Cancelled Dec 25-27 | | 0.128 | | 2162.2** | | 15.85 | | 0.00518 | | 21.35 |
| | | (0.138) | | (919.3) | | (12.46) | | (0.0221) | | (44.07) |
| State, Year, and Grade FEs | × | × | × | × | × | × | × | × | × | × |
| Detailed Controls | × | | × | | × | | × | | × | |
| Mean of Dep Var | .387 | .386 | 2,157 | 2,150 | 27.3 | 27.2 | .0261 | .026 | 53.7 | 53.5 |
| Observations | 51,110 | 51,311 | 51,110 | 51,311 | 51,110 | 51,311 | 51,110 | 51,311 | 51,110 | 51,311 |

Notes: This table presents additional estimates of the effect of conference attendance on future donations and staff employment. Columns 1, 3, 5, and 7 present estimates from OLS regressions controlling for a detailed set of observed characteristics. Columns 2, 4, 6, and 8 present reduced form estimates of the effect of flight cancellations on the dependent variable. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.9: Future Event Participation

| | Future Mission Trip | | | Future Conference Attended | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Attended Conference | 0.0245*** | 0.00944*** | | 0.0725*** | 0.0390*** | |
| | (0.00281) | (0.00282) | | (0.00472) | (0.00460) | |
| Share State Flights Not Cancelled Dec 25-27 | | | 0.00338 | | | 0.0778 |
| | | | (0.0450) | | | (0.0905) |
| State, Year, and Grade FEs | × | × | × | × | × | × |
| Detailed Controls | | × | | | × | |
| Mean of Dep Var | .0362 | .0363 | .0362 | .118 | .118 | .118 |
| Observations | 51,311 | 51,110 | 51,311 | 51,311 | 51,110 | 51,311 |

Notes: This table presents estimates of the effect of conference attendance on future sponsor organization short-term mission trips to foreign countries ("Global Programs") and future attendance of the conference. Columns 1 and 4 present estimates from OLS regressions with basic controls. Columns 2 and 5 present estimates controlling for a detailed set of observed characteristics. Columns 3 and 6 present reduced form estimates of the effect of flight cancellations on the dependent variable. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.10: Heterogeneity by Level of sponsor Organization Involvement

| | Amount Future Giving | | | | | | Ever Joined Staff | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Attended × Host Org Member | 2456.9*** | | | | | | 0.0344*** | | | | | |
| | (201.8) | | | | | | (0.00396) | | | | | |
| Attended × Not Host Org Member | 84.71 | | | | | | 0.0143*** | | | | | |
| | (158.9) | | | | | | (0.00362) | | | | | |
| Attended × Filled Departing Student Form | | 3050.0*** | | | | | | 0.0507*** | | | | |
| | | (253.7) | | | | | | (0.00470) | | | | |
| Attended × Did Not Fill Departing Student Form | | 734.8*** | | | | | | 0.0159*** | | | | |
| | | (162.6) | | | | | | (0.00365) | | | | |
| Attended × Ever Student Leader | | | 4138.5*** | | | | | | 0.0655*** | | | |
| | | | (281.3) | | | | | | (0.00496) | | | |
| Attended × Never Student Leader | | | 186.8 | | | | | | 0.00827** | | | |
| | | | (160.2) | | | | | | (0.00351) | | | |
| Share Flights Not Cancelled × Host Org Member | | | | 4259.9*** | | | | | | 0.0273 | | |
| | | | | (823.4) | | | | | | (0.0222) | | |
| Share Flights Not Cancelled × Not Host Org Member | | | | 630.0 | | | | | | -0.0110 | | |
| | | | | (847.4) | | | | | | (0.0220) | | |
| Share Flights Not Cancelled × Filled Departing Student Form | | | | | 5635.7*** | | | | | | 0.0529** | |
| | | | | | (849.4) | | | | | | (0.0216) | |
| Share Flights Not Cancelled × Did Not Fill Departing Student Form | | | | | 1751.3** | | | | | | -0.000463 | |
| | | | | | (835.4) | | | | | | (0.0212) | |
| Share Flights Not Cancelled × Ever Student Leader | | | | | | 6323.7*** | | | | | | 0.0623*** |
| | | | | | | (831.1) | | | | | | (0.0217) |
| Share Flights Not Cancelled × Never Student Leader | | | | | | 798.2 | | | | | | -0.0135 |
| | | | | | | (811.8) | | | | | | (0.0209) |
| State, Year, and Grade FEs | × | × | × | × | × | × | × | × | × | × | × | × |
| Detailed Controls | × | × | × | | | | × | × | × | | | |
| Grade Attendance Controls | × | × | × | | | | × | × | × | | | |
| Mean of Dep Var | 2,157 | 2,157 | 2,157 | 2,150 | 2,150 | 2,150 | 2,157 | 2,157 | 2,157 | 2,150 | 2,150 | 2,150 |
| Observations | 51,110 | 51,110 | 51,110 | 51,311 | 51,311 | 51,311 | 51,110 | 51,110 | 51,110 | 51,311 | 51,311 | 51,311 |

