UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Human Capital

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Sungwoo Cho

2024

ABSTRACT OF THE DISSERTATION

Essays on Human Capital

by

Sungwoo Cho

Doctor of Philosophy in Economics

University of California, Los Angeles, 2024

Professor Adriana Lleras-Muney, Chair

This dissertation consists of three essays on human capital. In Chapter 1, I investigate the impact of collaborating with robots on human capital, focusing on professional baseball umpires who were provided with and then deprived of robot assistance. Umpires displayed enhanced accuracy with robot assistance, but experienced significant declines in performance once it was removed. I argue that these findings can extend beyond baseball, suggesting broader implications for occupations sharing similar skill sets. In Chapter 2, joint with Felipe Gonçalves and Emily Weisburst, we investigate the effects of changes in workplace risk awareness on police behavior and public safety, focusing on incidents of police officer fatalities while on duty. Our findings reveal that following the death of a fellow officer, police officers reduce arrest activity for one to two months, indicating heightened fear. This reduction is most prominent for minor offenses and is more pronounced in smaller cities. Yet, we find no evidence of increased crime rates during this period, suggesting that reduction in arrests does not adversely impact public safety. In Chapter 3, joint with Anna Aizer, Shari Eli and Adriana Lleras-Muney, we utilize newly gathered data of 16,000 women who applied for Mothers' Pensions, America's first welfare program, to explore the impact of means-tested cash transfers on lifetime family dynamics and maternal welfare. In the short term, these transfers led to postponed marriage and reduced geographical mobility. However, in the long run, they had no discernible effect on remarriage probability, spouse quality, fertility, or maternal well-being, as measured by longevity and family income in 1940. With the absence of significant negative behavioral outcomes, we suggest that the benefits of such transfers may outweigh the costs, particularly if they yield even modest positive impacts on children.

The dissertation of Sungwoo Cho is approved.

Martha Jane Bailey

Felipe M. Gonçalves

Emily Karen Weisburst

Adriana Lleras-Muney, Committee Chair

University of California, Los Angeles

2024

To my family.

Contents

List of Figures

List of Tables

Acknowledgements

I wish to express my sincere appreciation to the many individuals for their invaluable support in completing this dissertation.

I extend my deepest gratitude to my advisor, Adriana Lleras-Muney, for her unwavering guidance and encouragement throughout this journey. Her invaluable insights and dedication have been instrumental in shaping not only the trajectory of this dissertation, but also my career. I cherished every conversation with her and aspire to one day become an advisor as supportive and insightful as she was to me. I am also very grateful to Felipe Gonçalves and Emily Weisburst for placing their trust in me, offering guidance and support and acquainting me with field of the economics of crime. Working with them has been an invaluable learning experience and I am profoundly thankful for the numerous opportunities they have provided me. I am also indebted to Martha Bailey for her encouragement and insight, which have played a significant role in shaping my academic journey. Her mentorship has been invaluable, and I am profoundly grateful for her guidance.

I would not have been able to complete two chapters in this dissertation without my coauthors. Chapter 2 is joint work with Felipe Gonçalves and Emily Weisburst that is submitted and Chapter 3 is a version of "The Impact of Cash Transfers to Poor Mothers on Family Structure and Maternal Well-Being," (joint with Anna Aizer, Shari Eli and Adriana Lleras-Muney) *American Economic Journal: Applied Economics* 16.2 (2024): 492-529. I would like to thank them all for all their work.

My experience at UCLA has been greatly enriched by the unwavering support of the numerous faculty members I have had the privilege of interacting with. I am especially grateful for the invaluable feedback provided by Daniel Haanwinckel, Kareem Haggag, Juliana Londoño-Vélez, Yotam Shem-Tov, Bernardo S. Silveira and participants in the Albert Family Fund Proseminar in Applied Microeconomics. I also have been very fortunate to have made wonderful friends during my six years in the program. I would like to thank them for our discussions and friendship. I would also like to thank Chiara Paz, Alice Wang and many staff members in the department for all of their support. I also acknowledge that I benefited from facilities and resources provided by the California Center for Population Research at UCLA (CCPR) and from financial support from the Economics Department at UCLA and from the Graduate Division at UCLA.

Finally, I would like to thank all my friends whom I have met throughout my life and who were always there for me. My deepest gratitude goes to my family for their boundless love and unwavering support during my Ph.D. Their encouragement and belief in me have been invaluable, and I could not have accomplished this without them. I will be forever grateful.

Vita

Education

M.A. Economics, Duke University	2018
B.S. Economics *with Distinction and High Honors*, University of Michigan, Ann Arbor	2016

Publications:

"The Impact of Cash Transfers to Poor Mothers on Family Structure and Maternal Well-Being" (with Anna Aizer, Shari Eli and Adriana Lleras-Muney), *American Economic Journal: Applied Economics*, 16 (2): 492–529.

Honors, Fellowships & Grants

Best Paper Award, Albert Family Proseminar in Applied Microeconomics, UCLA	2023
Graduate Representative, SEA Annual Meeting	2023
Dissertation Year Fellowship, UCLA	2023–2024
Summit Fellowship in Applied Economics, UCLA	2023
Lewis L. Clarke Graduate Fellowship, UCLA	2022
Graduate Research Mentorship, UCLA	2020–2021
Graduate Summer Research Mentorship, UCLA	Summer 2020
Outstanding Teaching in Remote Learning, UCLA	Winter 2020
Departmental Fellowship, UCLA	2018–2023
M.A. Merit Scholar Award, Duke University	2017
Duke Economics Master's Scholar Award, Duke University	2016–2018
Phi Beta Kappa, University of Michigan	2016

Presentations

University of Pennsylvania (Wharton, AI and the Future of Work)	2024
Data-Driven and Experiment-Based Evidence for Policy	2024

Monash University	2024
Hong Kong University of Science and Technology	2024
United States Military Academy at West Point	2024
Institut Mines-Télécom Business School	2024
Southern Economic Association	2023
Korea-America Economic Association (Job Market Conference)	2023
Association for Public Policy Analysis and Management	2022
Southern Economic Association	2021
American Law and Economics Association	2021

Relevant Experiences

Instructor, UCLA	2022–2023
Teaching Assistant, UCLA	2019–2024
Research Assistant to Professor Adriana Lleras-Muney, UCLA	2020–2021, 2022
Research Assistant to Professor James W. Roberts, Duke University	2016–2017
Legislative Assistant, National Assembly of the Republic of Korea	Summer 2014

Introduction

This dissertation consists of three essays in the intersection of labor economics and public finance with a particular focus on human capital and crime.

In Chapter 1, I ask how does working with robots change human capital? To examine how collaboration with robots affects human skills, I exploit a unique setting in which professional baseball leagues provided, and subsequently removed, access to robot assistance for umpires. Umpires demonstrated improved precision and accuracy in ball-strike decisions while using robot assistance, and their performance declined substantially below preassistance levels after it was removed. Both highly skilled and inexperienced umpires exhibited large declines in performance after the removal of robot assistance. Umpires who used robot assistance for longer periods of time faced a steeper decline in accuracy than those who used it for shorter periods. In addition, umpires who worked a full season with robot assistance did not fully return to their initial skill level by the end of the following season. By examining a canceled season during the COVID-19 pandemic, I reject that skill depreciation is solely a result of umpires simply not using their skills. Umpires also experience skill deterioration in determining whether a baserunner is safe, suggesting that the findings are widely applicable to various occupational settings with a similar skill set.

In Chapter 2, co-authored with Felipe Gonçalves and Emily Weisburst, we examine how changes in the salience of workplace risk affect police behavior and public safety. Specifically, we investigate cases of police officer deaths while on duty. Officers respond to a peer death by decreasing arrest activity for one to two months, consistent with heightened fear. Reductions are largest for low-level arrests and are more pronounced in smaller cities. Crime does not increase on average during this period, nor do we observe crime spikes in cities with larger or longer arrest declines. While shocks to perceived fatality risk generate substantial enforcement responses, officer fear is unlikely to harm public safety.

In Chapter 3, co-authored with Anna Aizer, Shari Eli and Adriana Lleras-Muney, we use newly collected data for 16,000 women who applied for Mothers' Pensions, America's first welfare program, to investigate the effect of means-tested cash transfers on lifetime family structure and maternal well-being. In the short-term, cash transfers delayed marriage and lowered geographic mobility. In the long run, transfers had no impact on the probability of remarriage, spouse quality or fertility. Cash transfers did not affect women's well-being, measured by longevity and family income in 1940. Given the lack of significant negative behavioral impacts, the benefits of transfers appear to exceed costs if they have, even modest, positive impacts on children.

Chapter 1

The Effect of Robot Assistance on Skills

> *"If men learn this, it will implant forgetfulness in their souls; they will cease to exercise memory because they rely on that which is written..."*
>
> ***- Socrates on writing, from Plato's Phaedrus***

1.1 Introduction

As early as the time of Socrates, humans have harbored a fear that new technologies may deteriorate human knowledge and skills. Recent popular discussion concerns the potential impacts of industrial robots and artificial intelligence (AI) technologies on human capital. Instead of fully displacing humans, however, numerous robots and technologies may collaborate with humans in practice, enhancing and complementing their work. Combining human expertise with robot assistance has the potential to yield gains that cannot be fully achieved by exclusively depending on either of them alone. However, heavy reliance on technology could lead to decreases in workers' skills and increases in workers' technological dependence. The deterioration of human capital caused by dependence on technology is especially important in the labor market and can carry significant consequences. For example, while automated systems improved safety records in the airline industry, pilots now experience difficulties manually operating aircraft, contributing to some recent disasters (Nicas and

Wichter, 2019).[1] Similarly, doctors' physical diagnosis skills have declined with increasing reliance on technologies, shifting focus away from patients (Aw, 2014).[2]

Even in a world where robots are readily available, human capital without robot assistance is essential. First, during emergencies or unexpected events, humans excel in making rapid, context-sensitive decisions and improvising solutions (e.g., Captain Sullenberger and the 2009 Hudson River plane crash). Humans rely on these abilities to manage unforeseen situations where robots may struggle to respond effectively. Second, in many sectors, particularly healthcare, education, and customer service, human interaction and empathy are necessary. Robots lack the capacity for human connection, making human capital indispensable for roles that require understanding and compassion. Third, human oversight and intervention are crucial in industries where judgment and ethics are paramount. Improving human capital is vital to ensure the quality of robot-assisted decisions and the ethical conduct in these industries.

In this paper, I ask three questions: First, what is the impact of robot assistance on individual workers' skill and productivity? Second, who are the biggest losers of adoption of robot assistance? Third, does longer work duration with robot assistance increase skill depreciation, and do workers bounce back and regain skills upon removal of the robots?

Examining the impact of technology assistance on human capital is empirically challenging. First, direct measures of individuals' skill and productivity are often unavailable. Even when data are available, it is difficult to clearly identify the effect of robots on individual skills versus measuring skills without robot assistance because individuals do not randomly adopt or drop robot assistance. Firms and humans choose robot assistance if they believe that it

[1] Over 60% of recent accidents can be attributed to pilots who have lost their practical skills. For example, an investigation by the National Transportation Safety Board identified pilot error as the primary contributing factor to the Asiana Airlines Flight 214 crash that occurred on July 6, 2013. The pilots did not properly monitor airspeed and altitude, and they allowed the aircraft to sink dangerously low and stall just before landing. One contributing factor was the overreliance on automation in the cockpit.

[2] Gong et al. (2019) argue that medical students now avoid specializing in radiology due to fear of job loss caused by AI assistance. However, Agarwal et al. (2023) find that radiologists-AI collaboration is still suboptimal due to radiologists' biases against AI.

improves their performance and stop using it when they do not. In addition, the durations of robot assistance or unemployment spells are usually not random, making it difficult to study the effect of duration on skill depreciation as those with lower skills may choose to be assisted for longer time, or those with lower skills may take longer time to be reemployed.

I address these challenges by studying the implementation of the automated ball-strike system (ABS, or "robot umpire") in professional baseball leagues. Professional baseball leagues provide several advantages when studying the impact of robot assistance. First, umpires are professional decision-makers with similar job tasks as judges and make decisions about whether a thrown pitch is a strike or a ball.[3] In this setting, I have unique and precise measures of individual productivity: I observe the decisions solely made by the umpire as well as whether the decision is correct or not as determined using post-processed data. Second, there is substantial variation in the implementation of robots. There are 11 different leagues in the Minor League Baseball (MiLB) system, with each league providing a stage for experimental rule changes. The implementation of robot assistance only occurred in the Single-A Florida State League starting in 2021 and the Triple-A Pacific Coast League in 2022. With umpires moving across leagues, I can observe worker-specific usage of technology and their behaviors before and after robot adoption, as well as *with* and *without* robot assistance. The *removal* of robot assistance enables a unique opportunity to evaluate human capital in isolation, disentangling it from the influence of human-robot collaboration. Third, a change to intermittent robot utilization in 2023 provides an ideal setting to study how skills depreciate as a function of how long individuals have been assisted by robots. In 2023, Triple-A leagues imposed a rotation system in which half of the season utilized robot assistance, and the other half did not. Umpires are quasi-randomly assigned to games and thus, the duration of robot assistance throughout a season is as-good-as random.

I examine over 62,000 professional baseball games played between 2017 and 2023 in Minor and Major League Baseball (MLB) and employ a difference-in-differences design exploiting

[3] In baseball, a "strike" is a pitch within the defined strike zone, and a "ball" is a pitch outside of this zone.

the staggered adoption of robots across leagues and the movement of umpires across leagues. This approach compares the change in the performance of umpires before and after being assisted by robots to the change in performance of umpires that are never assisted by robots.

First, I find that umpires perform with higher precision and accuracy when assisted by a robot. In the season with robot assistance, umpires' decision accuracy increases by 6.4 percentage points, on average. In particular, both Type I errors (e.g., incorrectly calling a strike) and Type II errors (e.g., incorrectly calling a ball) decrease significantly. Using an event-study specification, I confirm that there are no pretrends suggesting that the timing of robot implementation is as-good-as random and that umpire moves are not correlated with the adoption of robot assistance. While the umpires of different prior skill levels all benefit from robot assistance, the skills distribution, with robots, compresses by 36%, consistent with previous work on ChatGPT, surgical robots, and AI programs reducing skills gaps among workers, surgeons, and taxi drivers, respectively (Brynjolfsson et al., 2023; Kanazawa et al., 2022; Noy and Zhang, 2023; Tafti, 2022). When restricting attention to more ambiguous decisions, robot assistance improves the umpire's accuracy by 12.3 percentage points. Further, umpires suffer less from decision biases with robot assistance: negative autocorrelation in decisions, or the "gambler's fallacy", is reduced by almost 100% and omission bias, a tendency to call more pitches strikes when the next called ball would end an at-bat (i.e., to avoid issuing a walk), by 89%.

However, umpires who are previously assisted by robots experience significant skill declines when the robot is removed. When making calls without robot assistance, umpires suffer a decline of 2.0 percentage points in accuracy relative to preassistance levels. The size of the decline is roughly equal to the gap between the median and the bottom 5th percentile umpires. For pitches that are harder to judge, accuracy declines by 3.7 percentage points. Umpires also suffer more from decision biases showing increasing signs of omission bias by about 33%. Further, I find that robot-induced deterioration in skills differs across umpires with the largest skill decline occurring for highly skilled umpires. When the robot assistance is removed, the lower-skilled umpires' skills decline by a statistically insignificant 0.9 percentage points, whereas the high-skilled umpires lose skill by 4.7 percentage points. The distribution of skills across the umpires is compressed by 64.6% following the work with robot assistance at the expense of high-skilled umpires relative to the preassistance period.

Further, umpires subject to a full season of work with the robot do not bounce back quickly when recalled to the task without a robot. In the first game back, the umpires sustained a large skill decline of 3.6 percentage points. The decline becomes smaller as the season progresses: it is 1.7 by the middle of the season and 1.0, and still marginally significant, by the end of the season. These results demonstrate that umpires require nearly an entire season to approach their initial skill level after a loss following a season of robot assistance and never fully recover.

Understanding how skill declines with the intensity and the length of the robot assistance provides important insight into the trade-offs of assigning robot assistance to workers and the transitions of individual workers after assistance stops. Using the 2023 Triple-A implementation of robot assistance, I find that umpires who used robot assistance for longer periods experience more significant skill depreciation when making decisions without robot assistance. Umpires assisted by robots for just one game suffer a marginally significant skill decline of 0.3 percentage points while those assisted for two or three consecutive games experience a larger magnitude decline of 1.2 and 1.0 percentage points, respectively. The baseline results suggest that a full season of work with robot assistance induced a decline of 3.6 percentage points in skill in the first game back. This implies that as the duration of robot assistance increases, the decline in skill becomes more pronounced.

A related question is whether the skill depreciation is a result of the umpires simply not using their skills. If the decline in skill was caused by a lack of practice alone, one would expect at least some decline in skill based on results demonstrated in past literature on job loss and human capital depreciation (Benhenda, 2022; Dinerstein et al., 2022; Edin and Gustavsson, 2008; Jarosch, 2023). To examine whether robot assistance has a distinct impact from time off work, I examine a different treatment: some umpires abruptly experienced a year-long interruption in their careers due to the COVID-19 pandemic. In 2020, the Minor League canceled the season whereas the Major League had a delayed and shortened season. Minor League umpires experienced a marginally significant skill decline of 0.8 percentage points in the first month of the return relative to the Major League umpires. Taken together, these results imply that the effect of time away from the task is smaller than the effect of robot assistance, perhaps due to reduced incentives to practice skills from the introduction of advanced technology.

Home-plate umpires experience a sharp decline in their ability to accurately call pitches once they have worked with robot assistance, raising a question of whether it impacts other skills as well. Umpires are tasked with various responsibilities as they rotate through different positions such as those of the first-base umpire and home-plate umpire. I examine whether the skill as the first-base umpire in deciding whether a baserunner is safe or out when attempting to reach first base declined following the work with the robots. The skill required to perform this task closely resembles that of the home-plate umpire as it demands sharp visual perception. Umpires who have previously worked alongside robots experience an additional 0.1 replay review request per game (44.3%) when they officiate at first base, which is entirely due to an increase in overturned challenges. Collectively, these findings strongly imply that umpires with prior experience with robot assistance demonstrate higher rates of inaccuracies and a decline in their skills in this role, also suggesting transferability to a wide range of other occupational settings that require similar skill sets.

An important and interesting question is whether other aspects of the game are also affected by robot adoption. Robots might have affected the performance of other workers, changing the nature of the game and suggesting important complementarities across workers. If this occurred, however, then the effects of robots that I estimate could reflect both the effects of the robots and the effect of the responses of other players. Indeed, players exhibit different strategic responses to robot adoption in the game. To optimize productivity with an umpire assisted by a robot, pitchers adjust their strategy by altering their pitching behavior to aim at a specific region of the strike zone where the umpire is now more likely to call strikes. The findings regarding skill depreciation of umpires are, however, robust to reweighing the data to address changes to pitch distributions that the umpires face.

Finally, I show that the implementation of robot technology has positive impacts on professional baseball. Compared to the leagues that did not use robot assistance, leagues that adopted it experienced an increase of 12.4% in attendance, perhaps due to increased public attention to the technology. With an average family of four spending about $65 to attend a game, increased attendance translates to an increase in revenue ranging from $1.25 to $2.5 million. A conservative estimate still suggests that the league profits from adopting robot assistance despite the cost of the technology.

This study directly relates to several large strands of literature. First, this work provides new insights into understanding the effect of robot adoption on individual workers. Past studies have focused and found mixed evidence on the type of workers affected by the adoption and how it affects them typically using aggregate industry-level data. Acemoglu and Restrepo (2020), Acemoglu et al. (2023), Dauth et al. (2021), Acemoglu et al. (2020) and Humlum (2022) find negative effects on wages and employment especially among workers in the manufacturing industry in the US, Netherlands, Germany, France and Denmark, respectively. Similarly, Bonfiglioli et al. (2020) and Barth et al. (2020) find that low-skilled workers are negatively affected by robot adoption. However, Aghion et al. (2020) and Hirvonen et al. (2022) find non-negative effects even among low-skilled workers. These studies do not measure productivity directly, but instead rely on wages using aggregate data. I study a different question of what happens to workers when they are assisted rather than replaced by robots.

Past literature that relates closely to my study includes Kanazawa et al. (2022), Brynjolfsson et al. (2023), Tafti (2022), and Noy and Zhang (2023), which find the positive productivity effects of AI assistance programs, surgical robots and ChatGPT. They also suggest compressing effects on the productivity distribution with technologies. Unlike their studies, I am able to study what happens to skills after the *removal* of the technology. These results help us understand further the extent to which workers learn to rely more on those technologies. I also find that the distribution of skills contracts following robot implementation with better-skilled workers facing the larger skill decline. Collectively, highly skilled workers not only experience the least benefit from robots but also face the most significant losses when the robot is removed.

Second, my findings relate to the literature in labor economics on skill depreciation and duration dependence. Human capital depreciates when workers are unemployed or absent (Benhenda, 2022; Dinerstein et al., 2022; Edin and Gustavsson, 2008; Jarosch, 2023). Related literature on non-employment duration dependence also finds negative effects of the lengthy non-employment spells on callback and reemployment rates and on wages upon reemployment (e.g. Jacobson et al., 1993; Kroft et al., 2013; Maestas et al., 2015). However, estimating human capital depreciation is difficult due to the correlation between the duration of unemployment and the productivity of workers (i.e., those with lower skills may take longer time to be reemployed) (See Machin and Manning, 1999, for discussion). Two related studies use novel aggregate datasets exploiting quasi-randomly assigned time spent in employment (Dinerstein et al., 2022) and panel data tracking skills (Cohen et al., 2023) to address this challenge. They, however, reach mixed conclusions as Cohen et al. (2023) find no decline in cognitive and noncognitive skills during an unemployment spell, but Dinerstein et al. (2022) find productivity declines from not working among Greek teachers. I leverage the career pause brought on by COVID-19 to identify a modest dip in umpire skills. In contrast, I find larger skill depreciation following the removal of robot assistance. Further, using quasi-random variation in work duration with the robot, I study a similar, but different setting of robot assistance to find that longer periods with robot assistance induce larger depreciation of skills.

Third, the literature on peer effects in the workplace shows the positive effects of coworkers (Falk and Ichino, 2006, using lab experiments; Mas and Moretti, 2009, among supermarket cashiers; Cornelissen et al., 2017, in Germany; Cardoso et al., 2018, in Portugal; Hong and Lattanzio, 2022, in Italy). A notable exception is Guryan et al. (2009) which finds no evidence of peer effects among professional golfers. I document the negative effect of the removal of robot assistance which can be thought of as a functional and collaborative “peer." The finding suggests that productive peers can have detrimental effects on skill retention and can increase future dependence on the peer assistance. Additionally, I also find that players, who are *indirectly* affected, also strategically adjust their behaviors.

The firm-level analysis of robot adoption finds increases in output gains and productivity (e.g. Acemoglu et al., 2020, 2022; Dixon et al., 2021; Humlum, 2022; Koch et al., 2021). However, firms that implement robots are observably different from the firms that do not as they are often larger and more productive firms. I contribute here by studying quasi-random assignments of robots to different leagues and documenting an increase in attendance and total scores in a game that translates into large revenue gains for the league.

Finally, economics has a longstanding tradition of employing specific occupations as a context for studying fundamental economic questions. This approach not only provides valuable insights into the functioning of specific sectors but also offers broader lessons about economic principles and policies that can be applied to a wide spectrum of contexts. For example, orchestral musicians have been used to test for discrimination (Goldin and Rouse, 2000), taxi drivers have been examined to test labor supply models (e.g. Farber, 2005), and cashiers have been analyzed to explore peer effects (Cornelissen et al., 2017). Many sports settings have also been used to study various economic theories, ranging from game theoretical predictions to discrimination and corruption (e.g. Chiappori et al., 2002; Duggan and Levitt, 2002; Malueg and Yates, 2010; Price and Wolfers, 2010; Price et al., 2013). While I analyze the impact of robot assistance on umpires' skills in making ball-strike decisions, the skill sets and cognitive processes involved in these decisions extend beyond the realm of baseball. As the skill sets are commonly shared by professionals in various domains, the findings can provide broader lessons that are applicable to a wide range of contexts.

The rest of the paper is organized as follows. Section 1.2 describes the umpire's decision and explains the implementation of robots and Section 2.2 introduces the data. Section 2.3 presents the empirical strategy. Section 2.4 shows the main results and Section 1.6 discusses the potential mechanisms of the findings. Section 1.7 shows additional results and discusses external validity of the results. Section 2.7 concludes.

1.2 Background

1.2.1 Umpire Decisions

Baseball umpires are professional decision makers applying official rules to the game as it is played.[4] Becoming a professional baseball umpire requires knowledge of the intricacies of the game rules, experience, and fitness. Formal training programs and umpire schools offer instructions on every area of the game.[5] Many umpires start at amateur levels and work their way up into professional baseball leagues and ultimately to Major League Baseball.

A baseball season starts in mid-February with spring training, and the regular season starts around the last week of March. The regular season consists of 162 games for the Major League and between 132 and 150 games for the Minor League. The season concludes with the postseason of up to 22 games. On average, a Major League Baseball umpire works in about 112 regular season games, of which 28 are behind the home plate, calling balls and

[4] Umpires' decisions have been used to test for the gambler's fallacy (Chen et al., 2016), racial discrimination (Parsons et al., 2011), rational inattention (Bhattacharya and Howard, 2022), attention scarcity (Archsmith et al., 2021) and status bias (Kim and King, 2014). These decisions have also been used to find that increased monitoring improves productivity (Mills, 2017), and that hotter temperatures decrease skills (Fesselmeyer, 2021). The availability of data and identification advantages sports settings offer contributes to the production of credible evidence for addressing numerous challenging research questions (See Kahn, 2000). Umpires' decisions are particularly useful tools as they provide key advantages over other settings, namely, observing precise individual-level decisions and whether they are correct or not.

[5] Attending an umpire school is a necessary step for those aspiring to umpire in the Major League and Minor League. It is extremely difficult for an umpire to reach the Major League. For example, out of about 150 students in a class, only about 20 will receive recommendations to advance to the umpire evaluation course conducted by the Professional Baseball Umpires Corporation (PBUC). Additionally, about 45 other students will be designated for placement in independent leagues. One umpire from each class will become a Major League umpire, on average.

strikes. Umpires work as a crew of usually four umpires, each working behind home plate and behind first, second, and third bases. The umpire crew rotates positions throughout the series of games, resulting in a quasi-random assignment of umpires to games.

The primary task of the umpires is rule enforcement.[6] Home-plate umpires are tasked with determining whether thrown pitches are balls or strikes.[7] These split-second decisions could sometimes be subject to human error and subjective judgment, undermining accuracy, fairness, and consistency in determining the outcome of each pitch.

In particular, when a pitch is thrown and the batter does not swing, the umpire must determine if the ball crosses home plate through the strike zone, resulting in a strike, or if it crosses home plate outside the zone, resulting in a ball being called (See Figure A.1).[8] The strike zone is defined as an imaginary rectangular region over the home plate that extends roughly between the batter's shoulders and kneecaps, of roughly 20 by 25 inches in dimension. An umpire's ball-strike decisions are critical aspects of the game, as they directly affect the count of balls or strikes on the batter.[9] The count influences the strategies of both the pitcher and the batter, making it a pivotal element in the game's dynamics.

With umpires' decisions being such vital parts of the game, players, coaches, and fans closely monitor and scrutinize them for consistency and accuracy. Often, disagreements between players or coaches and umpires arise. Fans also criticize umpires for poor calls on even a single pitch. The Major League employs umpire supervisors and observers, and also uses the Supervisor Umpire Review and Evaluation (SURE) system to evaluate umpires. Umpires receive "report cards" following the game and also receive mid-year and postseason evaluations.[10] Therefore, umpires must maintain consistent and accurate strike zones within

[6] See Section A.2 for a short description of a baseball game.

[7] A pitcher throws a ball from the mound to the catcher, who sits 60 feet and 6 inches away. A fastball thrown at 100 miles per hour takes about 0.4 seconds to reach the catcher.

[8] All pitches are subject to the umpire's judgment. However, the umpire makes a call only when the batter does not swing.

[9] The count refers to the number of balls and strikes on the batter. For example, if the umpire calls two strikes on the batter, the count is "0-2," meaning there are no balls and two strikes.

[10] See https://umpscorecards.com/home/ for examples regarding how umpire performance is measured by

and across games.

1.2.2 Robot Assistance

Despite their efforts to strive for consistency and high accuracy, umpires often make controversial calls.[11] Because umpires' decisions are subject to human error and judgment, Major League Baseball started an experiment with an automated ball-strike system (ABS, or robot umpires) that could provide reliable assistance, ensuring accuracy, fairness, and consistency.

The robot utilizes sophisticated hardware and advanced software algorithms to track pitches and determine whether a pitch passes through the strike zone. As part of the "Hawk-Eye" tracking system, multiple cameras placed around the stadium capture the ball's flight path from different angles after a pitch is thrown. The system then identifies the pitch location and uses a calibrated strike zone, personalized for each batter, to make a ball-strike decision. The call is then communicated through an earpiece to an umpire who is standing behind home plate to physically make the decision.[12]

In 2019, Major League partnered with the independent Atlantic League of Professional Baseball (ALPB) to begin experimenting with robot assistance. Positive results in consistency and reliability led to an expansion of the experiment to Minor League; the Single-A Florida State League adopted it first in 2021, followed by the Triple-A Pacific Coast League in

fans.

[11] A controversial called third strike can directly change the result of the game. For example, a game between the San Francisco Giants and the San Diego Padres on September 28th, 2020 had 27 incorrectly called pitches that benefited the Padres by 1.85 expected runs in a game they won by 1.

[12] The home-plate umpire still stands behind the plate and makes decisions on other aspects of the game. The home-plate umpire can overrule the decision communicated by the robot assistance for an obvious mistake.

2022.[13,14] In 2023, both of the Triple-A Leagues (i.e., the International and Pacific Coast Leagues) implemented robot assistance in games held between Tuesday and Thursday of the week.[15,16]

1.3 Data

This study uses the universe of Major League and Minor League pitch-level data from the 2017 season to the 2023 season. The data are web-scraped from the MLB Stats API which provides a wide range of statistics and data related to all games and individuals involved in the league.[17]

First, I collected data from about 73,000 games, including the dates, weather conditions, time durations, attendances, teams and leagues, venues, and importantly, the umpires overseeing the games. I also collated information on personnel associated with each game that includes the names, ages, experiences, height, and handedness of the umpires and individual players.

I then gathered play-by-play data for over 22.5 million pitches that include detailed pitch characteristics.[18] Specifically, X and Y pitch coordinates of every pitch as it crosses the plate

[13] Minor League is a professional league below the Major League that is divided into four classes: Single-A, High-A, Double-A and Triple-A. Each class includes multiple leagues. Major League Baseball teams utilize the minor leagues to develop young players; players move through the ranks to eventually play in the Major League. Umpires also can be promoted and demoted through the major-minor league system. There are about 230 professional umpires working in both major and minor leagues. Compensation also varies significantly across levels: an average umpire in the lowest league earns about $3,000 per month, while a Major League umpire receives over $10,000 on average.

[14] On July 20, 2021, the league adjusted the size of the strike zone following inputs from players and umpires. The adjusted strike zone is only applied to the games utilizing robot assistance and not to other games. During those games, the strike zone is wider and shorter than the original strike zone. See Figure A.2.

[15] Games held between Tuesday and Thursday of the week used robot assistance, while games between Friday and Sunday were called by umpires without assistance. The system was adopted to compare two formats and to prepare Triple-A umpires for when they appear in Major League games as replacement umpires.

[16] Robot implementation is described in more detail in Section A.3. Also, see Figure A.3 for chronology of these events.

[17] The data are provided by the Statcast system.

[18] In a typical 9-inning game, about 300 total pitches are thrown, of which about 150 are called.

and the top and bottom of the strike zone are available.[19] Further, detailed descriptions of a pitch outcome (e.g., called strike, called ball, pitch resulted in hit or out, etc.) are collected. The main outcome is whether an umpire correctly called a pitch or not. To determine this, I utilize the exact pitch location, the dimension of the strike zone, and the umpire's call decision to generate the decision accuracy measure.[20] Various game-situational characteristics like ball counts and out counts, and whether there are runners on base at the moments of these counts, are also collected.

To generate the estimation sample, I drop non-regular season games that include spring training, All-Star Games, other exhibition games, and postseason games. I also drop games with missing personnel information, games in which I cannot identify whether a robot is used, and games with too few pitches.[21] Finally, I also drop outlier pitches that are inaccurately recorded due to calibration issues. The final dataset has over 62,000 games and 18.5 million total pitches, of which 8.8 million are called with about 680 umpires and 7,500 pitchers and batters.

1.3.1 Summary Statistics and Preliminary Evidence

Table 1.1 summarizes the pitch characteristics of the sample, separately by whether a game used robot assistance or not. Out of 62,678 games in the sample, 2,611 games are called with robot assistance. In these games, umpires are harsher relative to when the assistance is not used and call fewer strikes (31.2% vs. 33.6%). These differences in the calls may be due to the differential responses of pitchers: when pitchers know the robot is assisting to make a call, they tend to pitch further away from the strike zone (0.78 vs. 0.75 feet horizontally and 0.87 vs. 0.85 feet vertically). In other words, pitchers are more confident about receiving a strike

[19] In stadiums with advanced cameras equipped, the Statcast system tracks the pitches with an accuracy of better than one inch. See Section A.4 for details on how pitches are tracked in other stadiums.

[20] The X-and Y-coordinates utilized to generate the decision accuracy measure come from the MLB Stats API and are different from the coordinates plotted by the robot. These are also post-processed to correct for potential errors.

[21] In 2022, the Major League announced that the "select" games in Single-A Florida State League will use the "challenge" system detailed in Section A.3. However, these games are not specified in the data.

call at the edge of the strike zone with robot assistance. On average, umpires with assistance are no better than without in accuracy (93.0% vs. 93.0%), but after accounting for the pitch location and the stadium, they perform better (0.015 vs. -0.0007).[22] When a robot assists umpires calling the game, players have no incentive to argue the ball-strike decision and risk being ejected. Therefore, the number of ejections is fewer in these games (0.04 vs. 0.07).