Notes: This table presents heterogeneity in the estimates of the effect of conference attendance on future donations and staff employment by level of sponsor organization involvement. Columns 1-3 and 7-9 present estimates from OLS regressions controlling for a detailed set of observed characteristics, where the attendance

indicator is interacted with an involvement indicator. Columns 4-6 and 10-12 present reduced form estimates of the effect of flight cancellations interacted with involvement indicator on the dependent variable. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college with the exception of the flight cancellation regressions where they are clustered by state. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.A.11: Black Lives Matter Messaging
Difference-in-Differences (DD) and Triple Difference (DDD) Estimates

| | Giving 3 Years Post-Graduation (USD) | | Ever Joined Staff | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Black (Non-Hispanic) | -95.57*** | -92.16*** | -0.00150 | -0.00110 |
| | (24.48) | (24.02) | (0.00343) | (0.00341) |
| Black (Non-Hispanic)=1 × 2015=1 | -29.66 | -7.670 | -0.00270 | 0.00254 |
| | (41.54) | (104.0) | (0.00889) | (0.00779) |
| Attended Conference | | 75.62*** | | 0.00878*** |
| | | (16.04) | | (0.00239) |
| Attended Conference=1 × 2015=1 | | 36.69 | | 0.00512 |
| | | (57.06) | | (0.00539) |
| Black (Non-Hispanic)=1 × Attended Conference=1 × 2015=1 | | -23.19 | | -0.00561 |
| | | (93.98) | | (0.0104) |
| State, Year, and Grade FEs | × | × | × | × |
| Detailed Controls | × | × | × | × |
| Mean of Dep Var | 325 | 325 | .0261 | .0261 |
| Observations | 51,110 | 51,110 | 51,110 | 51,110 |

Notes: This table presents estimates of the effect of Black Lives Matter Messaging at Urbana 2015 on Black student donations and staff employment. Columns 1 and 3 present difference-in-difference (DD) estimates of outcomes for Non-Hispanic Black registrants compared to all other registrants for the 2015 conference, relative to the racial difference for conferences before 2015. Columns 2 and 4 present triple-difference (DDD) estimates of Non-Hispanic Black donations relative to other race-ethnic groups, for attendees relative to non-attendees, for registrants for the 2015 conference relative to registrants for conferences before 2015. All columns include detailed controls. Detailed controls include past amount of giving, past event and mission trip attendance, indicators for gender and race/ethnicity, college fixed effects, student leadership role indicators, membership in the sponsor organization or another organization, and number of days registered in advance of the conference. The number of observations with the detailed set of controls is reduced slightly due to omitted single-observation cells. Standard errors are clustered by college. Statistical significance reported as: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

# Bibliography

**Abaluck, Jason, and Jonathan Gruber.** 2016. "Improving the quality of choices in health insurance markets." National Bureau of Economic Research.

**Acemoglu, Daron.** 2002. "Technical change, inequality, and the labor market." *Journal of economic literature*, 40(1): 7–72.

**Agan, Amanda, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan.** 2022. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias."

**Albright, Alex.** 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions."

**Angelova, Victoria, Will Dobbie, and Crystal S Yang.** 2022. "Algorithmic Recommendations and Human Discretion."

**Angrist, Joshua D, Guido W Imbens, and Alan B Krueger.** 1999. "Jackknife instrumental variables estimation." *Journal of Applied Econometrics*, 14(1): 57–67.

**Anwar, Shamena, Shawn D Bushway, John Engberg, and RAND Social.** 2022. "The Impact of Defense Counsel at Bail Hearings."

**Arnold, David, Will Dobbie, and Peter Hull.** 2022. "Measuring racial discrimination in bail decisions." *American Economic Review*, 112(9): 2992–3038.

**Ashraf, Nava, Oriana Bandiera, Scott S Lee, et al.** 2015. "Do-gooders and go-getters: career incentives, selection, and performance in public service delivery." *Harvard Business School*, 1–60.

**Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett.** 2022. "The causal impact of removing children from abusive and neglectful homes." *Journal of Political Economy*, 130(7): 000–000.

**Baron, E Jason, and Max Gross.** 2022. "Is There a Foster Care-To-Prison Pipeline? Evidence from Quasi-Randomly Assigned Investigators." National Bureau of Economic Research.