Figure 1.1 plots the average accuracy of umpires with and without assistance as a function of the pitch's distance from the nearest border of the strike zone. Both with and without robot assistance, umpires perform extremely well when the pitches are obviously inside or obviously outside the strike zone. However, accuracy declines as the pitch gets closer to the border. Umpires with assistance are more accurate on these pitches; while not perfect, for outside pitches that are 0.2 inches away from the nearest border, they have an average accuracy of 77.4% relative to accuracy of 64.5% without robot assistance. Figure 1.2 reveals distinctive discrepancies between umpires with and without assistance. The "enforced" strike zone without a robot is oval-shaped despite the actual strike zone being rectangular. Umpires without assistance incorrectly call pitches that fall within the corners. Robot assistance, on the other hand, helps the umpires match the rule-defined strike zone well.[23]

This difference in the "enforced" strike zone leads the players to strategically adjust their behaviors. In games assisted by robots, pitchers pitch further away from the center of the strike zone (1.07 vs. 0.99 feet) on average, and pitches are also more likely to fall on the "edge" of the strike zone (61.8% vs. 59.5%) (Table A.1). The adjustment also increased "base on balls" in the games in these games, as shown by pitcher's walks allowed (2.20 vs. 1.94 per game) and batter walks (0.54 vs. 0.42 per game).[24]

[22] Robot assistance is not perfect for several reasons. Any technology has a margin of error. The system relies on a radar system that tracks the pitches, and it sometimes performs suboptimally in a crowded environment. Errors could also be due to calibration issues in these environments in locating pitches, and measuring the size of the strike zone depends on the batter's height and stance. These errors likely happen at the stadium-level, so I employ a team-by-year fixed effect to address this issue.

[23] The robot-assisted umpire fails the "Turing test," which is a test to assess a machine's ability to exhibit human-like intelligence or behavior.

[24] A batter's at-bat ends in a walk (or "base on balls") when the count reaches 4 balls. The batter is awarded first base by the umpire.

The estimation focuses on umpires. Figure A.4 plots the raw accuracy for umpires who are assisted by robots and those who are not in the year with robot assistance and years prior and after by pitch location. Panel A shows that the Major League umpires improve their accuracy slightly over the years. In contrast, Panel B displays a large increase in accuracy across pitch location when robot assists calling the game, and a drop in the year following relative to pre-robot year. This provides suggestive evidence that umpires' skills deteriorated following the implementation of the robot below preassistance levels.

1.4 Empirical Strategy

Main Model I exploit the staggered adoption of robot assistance over time and umpires moving across leagues in a difference-in-differences framework.[25] I observe umpires *before and after* the implementation of robots and compare them to those who were not assisted by the robot. Figure A.5 describes the potential umpire moves across leagues and robot implementation.[26] A total of 17 umpires called games in the Single-A Florida State League in 2021, 11 with more than 10 games with a robot, and 66 umpires called games in the Triple-A Pacific Coast League in 2022, 41 with more than 10 games with a robot. I compare these umpires to those who have never worked in these two leagues.

My primary specification estimates the effects for periods with a robot and without a robot after the implementation:

$$Y_{it} = \delta_0 D_{it}^{Robot} + \delta_1 D_{it}^{PostRobot} + \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it} \tag{1.1}$$

The main outcome Y_{it} is the decision accuracy of the umpire measured as whether a

[25] Table A.2 shows the summary statistics separately for the Minor League umpires who are never-assisted and assisted by the robots. I also show the outcomes for those who are assisted and subsequently move to leagues without . In the year prior to robot implementation, these umpires are not observably different, on average (residualized accuracy, 0.0 vs. 0.0 and 0.0001).

[26] For example, Dane Poncsak called games in the Double-A Northeast in 2021 prior to the implementation. He moved to the Triple-A Pacific Coast League in 2022, a league that implemented, and called games with the robot. In 2023, I observe him in Triple-A in games without a robot.

pitch is correctly called by umpire i in time t.[27] The indicator variables D_{it}^{Robot} specifies that a robot is helping to call the game and $D_{it}^{PostRobot}$ corresponds to the cases where umpires are calling the games without assistance and after having been assisted by a robot. The coefficient δ_0, therefore, measures the relative performance with a robot compared to without a robot. The coefficient δ_1 is the main object of interest and it captures the effect of robot assistance on individual skill after having worked with a robot. I include the year-by-month fixed effects, θ_t, and umpire fixed effects, γ_i, which account for variations in the outcome over time and over umpires. Standard errors are clustered at the umpire level.

The identification assumption of the empirical design is that the the adoption was not anticipated and that the decision to move or to stay in a given league was uncorrelated with the adoption of the robot. In particular, to identify δ_0, the umpires who are assisted by a robot are not on a differential trend compared to the umpires who are not assisted by a robot, conditional on controls and fixed effects. To identify δ_1, conditional on controls, fixed effects, and the indicator D_{it}^{Robot}, the performance of umpires who are no longer assisted by a robot, after having been assisted, is parallel to that of umpires who are not assisted. In other words, the decision to move or stay for the assisted umpires does not depend on robot assistance. To test for these assumptions, I estimate an event-study model to test for pretrends:

$$Y_{it} = \sum_{\substack{k\in\{-\underline{T},\overline{T}\}\\ k\neq -1}} \delta_k D_{it}^k + \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it} \tag{1.2}$$

where the indicator variables D_{it}^k specify that the umpire is k months away from robot implementation. I check that the coefficients, δ_k, prior to the implementation are not statistically significant.

Addressing Threats to Identification A potential threat to identification is that the composition of pitches changed at the time of the implementation of robot assistance. For

[27] See Footnote 20 for how the decision accuracy is measured. I use the post-processed data to determine the accuracy.

instance, players, especially the pitchers, might strategically respond to the implementation of the system changing the nature of the game. These changes in strategies can persist further into the future when players move across leagues. Pitchers could aim at different areas of the strike zone and batters can also be more or less aggressive in the game. These further affect situational characteristics of the game: for example, if the batters are more likely to swing, then the scores of the games can increase. Changes in accuracy of umpires as a result of robot implementation could then be partially due to the changes in the compositions of pitches and game situations.

To partially address these concerns, I include a rich set of control variables and fixed effects ($X_{i,p}$) to flexibly account for pitch location and potential umpire and situational biases. Most importantly, I control for the exact pitch location, which is the only decision guideline that governs whether a pitch should be called a strike or not. Since umpires may want to avoid calling strikes in certain situations (Moskowitz and Wertheim, 2011), I also control for pitch counts and other situational characteristics, including game score, outs, and runners on base.[28] I also control for the top-of-the-inning indicator to account for potential home-team bias and employ pitcher and batter fixed effects to consider any player-specific effects.[29] Further, I conduct a robustness check, reweighing the data to match the distribution of pitches across the leagues and years to address the concern of potential endogenous player responses.

Another potential threat to identification is that the umpires who move across leagues can be systematically different from umpires who do not.[30] For example, if the best umpires in the league are promoted, then comparing the umpires who move following the robot

[28] For example, in the most extreme counts (3-0 and 0-2), umpires are known to adjust the size of the strike zone. Umpires are more likely to call a pitch a strike on a 3-0 count, because the next called ball ends an at-bat. In a lopsided game where the score difference is large, umpires become more lenient in calling pitches to increase the pace of the game.

[29] In principle, the implementation of robot assistance is random, so these are not needed except to increase precision. I test the validity of this assumption by estimating models with and without pitch-level controls and I find similar results (See Section 1.5.1).

[30] Section 1.5.1 shows that the umpires who moved to different leagues or who are promoted to upper-level leagues perform neither better nor worse, on average.

implementation to those who do not can partially capture the differences in these umpires that are not related to the robot and result in an upward bias. On the other hand, if the poor-performing umpires are moving across leagues, then the estimation will result in a downward bias. To alleviate this concern, I also estimate models with varying control groups such as promoted umpires, only the Major League umpires, and only the Single-A umpires.

A third concern is that there may be measurement error in how the pitch coordinates are recorded. As detailed in Section A.4, the data include the X and Y coordinates for every thrown pitch, and there are two different coordinate systems: manually-plotted pitch coordinates and camera-tracked pitch coordinates. In the Major League and the Minor Leagues that adopted the robot, pitch coordinates are available from these cameras. However, in other stadiums without such technology, pitch coordinates are manually-plotted by stringers hired by the league. If the stringers plot the pitches to match the umpire's calls (i.e., plot an ambiguous pitch inside if the umpire calls strike and vice versa), then when the umpire moves to the league with advanced technology, his accuracy might drop as a result of better measurement.

To address this concern, first, I use the dosage model described in the next section that only utilizes games with "Hawk-Eye" technology plotting the pitch coordinates. I also conduct a robustness check using just the manually-plotted pitch coordinates which are available in all stadiums. Second, I employ home team(stadium)-by-year fixed effects, $\pi_{j,y(t)}$ to address how the pitch coordinates are recorded manually. I transform the manually-plotted coordinates to have the same units as the camera-tracked coordinates at the home team-by-year level (Section A.4.3). While I cannot employ stringer-FEs as the data do not contain this information, home team-by-year-FEs partially address the concerns associated with the data quality of pitch coordinates. For example, unique bias in recording coordinates that could depend on the stringers' vantage point can be relieved.

Dosage Model With the 2023 implementation of robots in Triple-A leagues which assigned robots to games held on Tuesdays, Wednesdays, and Thursdays, and the random assignment of umpires to games, the treatment dosage, or the share of games with the robot assisting calling the game is as-good-as random. I therefore also consider the alternative dosage model using just the data from the 2023 Triple-A leagues:

$$Y_{it} = \beta_0 D_{it}^{Robot} + \beta_1 RobotExposure_{it} + \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it} \tag{1.3}$$

where D_{it}^{Robot} is an indicator variable that the game used the robot and $RobotExposure_{it}$ is the share of games with robot assistance.[31,32] The coefficient β_0, estimates how well the umpire performs with robot assistance relative to without. The coefficient β_1 measures the effect of treatment dosage on individual skill. A set of control variables and fixed effects is identical to those employed in Equation 1.1 except that θ_t is now at the week-level.

1.5 Results

1.5.1 Do Humans Lose Skills Following Robot Adoption?

Table 2.2 displays the main results. First, umpires using robot assistance make calls remarkably well. In months when a robot helps call games, pitches are 6.4 percentage points (about 6.9%) more likely to be correctly called. This finding confirms the general belief that robots perform higher quality work than humans in repetitive and monotonous tasks. Robots can help humans achieve a high level of precision and consistency.

Following the implementation of robot assistance, umpires can move to another league where they are asked to perform the task of calling pitches again without assistance. In this

[31] The umpire crew rotates positions throughout the series of games, resulting in a random assignment of umpires to robot assistance. Throughout the season, umpires will work statistically identical number of games with and without robot assistance.

[32] $RobotExposure_{it} = \frac{\text{\# of games with robot}}{\text{\# of total games}}$. For example, $RobotExposure_{it}$ can only take the value 0 or 1 for the umpire's first game and 0, 0.5 or 1 for the umpire's second game, etc.

post-robot period but *without* robots, umpires are 2.0 percentage points (2.1%) less accurate relative to preassistance levels. In other words, the umpires' skill deteriorates following robot implementation below the preassistance levels. The magnitude of the decrease roughly matches the difference between umpires at the median and those in the bottom 5th percentile. The event-study version in Figure 1.3 also suggests the same pattern. Panel A shows that the umpires using robot assistance execute tasks with high precision, but when the robot is removed, they experience a significant skill decline. The figure also confirms that there is no pretrend. The finding suggests that overreliance on robots for tasks that were once performed solely by humans can lead to a loss of skills.

The skills gap between umpires with and without assistance is most pronounced for pitches that fall near the border of the strike zone (See Figure 1.1). Table A.3 restricts the sample to the pitches that are within 0.5 feet from the nearest border of the strike zone from the outside, and within 0.2 feet from the nearest border of the strike zone from the inside. The results are more striking, with a decline in accuracy of 3.7 percentage points (4.1%) following work with robot assistance for these pitches.

With a unique setting that allows me to observe decision accuracy, I can also classify decisions into four different categories: true positive (correctly called strike), true negative (correctly called ball), false positive (Type I error, incorrectly called strike) and false negative (Type II error, incorrectly called ball). As robots improve precision, umpires assisted by a robot have higher true positive and true negative decisions (2.2 and 4.1 percentage points). On the other hand, following the work with robot assistance but without robots, umpires have lower rates of true decisions without robots (1.0 and 1.0 percentage points). Umpires, however, have many more incorrectly called strikes (false positive, 2.5 percentage points) than incorrectly called balls (false negative, -0.5 percentage points) in these months.[33] These collectively result in higher rates of pitches being called strikes; umpires are therefore more lenient to pitchers. Since Figure A.6 shows that umpires tend to call more pitches as strikes

[33] See Figure A.7 for event-study versions of these outcomes.

for pitches that are harder to judge, I take this as suggestive evidence that, when umpires feel more uncertain about their decisions, they lean towards calling strikes.

Humans suffer from decision biases. Table A.4 displays whether incorrect decisions due to biases increase following robot implementation. First, Moskowitz and Wertheim (2011) suggests that umpires suffer from omission bias: they are less likely to call a pitch strike when the count is 0-2 when the next strike would end an at-bat, and more likely to call a pitch strike when the count is 3-0 when the next ball would end an at-bat.[34] I find that the share of wrong calls when the count is 3-0 decreases in the months when the robot helps, calling the game more accurately by 7.7 percentage points relative to the base mean of 8.7%.[35] After robots are removed, the rate of omission biases when the count is 3-0 increases by 2.9 percentage points (33.3%) relative to preassistance levels. Second, Chen et al. (2016) show that umpires suffer from the gambler's fallacy: they underestimate the probability of consecutive streaks happening by random chance. In particular, they are less likely to call a pitch a strike if the previous pitch was called a strike. The robot helps to reduce the share of incorrectly called balls following a previous pitch that was called a strike by 1.2 percentage points. Following the work with robot assistance, but without the robot, umpires have a marginally lower share of wrong calls when the count is 0-2 and when the previous pitch was called a strike. These results, however, should be interpreted with caution, as I also find that umpires lean towards calling strikes on more ambiguous pitches; these results can therefore be an artifact of umpires calling more strikes overall.

With the implementation of robots, coaches and players cannot benefit from questioning an umpire's call of ball or strike. Table 2.2 shows that whether a game has an ejection decreased by 2.4 percentage points, and the number of ejections by 0.029 incidents per game in the season with robots, respectively.[36] Figure A.9 also suggests the same pattern of reduction

[34] The size of the “enforced" strike zone changes based on game situations (Figure A.8).

[35] In a 3-0 count, an incorrectly called strike due to bias is considered an inaccurate call.

[36] In baseball, an “ejection" refers to the act of an umpire removing a player or coach from the game for a rule violation or misconduct. Ejections are typically the result of actions such as arguing with an umpire, using inappropriate language, displaying unsportsmanlike conduct, or violating specific rules.

in ejections. Umpires, however, revert back to the original level of ejections when robots are removed.

Robustness Checks

In this section, I conduct a number of robustness checks to verify the results.

Control Groups My estimation relies on umpires who move across leagues, particularly for the estimation of the skill decline that occurs after robot assistance is no longer available. These umpires might be systematically different in their ability if the leagues promote umpires who are better at the task. However, Table A.5 presents that the umpires who moved across leagues or who are promoted to upper-tier leagues have neither higher nor lower accuracy relative to all other umpires who stayed, on average (residualized accuracy, -0.0 and 0.0 vs. -0.0002). In addition, the identification strategy depends on the presence of parallel trends between the groups, making the differences in levels acceptable.

Table A.6 shows the results with varying control groups and replicates the baseline results. The results are robust to comparing treated umpires to the untreated umpires who are promoted (a decline of 2.7 percentage points) and to untreated umpires who are in the bottom and top leagues (declines of 3.1 and 2.7 percentage points, respectively).

What if the Players Respond Strategically? If a robot increases accuracy and consistency, players may need to adjust their strategies to optimize their productivity. In particular, pitchers can adjust their behavior to focus on hitting specific regions of the strike zone where the umpires assisted by a robot are more likely to call strikes.[37] Knowing that the umpire with robot assistance is more precise in its strike zone, pitchers might aim to hit those areas consistently to increase their chances of getting called strikes.

Table A.7 estimates the same specification, but for players who are exposed to a robot

[37] Pitchers are more than capable of adjusting strategies. See Figure A.10 for how pitchers pitch at different levels.

and move to another league without a robot. Pitchers respond strategically to the use of robot assistance. In the months with robots assisting calls of the game, pitchers pitch closer to the center of the strike zone (0.016 feet), and also closer to the border of the strike zone (0.017 feet), on average. As Figure 1.2 showed, the umpire calls pitches that fall in the corner strikes more often with a robot, making players strategically respond by pitching closer to the border. Figure A.11 shows the event-study versions of these outcomes at the game level. Panel B shows that the pitchers gradually learn and respond to the robot-assisted strike zone by pitching more frequently inside of the strike zone.[38]

The effects persist after these players move to another league without robot assistance. Pitchers maintain the adjusted pitching behavior by pitching closer to the center of the strike zone and to the border (0.008 and 0.003 feet), but Panel B of Figure A.12 show that pitchers readjust by the end of the season and revert back to the original behavior.

The endogenous player response might raise a concern that the baseline results are an artifact of the distribution of pitches changing and persisting. To address this issue, Table A.8 presents the results reweighing the data to have the same distribution of pitches as the prerobot period. The results are largely similar and even more pronounced, with effect sizes of 8.0 percentage points in increased accuracy of robot assistance, and a 2.4 percentage-point decline below preassistance levels following the work with robot assistance.

The main specification includes a rich set of pitch-level controls to account for different game situational characteristics and player behaviors. These are not needed in essence as the implementation of robot assistance is as-good-as random, but used to increase precision. I confirm and find similar results with and without pitch-level controls in Table A.9.

Data Quality If the stringers hired by the league align their pitch plotting with the umpire's calls (i.e., they record a pitch as inside the strike zone when the umpire calls a strike for ambiguous pitches), the umpires' accuracy could potentially decline once they transition

[38] See Section A.5 for a discussion on how strategic adjustment affects players' productivity.

to a league with robot assistance.

Table A.10 shows the results using only the manually-plotted pitch coordinates. The results are much like the baseline results, with slightly attenuated 3.3 percentage-point increase of accuracy with the robot relative to without the robot, and a decline of skill of 1.5 percentage points following the removal of the robot.

Two-way Fixed Effects Finally, I address the concerns raised by recent literature regarding difference-in-differences models (e.g. Goodman-Bacon, 2021). In the presence of heterogeneous treatment effects, a two-way fixed effects model can provide biased results. Figure A.13 presents four different estimators addressing this issue and finds a generally similar pattern of results (Borusyak et al., 2024; Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2021).[39]

1.5.2 Heterogeneity

Do High-Skilled Umpires Lose More Skill? A particularly important question is who loses more with robot implementation.[40] Table A.12 shows the results separately by accuracy level of umpires in the year prior to the implementation. First, in the months when a robot helps calling the game, accuracy increases by 10.5 percentage points for the bottom-quartile umpires and 8.4 percentage points for the top-quartile umpires. In months without robot assistance, the umpires with varying prior skills experience skill declines of varying degrees relative to the preassistance levels. The umpires who had the least skill lose the least, showing a statistically insignificant decline of 0.9 percentage points, while the better umpires lose more, with the middle-half and top-quartile umpires suffering drops of 4.6 and

[39] In these models, I estimate at the month-level instead of the pitch-level.

[40] For all analyses in Section 2.5, I focus on the umpires who have worked in Triple-A or the Major League. Personnel information is more complete and better recorded for this group of umpires. Further, implementation of robot assistance in the Single-A Florida State League happened in 2021, so the preperiod statistics comes from 2019 instead of 2020 due to the COVID-19 pandemic. Table A.11 shows the main results for this sample. For this sample, pitches are 8.9 percentage points more likely to be correctly called with robot assistance and umpires face a 3.6 percentage-point decline in skills relative to preassistance levels.

4.7 percentage points below the preassistance levels, respectively.

Brynjolfsson et al. (2023), Kanazawa et al. (2022), Tafti (2022) and Noy and Zhang (2023) also find that the productivity distribution compresses among taxi drivers, surgeons, and workers after AI, driver assistance, surgical robots and ChatGPT are used, respectively. In all four settings, advanced technology benefited the low-skilled workers. In the setting I examine, I find that the skill decline is the smallest for low-skilled umpires and larger for better-skilled umpires after robot removal. Panel A of Figure A.14 shows that the skills gap among umpires of different skill levels compresses. Prior to robot implementation, the gap between the best and worst-skilled umpires is 6.0 percentage points. When they return to making calls without robot assistance following work with it, the gap shrinks to 2.1 percentage points or by about 64.6%. Collectively, high-skilled workers not only benefit the least *with* robots, but also lose the most when the robot is removed.

Do Less Experienced Umpires Lose More Skill? Table A.13 shows that umpires get neither better nor worse with additional years of experience, on average. Returns to experience are positive and significant for relatively young and new umpires, but they eventually become negative: umpires start losing skills with additional years of experience as they age.[41] Therefore, more experienced umpires are often the less-skilled umpires.

Table A.14 displays the results separately by umpires' years of experience. A robot increases accuracy for umpires of all experience levels in the months when the robot is used. Following the work with robot assistance, umpires with the most experience (i.e., more than 9 years of experience in professional leagues) lose the least, showing a 1.8 percentage points decline relative to the preassistance levels in accuracy. Umpires who are relatively inexperienced (i.e., less than 6 years) experience a significant decline of skill of 5.3 percentage

[41] Williams (2019) finds that the best-performing umpires, on average, have fewer years of experience. Figure A.15 also shows the same patterns.

points.[42,43]

Do Umpires with Prior Major League Experience Lose More Skill? Among Triple-A umpires, those with Major League experience as replacement umpires are more likely to be next in line for promotion.[44] These umpires are also less likely to be replaced sooner than those without Major League experience as the Major League has yet to plan to adopt robot assistance system completely.

Table A.15 shows the results separately by whether the umpire had prior experience in the Major League or not. In the months with robot assistance, decision accuracy increases for both types of umpires (8.2 and 9.3 percentage points, respectively for those without and with Major League experience). However, the magnitude of skill decline relative to the preassistance levels is much larger for those without Major League experience (4.9 percentage points) than those with Major League experience (1.5 percentage points). These findings suggest that those with Major League experience, and potentially who are more likely to be promoted in the near future, try to retain skills more.

1.5.3 Does Longer Work With Robot Increase Skill Depreciation?

The duration of nonemployment spells affects future prospects of wages and reemployment opportunities (e.g. Kroft et al., 2013). While potential explanations include skill depreciation while unemployed (e.g. Benhenda, 2022; Dinerstein et al., 2022; Edin and Gustavsson, 2008; Jarosch, 2023), the length of the nonemployment spell is often correlated with skill, making it difficult to study the impact of duration on skill depreciation. The 2023 Triple-A implementation of robot assistance, paired with the umpire rotation system, makes a good

[42] A large share of umpires who are assisted by robots come from Triple-A minor leagues. For umpires to reach this level, they need several years of training through Single-A and Double-A leagues.

[43] A potential explanation for this result is that the task requires physical fitness (e.g., dynamic visual acuity) to make a decision immediately after observing a fast pitch.

[44] Ahead of the 2023 season, the Major League promoted 10 umpires from the minor leagues. All of these umpires had worked in the Major League as replacement umpires.

setting to study this issue as the length of duration with robots is as-good-as random.

Table A.16 reports the pitch characteristics for Triple-A games in 2023. A similar pattern emerges in these games as the full sample: the robot increases accuracy, on average (94.0% vs. 90.1%). A total of 71 umpires called at least one game in Triple-A leagues, with 64 umpires working with the robot for at least a game. The average umpire's share of games with the robot is 42.0%. Figure A.16 shows the distribution of the dosage. Panel A presents that the share is approximately normal over the season. In the first game of the season, the umpire's treatment dosage is either 0 or 1, and as the season progresses, the dosage becomes more concentrated around 50% (Panel B, Figure A.16). The dosage is as-good-as random, and more or less experienced umpires are neither more nor less likely to have more games with the robot (Figure A.17).

In 2023, umpires working in Triple-A minor leagues had between 1 and 3 consecutive games with robot assistance before having to call a game solely by themselves.[45] To compare with the baseline results of full-season work with robot assistance, I consider the following model:

$$
\begin{aligned}
Y_{it} = \eta_0 D_{it}^{Robot} + \eta_1 RobotExposure_{it} &+ \eta_2 \mathbb{1}(\text{Previous Game Used Robot})_{it} \\
&+ \eta_3 \mathbb{1}(\text{Previous 2 Games Used Robot})_{it} \\
&+ \eta_4 \mathbb{1}(\text{Previous 3 Games Used Robot})_{it} \\
&+ \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it}
\end{aligned}
\tag{1.4}
$$

where the omitted group is that the previous game did not use the robot.[46] I employ the same set of control variables and the fixed effects as previously explained.

[45] On average, the number of days since the last time umpires had to call a game without robot assistance is 8.5, 15.9 and 23.7 days for 1, 2 and 3 consecutive games of robot assistance, respectively.

[46] In particular, $\mathbb{1}(\text{Previous Game Used Robot})_{it}$ is an indicator variable that specifies that the robot was used in the previous game in a previous three-game span. It therefore includes cases where the robot was used in $t-3$ and $t-1$, but not cases where the robot was used in $t-2$ and $t-1$, as this is denoted by $\mathbb{1}(\text{Previous 2 Games Used Robot})_{it}$.

Table 1.3 shows the results separately by the length of work with robot assistance relative to umpires who are not assisted. When an umpire returns to the task after a game with robot assistance, their skill declines by a 0.3 percentage points. The magnitude of the decline of skill increases when an umpire is assisted in 2 or 3 consecutive games by a robot to 1.2 percentage points and a statistically insignificant 1.0 percentage point, respectively.

An average umpire works as a home-plate umpire for 28 games in a season. Unfortunately, the umpires working in Triple-A minor leagues in 2023 had between 1 and 3 consecutive games of work with the robot, so I cannot fully estimate the effect of longer durations of robot assistance. The baseline results suggest that umpires with a full season of work with the robot suffered a 2.0 percentage-point decline in skills relative to the preassistance levels, on average, in the following season; the findings therefore suggest that increases in work duration with robots increase skill depreciation.

Further, Table 1.4 presents the results using an alternative specification (Equation 1.3) exploiting the variation in treatment intensity to robots.[47] I confirm the findings that umpires' skills worsen following the implementation of robot assistance. In games with robot assistance, accuracy increases by 5.2 percentage points. However, full exposure (Dosage = 1) to a robot decreases accuracy by 2.0 percentage points.

Using the variation in dosage, I also plot the second- and third-degree polynomial estimates of dosage effects (Figure 1.4). A game with robot assistance corresponds to dosage of about 3.5% in a full season. At low dosages, skill seems to deteriorate rather linearly, with declines of 0.2, 0.3 and 0.6 percentage points at the dosage of 5%, 10% and 20%, respectively. The 25th and 75th percentiles of treatment dosage are about 36% and 50%. At these dosages, the skill declines are about 1.0 and 1.3 percentages points, respectively. At the dosage of 100%, the decline of skill is 1.9 percentage points. The overall shape of the decline suggests that effects plateau after a certain intensity of exposure.

[47] The results also serve as a robustness check to the main result. The specification only uses camera-tracked pitch coordinates in 2023, so the findings address concerns about data quality.

1.5.4 How Long Does It Take To Regain Skill?

Related questions include whether umpires can regain their skills after loss of robot assistance and, if so, how long it would take to return to their original skill levels. Learning by doing is an important source of productivity growth and returns to experience for workers (e.g. Arrow, 1962; Lucas Jr, 1988; Thompson, 2010; Yang and Borland, 1991). Whether workers return to a prior level of skill with recall after use of robot assistance, and who therefore "unlearned" by doing less, is relatively unknown.[48]

I investigate this question in examining the umpires who are assisted by robots in 2022 and subsequently moved to another league without robots. In particular, I employ the following event-study specification at the game level:

$$Y_{it} = \sum_{\substack{g \in \{-\underline{G}, \overline{G}\} \\ g \neq -1}} \xi_l D_{it}^g + \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it} \tag{1.5}$$

where D_{it}^g indicates that the umpire is g games away from the last game with the robot.

Figure 1.5 shows that these umpires do not recover within a season from the skill decline following the robot assistance. In the first game back after use of robot assistance, the umpires suffer a decline of skill of 3.6 percentage points. While the decline does become smaller in magnitude over the season, it is still statistically significant. By the 16th game, or about halfway through the season, skill deteriorates by 1.7 percentage points. Closer to the end of the season (namely, after 21 games since the return to calling without robot assistance), the skill decline is 1.0 percentage points.

Table 1.5 presents the same results; namely, that the decline in skill is larger early in the season and gradually decreases as the season continues. In the first five games returning to the task without robot assistance, umpires face a skill decline of 2.9 percentage points relative to preassistance levels. However, in the 16-20 games back and 21 or more games

[48] Surprisingly, over 40% of workers who are unemployed return to the original employer (Fujita and Moscarini, 2017).

since the return to calling without robot assistance, the decline is 1.7 percentage points and 1.0 percentage points, respectively. These findings imply that umpires require a considerable amount of time to regain their skills following the work with robot assistance. While they do make significant progress regaining their skill levels over the course of the season, they do not fully return to their initial skill levels by the end of the following season.

1.6 Mechanisms

1.6.1 Are Robot Effects the Same as the Time-Away Effects?

Is skill depreciation a result of time away from the task? Past literature on whether skill depreciates during unemployment spells remains inconclusive as Cohen et al. (2023) find that cognitive and non-cognitive skills do not depreciate for German workers, but Dinerstein et al. (2022) show that Greek teachers become less skilled while unemployed.

In 2020, the COVID-19 pandemic delayed and shortened the 2020 Major League Baseball season, but canceled the entire Minor League Baseball season. Therefore, the Minor League umpires experienced a season-long “unemployment" in which they did not engage in the task, unlike Major League umpires.

I compare the Minor League and Major League umpires in a difference-in-differences framework following the COVID-19-canceled 2020 season.[49,50] In particular, I employ the following specification to study whether the COVID-19 pause resulted in skill depreciation:

$$Y_{it} = \rho_0 D_{it}^{FirstSeason} + \rho_1 D_{it}^{SeasonsAfter} + \beta X_{i,p} + \pi_{j,y(t)} + \theta_t + \gamma_i + \epsilon_{it} \tag{1.6}$$

where the $D_{it}^{FirstSeason}$ and $D_{it}^{SeasonsAfter}$ indicate the 2021 season when the umpires

[49] Table A.17 compares the Minor League umpires who did and did not return from the COVID-19-canceled season. Out of 167 Minor League umpires who worked in 2019 and were never robot assisted, 84 returned in 2021 while 83 did not. In 2019, the year prior to the pause, these umpires are not observably different, on average.

[50] This analysis limits the sample to the never-assisted group.

returned from the pause and all following seasons, respectively.[51]

Table 1.6 reports that in 2021, the first season back from the canceled season, the Minor League umpires do not suffer from any skill depreciation, showing a statistically insignificant decline of 0.4 percentage points, on average.

However, the event-study version reveals that the umpires did marginally lose skill following the pause. Figure 1.6 shows that in the first month since the leagues restarted, the umpires faced a skill decline of 0.8 percentage points, statistically significant at the 10% level. They quickly bounced back, however, with a statistically insignificant decline of 0.5 percentage points starting from the second month of return.

Collectively, these results suggest that the skill depreciation stemming from robot implementation is larger than the time-away effects. A first potential explanation is a reduced incentive to practice skills: umpires may perceive that their skills are no longer relevant, leading to a faster skill depreciation compared to individuals who are temporarily off work. Another potential reason is loss of confidence: when umpires are continually assisted by the robot, they may lose confidence in their abilities to perform tasks manually. This lack of confidence could further erode their skills because they become less willing to engage in tasks that require their expertise.

1.6.2 Complementarity or Substitutability of Skills

Home-plate umpires experience a noticeable drop in their ability to accurately call pitches once they have worked with robot assistance. This raises the question of whether the introduction of robot assistance impacts other skills as well. If two skills are substitutable, a decline in skill in one area can lead to an increase in competency for the other skill. Conversely, when two skills are complementary, a decrease in proficiency in one task can also lead to a decline in skills for the other task.

[51] While the umpires can “practice" before the start of the season in the preseason (i.e., spring training), the same is true for the umpires assisted by a robot for the entire season.

The first-base umpire in baseball is responsible for several tasks during a game and his primary role is to make calls related to plays that occur at first base. The most crucial task of the first-base umpire is to make decisions on whether a baserunner is safe or out when attempting to reach first base. They must closely observe the timing of the runner's arrival at first base and determine whether the ball reached the base before the runner.

A team can request a review of a play that occurred at first base to determine if the call made by the umpire was correct. The replay officials review the available camera angles and video footage of the play and make a determination on whether the call on the field was correct or if it should be overturned. As replay reviews provide a mechanism to correct any errors made by the first base umpire during the game, the numbers of the requested reviews and overturned calls are good proxies for the first-base umpire's skills.

I compare the outcomes of umpires who are assisted by the robots in 2022 to those of who are not.[52] The outcomes are only available for the Major League games, so I restrict the set to all Major League games held in 2023.[53] In particular, I employ the following specification:

$$Y_{it} = \phi_0 D_{it}^{Treated2022} + \beta X_i + \theta_t + \epsilon_{it} \tag{1.7}$$

where the $D_{it}^{Treated2022}$ indicate that the umpire i is treated in 2022. I include team fixed effects to account for variations in the outcome across teams and control for umpire's years of experience. θ_t are the month fixed effects.[54]

Table 1.7 first establishes that there is no significant difference in the number of ground

[52] The umpire crew rotates positions throughout the series of games. For example, an umpire will work behind the home plate in one game and will work behind first base in the next.

[53] Replay reviews are only available in stadiums with cameras installed. Many minor league stadiums still lack this technology.

[54] The identification assumption is that umpires who have not previously received robot assistance serve as good counterfactuals for those who have been assisted. The two groups exhibit some observable differences, as the umpires who were assisted by robots in 2022 are, on average, younger and less experienced. However, in previous sections, I demonstrated that these umpires also perform better on average. As part of a placebo test, I found that the number of ejections does not differ between these two groups which show that the assisted umpires do not make more controversial calls. While the results should be interpreted with caution, I argue that mean comparisons between these groups provide insights into the impact of robot assistance.

outs, or calls for which the first-base umpire is responsible, between the two groups. However, the umpires who are previously assisted by the robots face about 0.1 more challenges (i.e., replay reviews) per game (44.3%), indicating that umpires' decisions are under increased scrutiny. Considering that the number of challenges being overturned has also increased by 0.09 per game (68.6%), these findings collectively imply that umpires who have received previous robot assistance exhibit higher rates of inaccuracies and diminished skills in the task.

As robots and automation systems are integrated into various industries and tasks, the demand for specific skills may shift to adapt to the new working environment. In particular, workers may need to adapt and acquire new skills to perform other tasks. The two tasks, calling pitches and determining whether a baserunner is safe or out, are similar and exhibit complementarities, suggesting that the skill shift following robot implementation may occur at a greater distance along the spectrum.