**Bénabou, Roland, and Jean Tirole.** 2011. "Identity, morals, and taboos: Beliefs as assets." *The Quarterly Journal of Economics*, 126(2): 805–855.

**Bundorf, M Kate, Maria Polyakova, and Ming Tai-Seale.** 2019. "How do humans interact with algorithms? Experimental evidence from health insurance." National Bureau of Economic Research.

**Cao, Sean, Wei Jiang, Junbo L Wang, and Baozhong Yang.** 2021. "From man vs. machine to man+ machine: The art and ai of stock analyses." National Bureau of Economic Research.

**Carey, Iain M, Fay J Hosking, Tess Harris, Stephen DeWilde, Carole Beighton, and Derek G Cook.** 2017. "Hospital admissions." In *An evaluation of the effectiveness of annual health checks and quality of health care for adults with intellectual disability: an observational study using a primary care database*. NIHR Journals Library.

**Carrell, Scott, and Bruce Sacerdote.** 2017. "Why do college-going interventions work?" *American Economic Journal: Applied Economics*, 9(3): 124–51.

**Castro, Francisca, and Renata Retamal.** 2022. "Does electoral behavior change after a protest cycle? Evidence from Chile and Bolivia."

**Chan, David C, Matthew Gentzkow, and Chuan Yu.** 2022. "Selection with variation in diagnostic skill: Evidence from radiologists." *The Quarterly Journal of Economics*, 137(2): 729–783.

**Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu.** 2022. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions." 1–22.

**Children's Bureau.** 2021. "Child Maltreatment 2019." U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau.

**Choi, Sukwoong, Namil Kim, Junsik Kim, and Hyo Kang.** 2022. "How Does AI Improve Human Decision-Making? Evidence from the AI-Powered Go Program." *Evidence from the AI-Powered Go Program (April 2022). USC Marshall School of Business Research Paper Sponsored by iORB, No. Forthcoming.*

**Chorniy, Anna, and Christopher Mills.** 2022. "More Money, Fewer Problems? The Effect of Foster Care Payments on Children's Quality of Care."

**Clingingsmith, David, Asim Ijaz Khwaja, and Michael Kremer.** 2009. "Estimating the impact of the Hajj: religion and tolerance in Islam's global gathering." *The Quarterly Journal of Economics*, 124(3): 1133–1170.

**Cuddy, Emily, and Janet Currie.** 2020. "Rules vs. discretion: Treatment of mental illness in us adolescents." National Bureau of Economic Research.

**Currie, Janet, and Cathy Spatz Widom.** 2010. "Long-term consequences of child abuse and neglect on adult economic well-being." *Child maltreatment*, 15(2): 111–120.

**Currie, Janet, and Erdal Tekin.** 2012. "Understanding the cycle childhood maltreatment and future crime." *Journal of Human Resources*, 47(2): 509–549.

**Currie, Janet M, and W Bentley MacLeod.** 2020. "Understanding doctor decision making: The case of depression treatment." *Econometrica*, 88(3): 847–878.

**De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova.** 2020. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores." *CHI '20*, 1–12. ACM.

**Doyle, Joseph J.** 2007*a*. "Can't buy me love? Subsidizing the care of related children." *Journal of Public Economics*, 91(1-2): 281–304.

**Doyle, Joseph J.** 2007*b*. "Child protection and child outcomes: Measuring the effects of foster care." *American Economic Review*, 97(5): 1583–1610.

**Doyle, Joseph J.** 2008. "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care." *Journal of political Economy*, 116(4): 746–770.

**Edwards, Frank, Sara Wakefield, Kieran Healy, and Christopher Wildeman.** 2021. "Contact with Child Protective Services is pervasive but unequally distributed by race and ethnicity in large US counties." *Proceedings of the National Academy of Sciences*, 118(30): e2106272118.

**Fitzpatrick, Maria, Katharine Sadowski, and Christopher Wildeman.** 2022. "Does Access to an Algorithmic Decision-Making Tool Change Child Protective Service Caseworkers' Investigation Decisions?"

**Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al.** 2019. "Toward understanding the impact of artificial intelligence on labor." *Proceedings of the National Academy of Sciences*, 116(14): 6531–6539.

**Fujiwara, Thomas, Kyle Meng, and Tom Vogl.** 2016. "Habit formation in voting: Evidence from rainy elections." *American Economic Journal: Applied Economics*, 8(4): 160–88.

**Goldhaber-Fiebert, Jeremy D, and Lea Prince.** 2019. "Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office." *Pittsburgh: Allegheny County*.