1.6.3 Are Umpires Learning from the Robot?

When automated systems are used to track pitches and make calls, umpires can receive immediate feedback on the accuracy of their calls compared to the technology. Over time, umpires may adapt their calling style to align more closely with the robot assistance, especially if the technology is proven to be highly accurate.

On July 20, 2021, following inputs from players and umpires, the league widened the strike zone by 2 inches on each side of plate and lowered the top of the strike zone by 3.5 inches (Figure A.2). If the umpires who are assisted by the robot adjust their calling style to match it, then their skills may have seemed to drop as they call with the adjusted rule instead of the old rule. Similarly, if the umpires did match the robot well, using the alternative definition of accuracy with the adjusted strike zone will show that the skills would have increased.

Table A.18 shows that even with the decision accuracy measure using the adjusted strike zone, the umpires experience a decline in accuracy following robot implementation (0.9

percentage points). Therefore, these estimates rule out umpires learning and adjusting to a new standard.

1.7 Discussion

1.7.1 Does the League Gain from Robot Adoption?

While the workers in firms that have adopted robots typically suffer wage and employment losses, robot-adopting firms enjoy increases in output gains and productivity (e.g. Acemoglu et al., 2022). In this section, I explore whether the adopting leagues benefited from implementing the robot.

I compare the Single-A Florida State League that adopted the robot in 2021, and the Single-A Carolina League that did not, in the following difference-in-differences framework:

$$Y_{it} = \tau_0 D_{it}^{FirstSeason} + \tau_1 D_{it}^{SeasonsAfter} + \beta X_i + \theta_t + \omega_i + \epsilon_{it} \tag{1.8}$$

where the $D_{it}^{FirstSeason}$ and $D_{it}^{SeasonsAfter}$ indicate the months in the 2022 and 2023 seasons, respectively.[55] θ_t and ω_i are the year-by-month and the league fixed effects, respectively. I also include team fixed effects to account for variations in the outcome across teams.

Table A.19 shows the results comparing the outcomes in those two leagues. First, league attendance increased by 265.2 per game (12.4%) in the first year of implementation of the robot. A back-of-the-envelope calculation suggests that this translates to an increase of \$1.25 million in just ticket revenue and \$2.5 million considering parking, food and beverages.[56] Figure A.18 shows the event-study version and confirms the finding. Panel A also suggests

[55] The Single-A Florida State League adopted the robot in 2021, and the “First Season" refers to the 2021 season. Other Minor League leagues do not provide good comparisons for several reasons. First, there are large differences in league quality across the league classes. Second, the Minor League has several experimental rule changes each year. For example, in 2021, the Single-A California League adopted onfield timers to reduce game length.

[56] A total of 589 games is played in the Single-A Florida State League in 2021 and an average ticket price in Single-A is about \$8.00. The average cost for a family of four to attend a game is about \$65.00 in total.

that the increase in attendance is concentrated at the beginning and end of the season, which insinuates that increased publicity about the robot might have prompted the increase. In comparison, the cost of “Hawk-Eye" installation is about $300,000 and the robot operator’s hourly wage is about $25.00.[57] The league can more than bear the cost of the robots in its first season of implementation just from increased revenue generated from ticket sales.

Upon the implementation of robot assistance, fans are more likely to enjoy games with fewer controversial calls and disputes, leading to increased viewership and attendance, which can positively impact revenue through ticket sales, merchandise purchases, and advertising revenue. In addition, the use of robot assistance could appeal to tech-savvy audiences and younger generations of viewers, who may be drawn to the game due to its integration with cutting-edge technology. Over time, the adoption of the robots could also lead to cost savings by reducing the number of onfield umpires needed for games.[58]

Second, the duration of the game increased by 21.1 minutes (12.7%) in the first season of implementation and the total score increased by 1.1 runs (12.9%). While the Major League wants to shorten game length, it also wants to generate more in-the-field-of-play offense.[59] Increased duration is likely an artifact of increased scoring and more activity in the game. Indeed, the number of total pitches thrown in a game also increased by 11.2 (4.1%) in the first season. Viewers generally prefer watching high-scoring to low-scoring games and thus, an increased total score in the first season is also likely to enlarge the fan base.

1.7.2 External Validity

The impact of robot assistance on umpires’ skills in making ball-strike decisions has significant relevance and transferability to a wide range of other occupational settings. Beyond the specific context of baseball, the skill set and cognitive processes involved in these decisions

[57] In 2013, the Premier League, an English football league, installed a “Hawk-Eye" system with 14 cameras for £250,000.

[58] Average umpires in Single-A and in the Major League earn $3,000 and $10,000 per month, respectively.

[59] The Major League experimented with various rule changes in the Minor Leagues, including limiting defensive shifts, limiting pickoff throws and moving the mound back to generate more offense.

are shared by professionals across various domains.[60] These skills encompass a combination of vision, physical conditioning, and mental sharpness.

Sharp visual perception and mental acuity required for umpires to make accurate ball-strike decisions are widespread attributes that apply across various professions. They are especially important in any occupation that relies on precise observation and decision making. Healthcare professionals, such as surgeons and radiologists, require sharpness of vision to accurately diagnose medical conditions and perform surgeries. In the aviation industry, pilots and air traffic controllers must possess sharp visual perception to navigate aircraft. Similarly, workers in manufacturing and quality control settings use visual perception to inspect products and ensure quality standards are met.

The cognitive demands on an umpire mirror those in professions that require quick and accurate judgments, especially in high-stakes settings. The cognitive processes involved in decision making, such as attention, are therefore fundamental beyond the setting of baseball. For instance, surgeons and air traffic controllers must maintain high level of focus throughout procedures and navigating air traffic as lapses in attention can lead to potentially catastrophic consequences.

The application of sharp visual perception and acute decision making extends across diverse occupational fields, emphasizing the broader implications and relevance of the findings of this paper. Hence, the deterioration of these skills could have significant consequences for accuracy and safety in numerous occupational domains.

1.8 Conclusion

How does working alongside robots influence human skills, and do individuals perform worse in tasks when robot assistance is removed? Even as robots become more accessible, the value of human skills endures, especially in domains that require adaptability and complex decision

[60] For the first-base umpire, the skill set closely resembles that of the home-plate umpire, as he must make decisions based on the timing of the ball and the runner's arrival at first base.

making.[61]

Importantly, I examine the impact of working with robots on human capital, focusing on the decision accuracy of umpires in professional baseball settings. By analyzing data from over 62,000 games played between 2017 and 2023 in both Major and Minor Baseball Leagues, I find that umpires perform more effectively with robot assistance, affirming the role of robots in enhancing efficiency and productivity across various contexts. However, umpires suffer a large skill decline below preassistance levels when they are required to perform tasks independently once again, suggesting that relying too heavily on robots can have detrimental effects on skill retention.

Second, I investigate the impact of robot implementation on different groups. The umpires with the highest skills see the smallest improvements with robot assistance and experience the most pronounced decline once the assistance is removed. Consequently, the skills gap diminishes with the introduction of robot assistance and persists even after its removal.

Third, I study the impact of the duration of working with the robot on the decline of skill and the time umpires take to recover their skills after the recall following the work with robot assistance. I find that as the duration of robot assistance lengthens, the decline in skill becomes increasingly pronounced, but plateaus over treatment intensity. On the contrary, although umpires make substantial progress, they do not completely return to their original skill level by the conclusion of the season.

Fourth, by studying a canceled season during the COVID-19 pandemic, I reject that skill depreciation can be attributed solely to umpires not utilizing their skills, as skill depreciation resulting from robot implementation is more substantial than the effects of time away. Further, after working with robots, umpires also exhibit a decline in their ability to determine whether a baserunner is safe, suggesting applicability to a broad spectrum of other occupational settings that demand similar skill sets.

[61] In 2018, the automobile manufacturing company, Tesla, failed to meet production targets after relying too much on robots, and replaced the automation system with humans. The company's CEO, Elon Musk, stated that “humans are underrated."

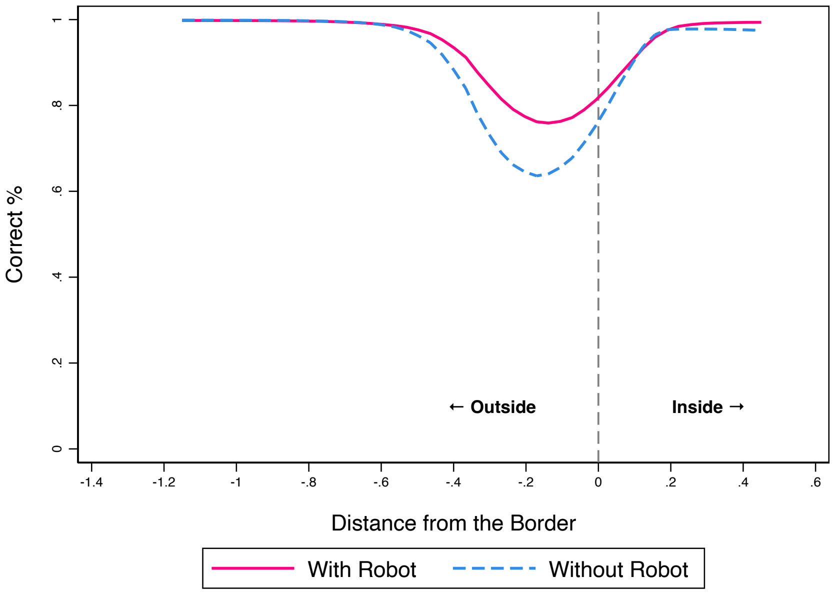
Finally, I find that robot implementation affects other workers who respond strategically and benefits the implementing firm. With robots increasing the accuracy and consistency of calls, pitchers adjusted their strategies to optimize their chances of getting favorable calls. In addition, game attendance and total scores in games increased following robot implementation suggesting an increase in revenue.

These findings raise important questions for future research. Robots and automation can often perform tasks more quickly, accurately, and consistently than humans. This increased efficiency can lead to higher productivity and output gains for businesses (Acemoglu et al., 2020, 2022; Dixon et al., 2021; Humlum, 2022; Koch et al., 2021). However, one of the most significant concerns with automation is the potential displacement of human workers. Jobs that involve repetitive or rule-based tasks are more susceptible to automation and workers in heavily impacted industries can suffer unemployment or wage decline (Acemoglu and Restrepo, 2020; Acemoglu et al., 2020, 2023; Barth et al., 2020; Bonfiglioli et al., 2020; Dauth et al., 2021; Humlum, 2022).

Understanding the impact of automation on human workers' skills when they are assisted by robots is crucial but understudied. First, identifying the skills affected by automation allows policymakers to identify potential skill gaps and develop strategies to help workers adapt to the changing job market. Second, analyzing the skills that are becoming less relevant due to automation can inform targeted reskilling programs to help affected workers transition into new roles. Studying the impact of automation on human workers' skills is essential for proactively addressing the challenges and opportunities brought by technological advancements. More research, therefore, is needed to provide evidence on which skills are affected and how to plan for the future labor market. The finding that robots decrease human skills in tasks they assist is important for future policies. As human capital without robot assistance is critical even when robots are readily accessible, effective labor policies need to address skill decline for a smooth but inevitable transition to the future labor market with robots.

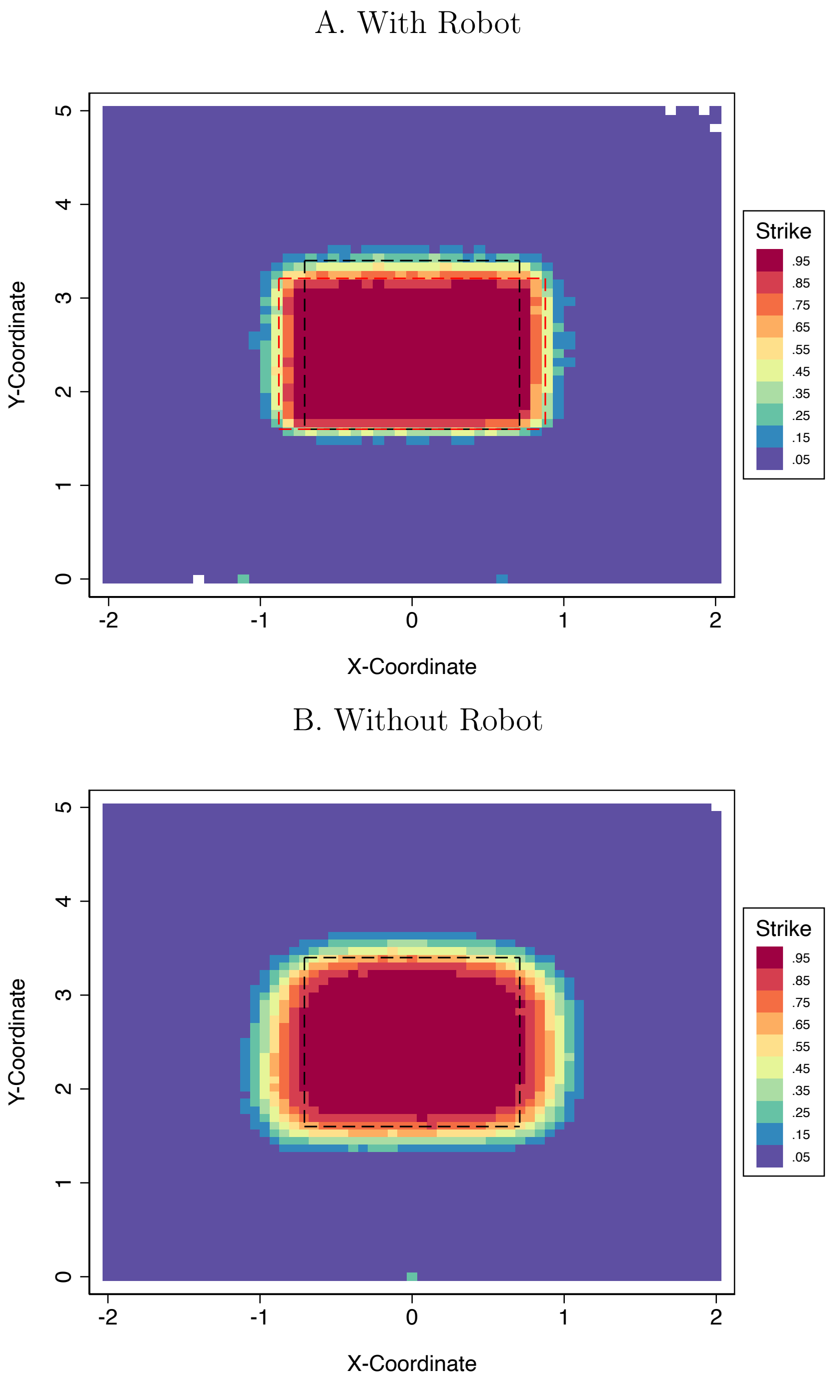
1.9 Figures & Tables

Figure 1.1: Pitch Distribution



Notes: X-axis is the pitch distance from the nearest border of the strike zone in feet and y-axis shows the average accuracy rate. To the left of origin are pitches falling outside of the strike zone, and to the right are pitches falling inside the strike zone.

Figure 1.2: Called Strike Heatmaps

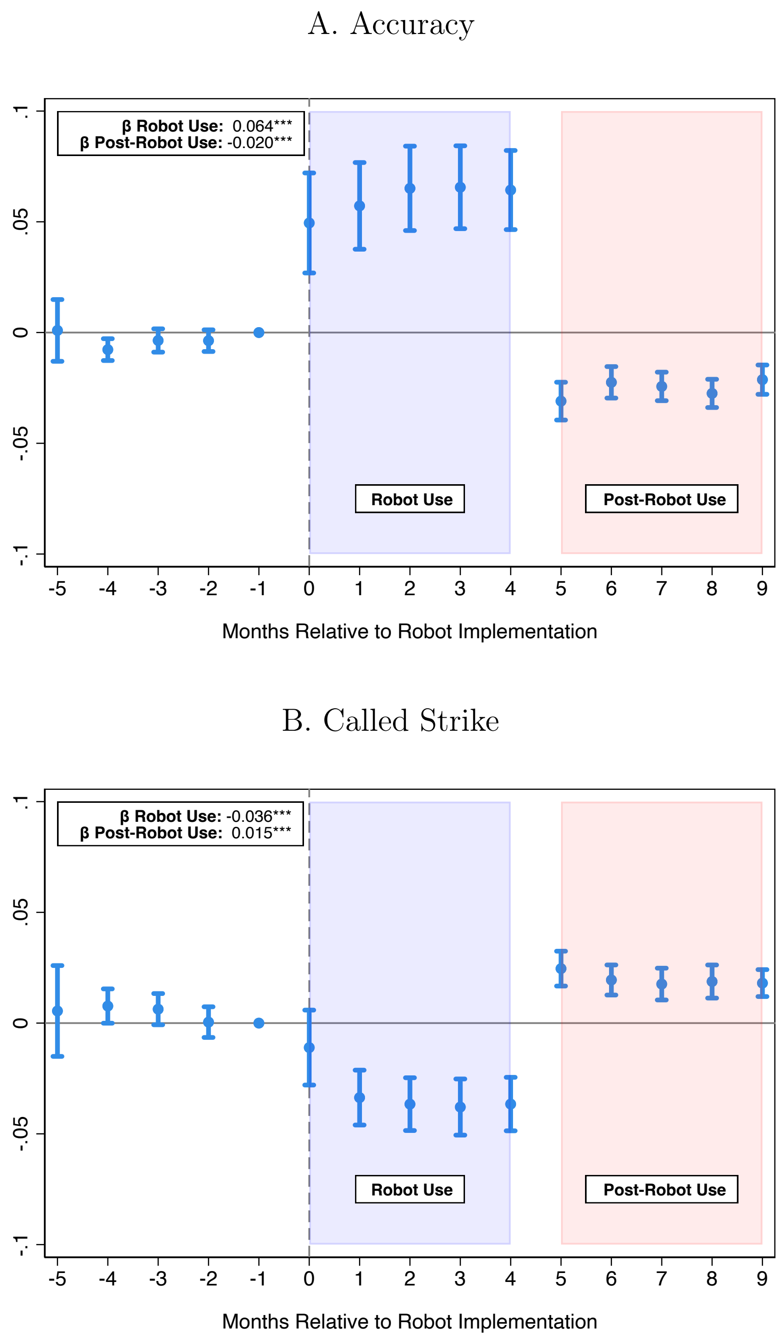


A. With Robot

B. Without Robot

Notes: The figures plot the share of pitches that are called strike by pitch location. The black dotted line shows the strike zone for an average batter and the red dotted line in Panel A shows the adjusted strike zone that was implemented on July 20, 2021 for games with robot.

Figure 1.3: Do Umpires Lose Skill? - Event-Study

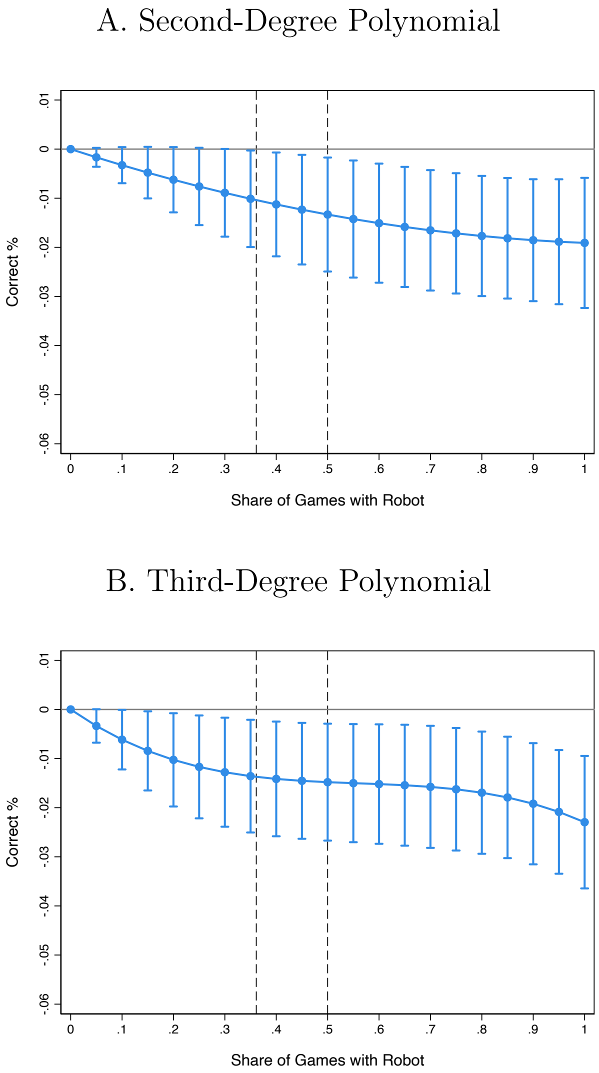


A. Accuracy

B. Called Strike

Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Use" indicate that robot is assisting umpires calling the game and "Post-Robot Use" indicate that the umpire returned following robot-assistance. X-axis is months relative to the first month of robot implementation. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. * p<0.1,** p<0.05, *** p<0.01.

Figure 1.4: Do Skill Depreciation Vary with Dosage?

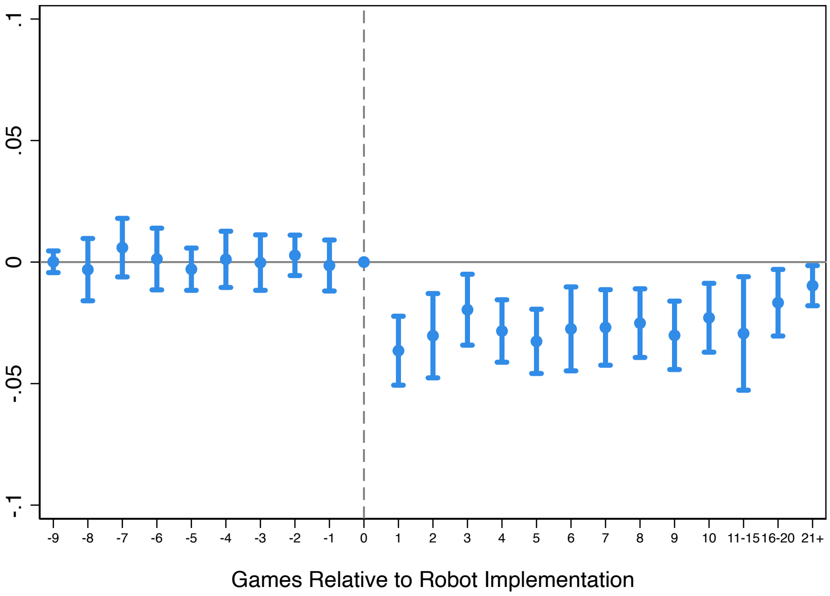


A. Second-Degree Polynomial

B. Third-Degree Polynomial

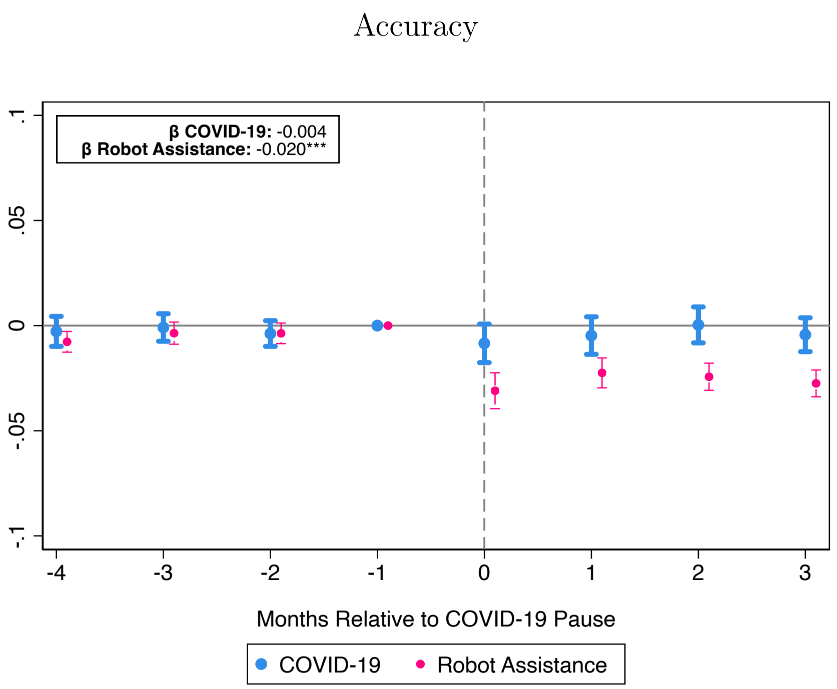
Notes: X-axis plots the dosage where it is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$. Y-axis plots the treatment effect on decision accuracy. The figure plots the second and third-degree polynomial estimates of dosage effect. The regression also includes a vector of covariates at the pitch-level, week, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. The gray dotted lines represent the 25th and 75th percentile of dosage. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball.

Figure 1.5: Do Umpires Regain Skill?



Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. X-axis is the number of games relative to the last game with the robot. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball.

Figure 1.6: Event-Study - Umpire Skills Following COVID-19



Notes: All regressions include year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. X-axis is months relative to the first month of 2021 season following the COVID-19 pause. Blue lines show the effect of returning from COVID-19 pause and red line shows the effect of robot implementation in comparison. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. * p<0.1,** p<0.05, *** p<0.01.

Table 1.1: Summary Statistics

	Full Sample		With Robot		Without Robot	
	Mean	SD	Mean	SD	Mean	SD
Pitch Characteristics						
Called Strike	0.335	(0.472)	0.312	(0.463)	0.336	(0.472)
Called Correctly	0.930	(0.255)	0.930	(0.255)	0.930	(0.255)
Residualized Accuracy	-0.0000	(0.2392)	0.0152	(0.2461)	-0.0007	(0.2388)
Horizontal distance	0.753	(0.453)	0.782	(0.495)	0.752	(0.451)
Vertical distance	0.850	(0.567)	0.869	(0.584)	0.850	(0.566)
Game Characteristics						
Ejection by Umpire	0.066	(0.302)	0.042	(0.243)	0.067	(0.304)
Number of Games	62,678		2,611		60,067	
Number of Pitches	8,864,801		390,823		8,473,978	
Number of Umpires	678		121		666	

Notes: A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. "Residualized accuracy" residualizes whether a decision was correct for pitch location and team-by-year fixed effect to account for pitch coordinates that may depend on calibration for each stadium and stringer plotting coordinates. Distances are in feet. Robot is used in games in Single-A Florida from 2021, in Triple-A Pacific Coast League in 2022 and in select games in both Triple-A leagues in 2023.

Table 1.2: Do Umpires Lose Skills Following Robot Implementation?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.064***	(0.009)	-0.020***	(0.003)	0.932	8,864,667
Correctly Called Strike	0.022***	(0.005)	-0.010***	(0.002)	0.281	8,864,667
Correctly Called Ball	0.041***	(0.006)	-0.010***	(0.002)	0.651	8,864,667
Incorrectly Called Strike	-0.058***	(0.009)	0.025***	(0.003)	0.059	8,864,667
Incorrectly Called Ball	-0.006***	(0.001)	-0.005***	(0.001)	0.008	8,864,667
Called Strike	-0.036***	(0.006)	0.015***	(0.003)	0.340	8,864,667
Game-level Outcomes						
$\mathbb{1}(Ejection)$	-0.024***	(0.007)	0.001	(0.006)	0.054	62,539
Number of Ejection	-0.029***	(0.010)	0.003	(0.008)	0.066	62,539

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and home team-by-year fixed effects. Game-level outcomes exclude pitch-level covariates. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table 1.3: Do Skill Depreciation Vary with Length of Work Duration With Robot?

	Treated Umpires									
			Num. of Consecutive Games w. Robot							
	With Robot	S.E.	1 Game	S.E.	2 Games	S.E.	3 Games	S.E.	Baseline Mean	N
Pitch-level Outcomes										
Correctly Called	0.049***	(0.001)	-0.002	(0.001)	-0.012***	(0.003)	-0.010	(0.008)	0.916	300,427
Correctly Called Strike	0.031***	(0.002)	-0.002	(0.002)	-0.010**	(0.004)	-0.006	(0.008)	0.244	300,427
Correctly Called Ball	0.018***	(0.002)	-0.000	(0.002)	-0.001	(0.004)	-0.003	(0.008)	0.673	300,427
Incorrectly Called Strike	-0.061***	(0.001)	0.001	(0.001)	0.008**	(0.003)	0.011	(0.008)	0.076	300,427
Incorrectly Called Ball	0.012***	(0.001)	0.001	(0.001)	0.004***	(0.001)	-0.002	(0.002)	0.007	300,427
Called Strike	-0.030***	(0.002)	-0.001	(0.002)	-0.003	(0.005)	0.005	(0.008)	0.320	300,427

Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. “With Robot” indicate that the robot is assisting umpires calling the game. “Num. of Consecutive Games w. Robot” indicate that the umpire was assisted by the robot for the last specified number of games. Omitted group is the previous game that did not have the robot. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the first game of the year without robot-assistance. * p<0.1,** p<0.05, *** p<0.01.

Table 1.4: Do Skill Depreciation Vary with Treatment Intensity?

	Treated Umpires					
	With Robot	S.E.	Dosage	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.052***	(0.001)	-0.020***	(0.007)	0.885	330,185
Correctly Called Strike	0.034***	(0.001)	-0.012	(0.007)	0.211	330,185
Correctly Called Ball	0.018***	(0.001)	-0.008	(0.008)	0.674	330,185
Incorrectly Called Strike	-0.064***	(0.001)	0.018***	(0.006)	0.106	330,185
Incorrectly Called Ball	0.012***	(0.000)	0.002	(0.003)	0.009	330,185
Called Strike	-0.030***	(0.001)	0.006	(0.008)	0.318	330,185

Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "With Robot" indicate that robot is assisting umpires calling the game and "Dosage" is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the first game of the year without robot-assistance. * p<0.1,** p<0.05, *** p<0.01.

Table 1.5: Do Umpires Regain Skill?

	Num. of Games Since Return											
	1-5 Games	S.E.	6-10 Games	S.E.	11-15 Games	S.E.	16-20 Games	S.E.	21+ Games	S.E.	Baseline Mean	N
Pitch-level Outcomes												
Correctly Called	-0.029***	(0.004)	-0.027***	(0.005)	-0.029**	(0.012)	-0.017**	(0.007)	-0.010**	(0.004)	0.929	4,800,807
Correctly Called Strike	-0.013***	(0.004)	-0.007*	(0.004)	-0.006	(0.007)	-0.011	(0.008)	-0.005*	(0.003)	0.275	4,800,807
Correctly Called Ball	-0.016***	(0.006)	-0.019***	(0.005)	-0.023**	(0.010)	-0.006	(0.005)	-0.004	(0.004)	0.654	4,800,807
Incorrectly Called Strike	0.031***	(0.004)	0.026***	(0.004)	0.006	(0.005)	0.017***	(0.007)	0.010***	(0.004)	0.063	4,800,807
Incorrectly Called Ball	-0.001	(0.002)	0.001	(0.001)	0.023*	(0.013)	-0.000	(0.003)	-0.001	(0.001)	0.008	4,800,807
Called Strike	0.018***	(0.005)	0.019***	(0.005)	0.000	(0.008)	0.007	(0.005)	0.005	(0.003)	0.338	4,800,807

Notes: All regressions for pitch-level outcomes include year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. “Num. of Games Since Return’ indicate that the umpire is calling the game in the specified numbers of game since the return. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table 1.6: Do Umpires Lose Skill? - From COVID-19 Pause

	Treated Umpires			
	First Season	S.E.	Baseline Mean	N
Pitch-level Outcomes				
Correctly Called	-0.004	(0.003)	0.907	2,817,021
Correctly Called Strike	0.001	(0.003)	0.261	2,817,021
Correctly Called Ball	-0.005**	(0.002)	0.647	2,817,021
Incorrectly Called Strike	0.003	(0.003)	0.082	2,817,021
Incorrectly Called Ball	0.000	(0.001)	0.010	2,817,021
Called Strike	0.005**	(0.002)	0.343	2,817,021

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "First Season" indicates that umpires returned from a year-long COVID-19 interruption relative to Major League umpires who didn't have such canceled season. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. * p<0.1,** p<0.05, *** p<0.01.

Table 1.7: Do Umpires Lose Other Skill? - First Base Umpire and Replay Review

	Post-Robot Umpire	S.E.	Baseline Mean	N
First Base Umpires				
Num. of Ground Outs to 1B	-0.601	(0.399)	19.769	2,403
Num. of Challenges for Calls on the 1B	0.098**	(0.041)	0.221	2,403
Num. of Challenges Overturned	0.094***	(0.028)	0.137	2,403
Num. of Challenges Upheld	0.005	(0.029)	0.084	2,403
Num. of Ejections	-0.013	(0.013)	0.060	2,403

Notes: All regressions include a control for umpire experience and month and team fixed effects. Standard errors are clustered at the umpire-level. "Post-Robot Umpire" indicate that the umpires are assisted by the robots in 2022. The outcomes are from the Major League in 2023 season. "Num. of Ground Outs to 1B" include all ground outs with a play at the first base. "Num. of Challenges for Calls on the 1B" are the number of replay reviews requested by the teams for plays happening at the first base. "Num. of Challenges Overturned" and "Num. of Challenges Upheld" are the number of replay reviews that are overturned and upheld for plays happening at the first base, respectively. "Num. of Ejections" are the number of ejections in a game. * p<0.1,** p<0.05, *** p<0.01.

Chapter 2

The Impact of Fear on Police Behavior and Public Safety

(with Felipe Gonçalves and Emily Weisburst)

2.1 Introduction

The job of a police officer is dangerous, with a fatality rate that ranks among the top twenty across professions in the United States.[1] It is also high stakes – law enforcement actions have the capacity to improve public safety but may also impose large economic, social, and human consequences for sanctioned individuals and their broader social networks.[2] An open question is whether an emotional response to perceived on-the-job risk could reduce the social optimality of officer decisions. This issue is particularly important in the U.S., where officers make over 10 million arrests each year, most of which are for lower-level misdemeanor offenses, and rates of police use of force, incarceration, and crime are high relative to other countries.[3]

[1] Stebbins, Samuel, Evan Comen and Charles Stockdale. 1/9/2018. "Workplace fatlities: 25 most dangerous jobs in America." *USA Today.* https://www.usatoday.com/story/money/careers/2018/01/09/workplace-fatalities-25-most-dangerous-jobs-america/1002500001/

[2] See Bacher-Hicks and de la Campa (2020a,b); Gonçalves and Mello (2023); Mello (2018); Weisburst (2024) for examples.