**Green, Ben, and Yiling Chen.** 2019. "The Principles and Limits of Algorithm-in-the-Loop Decision Making." *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

**Grimon, Marie-Pascale.** 2022. "Effects of the Child Protection System on Parents."

**Gross, Max, and E Jason Baron.** 2022. "Temporary stays and persistent gains: The causal effects of foster care." *American Economic Journal: Applied Economics*, 14(2): 170–99.

**Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad.** 2020. "Managing intelligence: Skilled experts and AI in markets for complex products." National Bureau of Economic Research.

**Hoffman, Mitchell, Lisa B Kahn, and Danielle Li.** 2018. "Discretion in hiring." *The Quarterly Journal of Economics*, 133(2): 765–800.

**Hungerman, Daniel M, and Vivek S Moorthy.** 2020. "Every Day is Earth Day: Evidence on the Long-term Impact of Environmental Activism." National Bureau of Economic Research.

**Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin.** 2020. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment."

**Kahneman, Daniel.** 2011. *Thinking, fast and slow.* Macmillan.

**Khan, Muhammad Yasir.** 2020. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." *Unpublished manuscript.*

**Kim, Hyunil, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake.** 2017. "Lifetime prevalence of investigating child maltreatment among US children." *American journal of public health*, 107(2): 274–280.

**Kisat, Faizaan.** 2021. "Loan Officers, Algorithms, & Credit Outcomes: Experimental Evidence from Pakistan." *Algorithms, & Credit Outcomes: Experimental Evidence from Pakistan (October 21, 2021).*

**Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan.** 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics*, 133(1): 237–293.

**Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan.** 2016. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807.*

**Kling, Jeffrey R.** 2006. "Incarceration length, employment, and earnings." *American Economic Review*, 96(3): 863–876.

**Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan.** 2017. "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables." 275–284.

**Lambrecht, Anja, and Catherine Tucker.** 2019. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads." *Management science*, 65(7): 2966–2981.

**Larreboure, Magdalena, and Felipe González.** 2021. "The impact of the Women's March on the US House Election." *Documento de Trabajo IE-PUC*, , (560).

**Lieder, Falk, Thomas L Griffiths, Quentin J M Huys, and Noah D Goodman.** 2018. "The anchoring bias reflects rational use of cognitive resources." *Psychonomic bulletin & review*, 25(1): 322–349.

**Lowe, Matt.** 2020. "Religious Revival and Social Order." Working paper.

**Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott.** 2013. "Do political protests matter? evidence from the tea party movement." *The Quarterly Journal of Economics*, 128(4): 1633–1685.

**Montero, Eduardo, and Dean Yang.** 2021. "Religious Festivals and Economic Development: Evidence from Catholic Saint Day Festivals in Mexico." National Bureau of Economic Research.

**Mueller-Smith, Michael.** 2015. "The criminal and labor market impacts of incarceration." *Unpublished working paper*, 18.

**Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan.** 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science*, 366(6464): 447–453.

**Paxton, Pamela, Kristopher Velasco, and Robert W Ressler.** 2020. "Does use of emotion increase donations and volunteers for nonprofits?" *American Sociological Review*, 85(6): 1051–1083.

**Pierson, Emma, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer.** 2021. "An algorithmic approach to reducing unexplained pain disparities in underserved populations." *Nature Medicine*, 27(1): 136–140.

**Porter, Catherine, and Danila Serra.** 2020. "Gender differences in the choice of major: The importance of female role models." *American Economic Journal: Applied Economics*, 12(3): 226–54.

**Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan.** 2019. "The algorithmic automation problem: Prediction, triage, and human effort." *arXiv preprint arXiv:1903.12220*.

**Sinding Bentzen, Jeanet.** 2019. "Acts of God? Religiosity and natural disasters across subnational world districts." *The Economic Journal*, 129(622): 2295–2321.

**Snyder Jr, James M, and Hasin Yousaf.** 2020. "Making rallies great again: The effects of presidential campaign rallies on voter behavior, 2008-2016." National Bureau of Economic Research.

**Stevenson, Megan, and Jennifer L. Doleac.** 2019. "Algorithmic Risk Assessment in the Hands of Humans." *SSRN Electronic Journal*.

**Suzuki, Mizuhiro.** 2021. "Affording expensive ceremonies: Evidence from quinceaneras in Mexico."

**Wildeman, Christopher, Natalia Emanuel, John M Leventhal, Emily Putnam-Hornstein, Jane Waldfogel, and Hedwig Lee.** 2014. "The prevalence of confirmed maltreatment among US children, 2004 to 2011." *JAMA pediatrics*, 168(8): 706–713.