[3] In 2019, annual arrests in the U.S. exceeded 10 million, with less than 20% of arrests corresponding to serious felony offenses (FBI UCR, https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-29). In 2021, the U.S. ranked 6th among all countries in the share of population incar-

Research at the intersection of economics and neuroscience suggests that fear can affect both the interpretation of and response to risk (See Camerer et al. 2005 for a review), and economists have found that heightened emotional states can have real-world impacts, often within short time horizons.[4] Moreover, a broad literature has identified non-pecuniary drivers of workplace behavior, including social connectedness (Bandiera et al., 2010), interpersonal comparisons (Ager et al., 2022), and grief from personal loss (Graddy and Lieberman, 2018). Because of officers' significant enforcement discretion, changes in their emotional mindset have the potential to affect decisions such as when to make an arrest, which could have material consequences for public safety.

This paper asks two questions: First, how do changes in the salience of fatality risk impact officer behavior? Second, do these changes have downstream consequences for public safety? We examine cases of police officer deaths in the line-of-duty. We show that, after the death of an officer, fellow officers markedly reduce their arrest activity. By making fewer arrests, officers remove themselves from potentially risky interactions with suspected offenders, consistent with risk mitigation due to heightened fear. Despite this decline in arrests, crime does not increase on average in the ensuing period, nor do we find crime increases in cities with larger or longer arrest declines. This lack of a crime effect may be attributable to the types of arrest reductions we observe, which are most concentrated among low-level offenses and are potentially less instrumental to improving public safety. These impacts suggest that, while shocks to perceived fatality risk can lead to substantial enforcement responses, officer fear does not ultimately contribute to higher rates of crime.

A preoccupation with fatality risk is central to police culture, and officers often view their work in "life-or-death" terms. They are formally instructed on the potential perils of their

cerated (World Prison Population Brief, https://www.prisonstudies.org/highest-to-lowest/prison_population_rate). Cross-country comparisons of crime show that the U.S. homicide rates rank 40th among 229 countries (World Bank, https://data.worldbank.org/indicator/VC.IHR.PSRC.P5). See Hirschfield (2023) for discussion of American police killing rates in a cross-country perspective.

[4] For applications in finance, see Cohn et al. (2015); Duxbury et al. (2020); Lo et al. (2005). For emotional responses to football game losses, see Card and Dahl (2011) and Eren and Mocan (2018). Further, exposure to violent crime and war have been found to alter risk preferences (Brown et al., 2019; Voors et al., 2012).

work and on self-protection in the field, beginning with their police academy training. When an officer dies while on duty, their department typically honors them with a formal police funeral, including dress uniforms, dedicated music, a 21-gun salute, and a symbolic "end of watch call" to the fallen officer.[5] A majority of officers (84%) cite that they worry about their safety on the job, and officers feel that the public does not understand the risks and dangers inherent in their occupation, or the challenges of policing more broadly.[6]

Theoretically, it is unclear how on-the-job fear will affect enforcement and public safety, and previous studies focusing on individual cities have found varied impacts. Peer officer injuries may increase arrests and use of force (Holz et al., 2023), while evidence from peer deaths has shown that officers subsequently reduce their enforcement activity (Chalfin et al., 2021b; Sloan, 2019; Sullivan and O'Keeffe, 2017). The criminal environment could change if officers adjust their arrest activity and their altered enforcement is important for incapacitating or deterring crime. Conversely, crime may not change if any marginal changes in enforcement are not central to maintaining public safety.

We examine a sample of 1500 municipalities between 2000 and 2018, and our empirical strategy uses a difference-in-differences design that exploits the staggered occurrence of line-of-duty deaths across agencies.[7] We first document that a line-of-duty death is followed by a significant 10% decline in police arrest activity over one to two months. This effect is present for arrests of all offense types, including serious violent and property crime. While the percentage change across all categories is similar, the reduction in number of arrests is substantially greater for lower-level offenses. Using a series of event-study specifications, we

[5] Ethnographic research also highlights that officer deaths become a part of a department's "organizational memory" long after the deaths occur (Henry, 2004; Marenin, 2016; Sierra-Arévalo, 2019, 2021).

[6] Pew Research (2017) "Behind the Badge": https://assets.pewresearch.org/wp-content/uploads/sites/3/2017/01/06171402/Police-Report_FINAL_web.pdf

[7] A growing literature in economics examines the impact of unexpected deaths of individuals, which have been used to identify productivity spillovers (Azoulay et al., 2010; Jaravel et al., 2018), labor market frictions (Jäger and Heining, 2022), and the impact of leaders (Bennedsen et al., 2020; Jones and Olken, 2005). In contrast to this previous work, the death events we study occur at work, leading to a shift in fellow employees' perception of their workplace safety. In addition, we provide evidence that our effects are driven by an emotional response rather than productivity spillovers or direct incapacitation from the deceased individual, most similar to Graddy and Lieberman (2018) on artist productivity after the death of a family member.

confirm that these events are not preceded by significant changes in crime or arrest activity, suggesting that their timing is exogenous to the criminal environment.

To further characterize the behavioral arrest response, we ask how our estimated effects vary by city characteristics. We use a synthetic control approach to estimate event-specific treatment effects, where each city's arrest and crime rates are compared to a weighted average of outcomes for cities without an officer death over the same period. We find the biggest arrest declines in smaller cities and those with fewer crimes per capita, consistent with a peer death being a greater shock to officers operating in a less-active criminal environment.

We provide several pieces of evidence that the observed arrest reductions are due to behavioral changes in enforcement rather than changes to police manpower. First, using conservative assumptions, our estimated effect size is too large to be driven by officers taking time off following the death of their peer. Second, arrest declines occur even in officer fatality cases where the offender is apprehended quickly, suggesting that the effect is not due to peer officers being diverted to investigate the incident. Third, no arrest decline occurs after accidental officer deaths (e.g. car accidents), which should have a similar incapacitation effect due to the deceased officer and any bereavement leave of peers. This result also rules out productivity spillovers from the deceased officer as an important channel. Finally, using police employment data from Florida, we find that a peer death does not result in officer quits or hires. This null employment-level response from officers is consistent with the literature on how workplace risk is priced into wages, which yields the smallest wage-risk gradients among occupations with higher baseline levels of fatality risk (Viscusi and Aldy, 2003).

We next turn to how officer deaths impact public safety. In contrast to the observed arrest decline, we find small and statistically insignificant impacts on reported crimes. Our 95% confidence intervals rule out short-term (long-term) increases of greater than 3.6% (2.6%) in felony "index" crimes, the most serious violent and property crimes defined by the Federal Bureau of Investigation (FBI).[8] Our results are robust to a variety of specification choices, and

[8] Index crimes include murder, rape, aggravated assault, robbery, burglary, theft, and motor vehicle theft. We consider murder separately from other violent crime to account for changes in this outcome related to the

they suggest that public safety is not worsened by heightened officer fear. Despite significant variation across cities, we find no evidence of crime increases in cities whose characteristics predict a bigger arrest effect, again leveraging our synthetic control estimates. To further probe the lack of crime impact, we directly stratify cities by their estimated magnitude and length of arrest decline. We fail to find evidence of a threshold arrest decline magnitude or duration above which crime increases, even when examining arrest reductions that are larger than 30% or persist for five or more months.

A key challenge with studying how changes in policing impact crime is that measured crime is partly a function of police reporting. If officers respond to a peer death by reducing their propensity to record crimes, this could bias us away from finding an increase in crime (Levitt, 1998; Mosher et al., 2010). To address this concern, we hand-collected a large data set of 911 calls from 56 police departments. These calls originate with civilians and therefore are not directly affected by changes in officer reporting behavior.[9] We find that 911 call volume does not significantly change after an officer death, and the propensity of officers to write a crime report conditional on a call does not decrease after a peer death.

Our study contributes to a growing literature on how police enforcement responds to sudden shocks to the salience of workplace risk and the downstream impact on crime. Previous studies all focus on single cities and either observe a large sample of lower-level incidents (Holz et al., 2023; Sloan, 2019) or a single prominent incident (Chalfin et al., 2021a; Sullivan and O'Keeffe, 2017). Relative to this important work, we emphasize two central contributions of our study. First, we examine a set of almost two-hundred officer deaths that occur across a wide range of police departments and include a rich set of outcomes which track police behavior and crime in multiple dimensions. Notably, we observe many incidents that occur in medium and small-sized agencies, which are unstudied in this context. This breadth

officer death itself (see Section 2.4).

[9] Note that Ang et al. (2024) find a reduction in civilian willingness to contact the police after the George Floyd murder, so police behavior can impact civilian reporting practices. The specific issue we address here is that we avoid changes in crimes due to officer reporting propensity *conditional on a civilian call.*

of settings allows us to examine how officer responses vary with department and incident characteristics. It also allows us to show a robust null crime impact, regardless of the decline in enforcement. Second, we improve on the measurement of crime outcomes by linking the line-of-duty deaths to 911 calls, which addresses long-standing concerns that changes in enforcement behavior may coincide with changes in reporting practices.

More broadly, our study relates to the labor economics literature on workplace safety, to which we make two important contributions. First, while many studies have examined variation across occupations and firms in measures of fatality risk to identify a wage-risk gradient (see Viscusi and Aldy 2003 for a review), we innovate on this approach by using variation within a given workplace in the *salience* of workplace risk. Second, the vast majority of the literature takes injury or fatality risk as an exogenous feature of a job rather than a characteristic that is partly determined by worker behavior.[10] We contribute to this understudied area by documenting a direct behavioral response to workplace risk and how it affects the relevant total productivity measure of public safety.

Our primary focus is on identifying the impacts of an officer death. However, if police officers are the only individuals who directly respond to these events and their response is solely manifested by a reduction in arrests, then our results can be interpreted to indicate the impact of a marginal change in arrests on crime.[11] In particular, the declines in arrests we observe are driven by a reduction in the enforcement in low-level offenses, which represent the majority of total arrests, but may have limited public safety value. Viewed through this lens, our work contributes to an important open question in the economics of crime of how

[10] Two notable exceptions are Guardado and Ziebarth (2019), who use employee weight (and how it varies with compensation) as an indirect measure of investment in safety, and Kohlhepp and McDonough (2022), who studies the overtime decisions of traffic officers and its impact on injury risk.

[11] A number of papers have studied other institutional changes that affect officer enforcement behavior, and this work finds mixed effects of these changes on crime (Mas, 2006; McCrary, 2007; Owens et al., 2018). We likewise complement a growing literature on the impact of heightened public scrutiny on police behavior (Ba, 2020; Heaton, 2010; Prendergast, 2001, 2021; Rivera and Ba, 2019; Shi, 2009), which finds that following a scandal, officers often reduce discretionary enforcement and crime increases (Cheng and Long, 2018; Devi and Fryer Jr, 2020; Premkumar, 2020), but that victim and community trust in police may also decline (Ang et al., 2024).

changes in police *enforcement* impact crime (Bacher-Hicks and de la Campa, 2020a; Chalfin et al., 2021b; Chandrasekher, 2016). While suggestive, these estimates point to the potential for reforms which reduce the scope of arrest activity without the cost of elevated crime rates.

2.2 Data

This study combines data from several sources. Our sample includes 1,578 municipal police departments that report at least 9 years of continuous monthly data between 2000-2018 to the Federal Bureau of Investigation (FBI) Uniform Crime Report (UCR) program.[12]

A total of 135 officer death events occur within 82 police departments during our sample period. A detailed depiction of the sample construction and sample restrictions is included in Figure B.1, and additional description of data sources can be found in Appendix B.5.

Information on officer deaths is derived from the Law Enforcement Officers Killed or Assaulted (LEOKA) series of the Federal Bureau of Investigation (FBI) Uniform Crime Report (UCR). The analysis considers only officer deaths that result from felonious killings and excludes deaths resulting from accidents. This data is linked to information collected on officer deaths by the Officer Down Memorial Page website to determine cause of death.[13]

The arrest and crime data at the month by department level is also sourced from the FBI UCR data on crime reports and arrests. These national data are self-reported to the FBI by individual police departments with limited auditing and therefore have notable data quality issues. To address concerns about reporting accuracy and quality, we first restrict to the agencies who report complete and continuous data on *both* arrests and crimes at the monthly level. Our sample period is 2000-2018. We include agencies whose records span at least nine consecutive years and include the latest year of data, 2018, meaning that each agency's panel

[12] Agencies in our analysis sample have an average panel length of 18.7 years, meaning that very few departments in the sample have fewer than the maximum 19 years of data.

[13] We exclude 16 officer fatalities coded in the LEOKA data that could not be verified by either the Officer Down Memorial Page or an external source.

starts between 2000 and 2010.[14] Our sample restriction differs from prior work that typically relies on *annual* data reporting or the population of municipalities.

Our crime and community activity outcomes also include records of 911 calls for 56 cities in our sample. We have hand-collected these records through open records requests to police departments across the U.S., as this data is not available in any systematic or aggregated form at the national level. To our knowledge, this collection represents the largest sample of 911 calls that has been used in a quantitative research study to date. This data covers the period of 2005-2018, though the number of years varies by city. These data largely originate from departments' "computer-aided dispatch" systems for routing officers to calls, and in some cities the data include cases that are officer-initiated, such as a dispatch call to assist another officer. We remove calls whose descriptions are indicative of an officer-initiated interaction, and we construct an agency-by-month count of civilian-initiated calls.[15]

We also incorporate data on traffic stops collected by the Stanford Open Policing Project through open records requests. As a complement, we measure traffic fatalities in each city in our sample using data from the Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA).

Lastly, we include data on yearly demographic characteristics of cities from the U.S. Census and the American Community Survey. These variables allow us to control for changing demographic composition in the cities covered by our analysis sample (see Section 2.3).

[14] We also clean the data to exclude a minority of observations where a police department lists crime or arrests as having a negative value. These negative values are very rare in practice. These missing values mean that the number of observations may differ slightly by crime or arrest outcome in our models. Negative numbers can be used to correct earlier reports of arrests or crimes that were misreported by an agency; however, they are not linked to a particular misreported month, so they cannot be used to update the crime or arrest data manually.

[15] A previous version of this study included a section with a case study of a single officer fatality in Dallas, TX. These analyses were based on public records requests made to the Dallas Police Department. We requested the same data for the time period around the fatality multiple times, and upon further inspection, we found that our results varied significantly when using different versions of the records provided by the department. We have therefore decided to remove this section from the study.

Summary Statistics – Approximately 7 officer deaths occur in each year within our sample of 1,578 police departments, though there is variation in the number of deaths across years.[16] Figure B.2 shows that there may be some seasonality in this outcome throughout the year, with the highest number of deaths observed in the winter and summer months. Over 90% of the officer deaths in our sample result from gunshot wounds (Table 2.1). Similar to the national statistics, officers who are killed in the sample are demographically representative; the average officer death is of a 37 year old white male with 11 years of experience.

Table B.1 summarizes demographic characteristics of the sample at the yearly level. The average city has 41 thousand residents, is 68% white, has a 13% poverty rate, and a median household income of $46 thousand dollars. In contrast, treated law enforcement agencies serve populations that are larger, more racially diverse, and more likely to live in poverty; on average, these cities have 240 thousand residents, are 54% white, and have a 16% poverty rate. Treated cities are defined by having an officer death; in turn, these departments also experience more officer assault injuries each year (75 vs. 11 in the full sample).

Our estimation focuses on arrest and crime outcomes at the department by month level. Table 2.1 shows that the average department in our sample reports 0.2 murders, 18 other violent crimes and 122 property crimes per month. The average department makes 152 arrests per month, of which 83 are for "quality of life" or low-level offenses, 0.17 are for murder, 8 are for other violent crimes, and 20 are for property crimes.[17] For the sub-sample of agencies that have traffic stop and traffic fatality data, the average department makes over 6,200 monthly traffic stops and experiences 0.26 monthly fatal traffic accidents. Consistent with the fact that treated agencies serve much larger cities, treated agencies also have substantially higher levels of reported crime and make more arrests and traffic stops than the average department

[16] As noted above, the national total is approximately 60 deaths per year. Our sample is restricted to cities that regularly report monthly FBI crime data, and cover a sub-set of the country. See the Data Appendix for additional details on sample construction.

[17] In this paper, we exclude murder arrests and murder crimes from index violent crime or arrest sums and measure these outcomes separately. We do this to easily see the effects on murder (which is related to the officer death treatment) separately from other violent crimes.

in the sample.

Given the clear baseline differences between our treatment and control agencies, we employ a difference-in-differences model which includes detailed controls and department-specific fixed effects, as we discuss in Section 2.3. Our findings are robust to restricting the sample to include only treated agencies and solely exploiting variation in the timing of officer deaths, which provides reassurance that the baseline differences across the treatment and control agencies do not bias the results (see Table B.2, specification (2)).

To provide a simple presentation of the time path of crime and arrests and our empirical strategy, Figure 2.1 plots the raw data around officer fatality events, comparing average logged outcomes in the treated year to the year prior for treated agencies. While these plots are not adjusted for any covariates or fixed effects, they accord with the overall pattern of findings in the study.[18] Panel A of Figure 2.1 shows that total arrests decline in the month of an officer death and month after, with a drop of ≈ 0.1 log points or 10% in the first month. Despite this drop in total arrests, Panels B does not appear to show a temporary or systematic increase in serious felony or index crimes.

2.3 Empirical Strategy

Our empirical strategy exploits the staggered occurrence of officer deaths over time in a difference-in-differences framework. A baseline regression will allow for effects to vary by the time horizon from the date of the incident:

$$\begin{aligned} Y_{it} =& \delta_0 D_{it}^{0} + \delta_1 D_{it}^{1} + \delta_{2-11} D_{it}^{2-11} + \delta_{12+} D_{it}^{12+} \\ &+ \beta X_{i,yr(t)} + \pi_{i,m(t)} + \theta_t + \gamma_i t + \epsilon_{it} \end{aligned} \tag{2.1}$$

In our primary specifications, we define our outcomes as $Y_{it} = log(y_{it} + 1)$ to approximate percentage changes and account for zero values for each outcome category, y_{it}; however, we

[18] The log transformation used is $ln(y + 1)$ to permit zeros in the outcome.

show that our results are robust to other functional forms in Section 2.4.1. The dummy variables D_{it}^{0}, D_{it}^{1}, D_{it}^{2-11}, D_{it}^{12+} indicate that a department is 0, 1, 2 to 11, and 12 or more months after the occurrence of an officer death, respectively. The coefficients δ_{it}^{k}, which indicate the time-path of the effect, are the main object of interest.

We include a vector of covariates at the department-by-year level, $X_{i,yr(t)}$ to account for city-level demographic variation (summarized in Table B.1). These controls include city-by-year resident age, sex, and race shares, as well as total population, median household income, poverty rate, and unemployment rate. City-by-month fixed effects, $\pi_{i,m(t)}$, remove all within-city seasonality in the outcome that is constant across years. We also include fixed-effects for year-by-month, θ_t, which account for sample-wide time variation in the outcomes.

Lastly, we include department-specific linear time trends $\gamma_i t$. Crime and arrests are decreasing nationally during our sample period, and locations with higher baseline crime levels are experiencing bigger declines (Ellen and O'Regan, 2009; Friedson and Sharkey, 2015), suggesting the need to account for cross-city differences in the time path of crime and arrests. We include this set of controls to isolate deviations from these downward trends due to treatment. Importantly, including time trends leads to more *conservative* estimates of the arrest declines, because without them, earlier periods of arrests prior to a officer death (contained in D_{it}^{0}) may be inflated upward. Indeed, we find consistent results albeit with larger arrest declines when these controls are omitted (Table B.2, specification (12) and Figure B.6). We also show that our baseline results are robust to a parsimonious model with no control variables or time trends, where treatment agencies are matched to control agencies using a nearest neighbor algorithm (Table B.2, specification (13) and Figure B.7).

We consider an officer death event to be any instance where one or more officers in a department died in a particular month.[19] Some cities experience officer deaths at multiple points in time within our sample period. We allow these events enter our specification

[19] In Table B.2, we show that our results are robust to counting each officer death in a city-month as its own event.

additively, denote each officer death event by d, and maintain one panel per city:

$$\begin{aligned} Y_{it} = &\sum_{d} \left(\delta_0 d^{0}_{idt} + \delta_1 d^{1}_{idt} + \delta_{2-11} d^{2-11}_{idt} + \delta_{12+} d^{12+}_{idt} \right) \\ &+ \beta X_{i,yr(t)} + \pi_{i,m(t)} + \theta_t + \gamma_i t + \epsilon_{it} \end{aligned} \tag{2.2}$$

The interpretation of our coefficients δ_k is that they represent the time-path of the effect of the average officer death event (Neilson and Zimmerman, 2014; Sandler and Sandler, 2014). This formulation is equivalent to calculating time period lag variables for each event and then summing these lag variables across multiple events within a police department panel.

A key assumption of our empirical design is that the occurrence of an officer death is not correlated with time-varying shocks to the outcome. A partial test of this assumption is to check that an officer death does not appear to impact an outcome *prior* to the date of the incident. To evaluate this hypothesis, we will also run an event study version of the above regression, where we include indicators for each month around the date of the incident:

$$Y_{it} = \sum_{d} \sum_{\substack{k \in \{-6, 6+\} \\ k \neq -1}} \delta_k D^{k}_{idt} + \beta X_{i,yr(t)} + \pi_{i,m(t)} + \theta_t + \gamma_i t + \epsilon_{it} \tag{2.3}$$

To test that our treatment does not have significant pre-trends, we check that the values of δ_k for $k < -1$ are statistically insignificant. We include event study coefficients that span the 6 months before and after treatment, where δ_{-6} and δ_{6+} are book-end coefficients which include all periods prior to period -6 and after period $+6$, respectively.

We conduct a number of robustness checks to verify the validity of our results and assumptions of our specification which are detailed in Section 2.4.1. These include restricting the analysis to treated cities, estimating the model outcomes in levels and per capita terms, entering multiple officer deaths within a department-month additively, and creating a separate panel for each officer death treatment (vs. each treated city). We pay careful attention to issues raised surrounding difference-in-differences event study models in the literature

(Borusyak et al., 2024; Goodman-Bacon, 2021; Sun and Abraham, 2021) and include checks to address these concerns. Lastly, as referenced above, we re-estimate a parsimonious version of the model with no demographic or time trend control variables, which compares matched treatment and control agencies selected using the nearest neighbor matching algorithm. We likewise display analogous estimates using synthetic control methods, which construct a weighted control group for each treated unit.

2.4 Results

Table 2.2 presents the central results. First, we examine murder crimes and arrests, as these outcomes capture the study treatment of a line-of-duty officer death. These analyses serve to validate the construction and linkage of our data, since our records of officer deaths and outcomes originate from different sources.[20] The top panel shows that the death of an officer while on duty coincides with a 39% increase in murder and a 11% increase in murder arrests. *We interpret this concurrent increase in murder as being a function of the officer death itself.* Indeed, if we adjust the murder outcome to subtract the number of officers killed in a fatality event, there is no significant change in murder in the focal month, as shown in Panel B of Figure 2.2 and the second line of Table B.2, specification (1). Likewise, when the outcome is estimated in levels, the first month coefficient on reported murder is statistically indistinguishable from 1 (Table B.2, specification (8)), corresponding to the treatment of the officer death itself. We confirm the unexpected nature of treatment in Figure 2.2, which shows that there are no pre-trends in this outcome preceding an officer death.

Arrest activity is highly responsive to an officer death in the short-term. Arrests decline by 9.5% in the month of an officer death, and these declines are similar in percentage magnitude across index (8.3%) and non-index (8.9%) arrests. The arrests for the lowest level offenses, "quality of life" arrests, decline at a rate of 9.4%. While the percentage declines are similar

[20] For all analyses where violent crimes and arrests are the outcome, we exclude murder offenses.

in magnitude across categories, the volume of arrests is greater for non-index and quality of life offenses, so these categories experience a greater decline in total volume. Declines in traffic stops are large, but they are insignificant. The magnitude of each of these coefficients are roughly halved in the second month after the officer death. For nearly all arrest types, the effects are smaller and insignificant three to twelve months (the long-term effect) after the incident.[21] Overall, the event study versions of the arrest results in Figure 2.3 confirm the pattern of decreases in the first two months following an officer death and also provide evidence that there are no pre-trends in these outcomes.

Relative to the treatment group mean, the arrest decline in the two months following an officer death corresponds to an average decrease of 134 arrests, of which 19 arrests are for index violent and property crimes, 70 arrests are for "quality of life" offenses, and 44 arrests are for other non-index offenses in each treated city.[22] Collectively, the estimates show that police reduce their enforcement activity following an officer death over the short-term and that this reduction is driven by a decline in enforcement of less serious offenses.

Why might officers reduce the number of arrests they make following the death of a peer? When an officer is killed, peers in their department are sharply reminded of the potential risks of working in law enforcement. This salient shock to the perception of risk could increase fear among officers and cause them to take new protective actions. Officers have a high degree of discretion over the ways in which they engage in their jobs in the field. In particular, interactions with civilians which are "officer-initiated," which can include arrests, do not occur unless an officer is motivated to participate in the activity. Following the death of a colleague, officers may feel that engaging in an adversarial interaction with a suspect is not worth the potential risk of injury or death that could occur during that interaction. Our finding that officers decrease low-level arrests suggests that they may adjust their threshold

[21] An exception is the long-term coefficient for violent arrests; however, this long-term effect is not visible in the event-study version of the model, where there is no evidence of joint significance of post-period indicators (Figure 2.3).

[22] The sub-category arrest counts are calculated from the coefficients on each arrest type and therefore do not sum directly to 134.

for what types of offenses are serious enough to be worth their enforcement effort.

How do crime outcomes change after an officer death? Crime rates may be viewed as a marker of police effectiveness, we are interested in how changes in the emotional state of officers could have consequences for public safety. The third panel of Table 2.2 shows that crime and community activity *do not* increase in the ensuing period. We find small and statistically insignificant estimates for both violent and property crimes. Our estimates rule out increases in index crimes of more than 3.6% (4.6%) in the month of an officer death (month after) with 95% confidence. Over the longer-term, the estimates rule out a 2.6% increase in index crime. While we observe a negative and significant long-term coefficient for violent crime, this effect is not evident or significant in the dynamic event study estimation (Figure 2.4).[23] Here, the lack of evidence of pre-trends is especially important; these plots confirm that officer deaths do not occur after an uptick in crime.

We next investigate changes in 911 calls for service. This outcome is a function of crimes that occur and victim decisions to report these crimes but is not a function of police decisions to officially record crimes or police enforcement decisions. This less "filtered" proxy for criminal activity also does not increase after an officer death. Our point estimate for the short-term 911 call response is close to zero, and we can rule out a greater than 3.9% (4.9%) increase in 911 calls in month 0 (month 1) and a 3.3% increase over the remainder of the year after an officer fatality.

Lastly, we find that the number of fatal traffic accidents does not increase. The traffic fatality outcome has the advantage that it is a function of traffic offenses and is a proxy for reckless driving but is not related to victim or police reporting, as nearly all fatal traffic accidents are reported. Despite the large decrease in the traffic stop point estimates following an officer death, fatal traffic accidents do not change.[24] Here, we can rule out traffic fatality

[23] In robustness Table B.2, the long-term violent crime effect has small negative coefficients, but significance varies across specifications. Since this coefficient is unstable, we hesitate to interpret it as a true causal effect.

[24] While enforcement of traffic offenses has been shown to affect traffic offending (Gonçalves and Mello, 2023), existing studies primarily focus on state highway patrols, which play a larger role in traffic enforcement than municipal police forces, which are the focus of this study.

increases of more than 6.5% in the first month, 4.4% in the second month, and 0.04% in the remainder of the year, with 95% confidence. The long-run impact on traffic fatalities is a marginally-significant *decline* of 2.5%, though we caution against interpreting this finding as a treatment effect given the time lag and lack of a short-term effect.

2.4.1 Robustness Specification Tests

In this section, we conduct several robustness checks to scrutinize our results. First, in Figure B.3, we re-estimate the model dropping one treatment event at a time and plot the distribution of results. This exercise confirms that the estimates are not driven by outlier observations, as this distribution is substantively close to the baseline estimate. Moreover, the alternative estimates are well within the confidence intervals given by the baseline model.

Next, we randomize the timing of officer deaths among treated agencies (holding the number of deaths per agency fixed) and re-estimate the model 100 times using these placebo treatments in Figure B.4. Our model estimate for the first month decline in arrests lies well outside the distribution of estimates in the placebo distribution, confirming that the results we find are actually a function of the treatment and are unlikely to be driven by chance.

Table B.2 includes a number of alternative specification tests, all of which find similar results to our preferred specification. The first specification replicates the baseline results and also includes an adjusted measure of the murder outcome that excludes officer fatalities (1). Using this adjusted outcome, we find no evidence that murders increase, confirming that the spike in murder is due to the treatment of the officer fatality itself.

Next, we show that the results are robust using only the sample to treated cities (2). Our estimates are robust to an alternative model that constructs a panel for each officer death treatment, rather than a panel for each city (3), and the results are also similar when we consider multiple officer deaths from the same event additively rather than as a single event (4). Our estimates are also similar when excluding the city-by-calendar month fixed effects from the model which adjust for seasonality in outcomes that may differ by department (5).

Additionally, we show that the results are robust to adding state-by-year fixed effects to the model, which flexibly control for state-level policy changes (6). Further, excluding arrests for driving under the influence (DUI), the single offense for which we observe the strongest arrest decline (see Section B.2.1 below), does not change the pattern of the results in (7).

The results are also largely similar when we alter the measurement of the key outcomes.[25] For example, the estimates are consistent when we use counts of arrests and crimes as outcomes (8); however, the standard errors are substantially larger, leading to less significant effects for our arrest declines. The results are also highly robust to a per capita model (9) and an inverse hyperbolic sine model (10).

Recent research documents potential issues with the standard difference-in-differences design and suggest modified specifications, and we consider the robustness of our estimates to these approaches (e.g. Borusyak et al., 2024; Goodman-Bacon, 2021). Sun and Abraham (2021) show that event study designs in the presence of treatment effect heterogeneity can produce estimands for each event-time coefficient that are contaminated by coefficients for other time periods. To address this concern, we present their estimator in (11), which explicitly constructs each event-time estimand as a positively-weighted average of cohort-specific treatment effects. We also present a graphical version of their approach with pre-period coefficients in Figure B.5. This methodology confirms our baseline findings, though their specification does require treating each line-of-duty death as its own panel.

The final issue we address relates to department-specific time trends in our outcomes. As we discuss above, crime is decreasing overall during our sample period, and this decline may be more pronounced in treated cities and bias our estimates of arrest and crime impacts. Our baseline specification includes city-specific linear time trends to address this issue, but we consider two alternative specifications to probe the robustness of our results to different ways

[25] Recent work has highlighted concerns with the use of the natural logarithm of y plus a constant as a regression outcome (Chen and Roth, 2023; Mullahy and Norton, 2022), especially in cases where a large share of observations have a y of zero. In our setting, only 99.96% (99.99%) of observations have a count of 0 crimes (arrests), alleviating concerns that our choice of constant in the logarithm is affecting our results. We present here alternative specifications that are recommended by these studies.

of accounting for secular trends. In model (12), we show our baseline specification without controls for department-specific linear time trends. The size of the arrest declines are *larger* in this specification (or less conservative), and we continue to find no positive crime effects in any period and a long-term decline in violent crime. We show in Figure B.6 that the event study estimates without linear time trends look similar to the baseline results.

In model (13), we use a nearest neighbor matching approach to directly match pre-period trends of treated and untreated departments. We use the nearest neighbor matching algorithm to match each treatment event to similar control agency panels using demographic information in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment.[26] Importantly, these models benefit from the matching algorithm's ability to select control agencies with similar pre-treatment levels and trends. After matching, we run a new difference-in-differences regression with this nearest neighbor sample, which *excludes all demographic covariates and time trend controls*. These results are also shown in Figure B.7. In this parsimonious model, we find results that are consistent with our baseline model.

2.5 Heterogeneity

Next, we consider how our results vary by characteristics of the city and line-of-duty death. In particular we ask whether the null finding of no increase in crime persists for subsamples of cities with particularly large or sustained declines in arrests following an officer death.

In Appendix B.2, we explore heterogeneity by arrest and crime sub-type as well as arrestee demographic characteristics. The largest percent declines in arrests (greater than 10%) are for driving under the influence (DUI) and drug sale and possession. We also find a large decline in arrests for weapons violations. Note that these arrest types are more likely to arise from an officer's discretionary interaction rather than a civilian call for service, consistent

[26] The matching variables are lagged values of log counts of violent and property crimes and arrests for periods -1, -2, and -3, and the slope of these outcomes between periods -3 to -12, as well as the treatment year city-level poverty rate, share white, share with a high school degree or less education, and log population. For each treatment event, we keep the 5 "nearest" agencies as controls.

with our hypothesis that officers primarily cut back on discretionary enforcement. We find similarly sized declines in arrests for all race, age and gender subgroups of arrestees, which translate to larger per capita declines in arrests for Black and male individuals given higher baseline rates for these groups.

2.5.1 Size of Arrest Decline and Crime Effect

To identify heterogeneity in treatment effects, we estimate an individual arrest treatment effect for each death event in our sample using a synthetic difference-in-differences design (Arkhangelsky et al., 2021). This approach constructs a control unit for each treatment that is a weighted average of multiple control units that minimizes the difference in a set of pre-period characteristics between the treated agency and the weighted "synthetic" control unit (see Appendix B.3 for additional detail).[27] We then take the difference between treatment and synthetic control in the post-period to identify the effect on arrest and crime rates for each death event, which we denote generically by $\hat{\tau}_i$.

We plot the average of arrest and crime outcomes across our treatment events versus the synthetic controls in Figure B.7. The plot confirms that, as in our nearest neighbor approach, treated and synthetic control agencies are well-matched on pre-period trends, and our post-period effects are consistent with our baseline results, showing a one to two month arrest decline but no change in crime.

We then ask how these event-specific treatment effects, $\hat{\tau}_i$, vary with city and incident characteristics. We focus on the first-month arrest effect estimates and regress these estimates on city and incident characteristics in Table B.3, where observations are weighted by the

[27] For each event, we restrict attention to a "donor pool" that consists of the 100 nearest neighbor cities, identified using the matching procedure described in robustness check (13) above (of Table B.2), who do not have an officer death of their own in the year before and after the treated agency's event. Then, we implement a synthetic difference-in-differences estimation method for each treatment event. The synthetic control weights are determined from the matching variables: log population, city-level poverty rate, share white and share with a high school degree or less education. From the procedure, we obtain the treatment and control series for each time period. We conduct a placebo method for estimating standard errors by replacing each treatment unit with a control unit and we repeat this procedure 100 times.

inverse variance of each $\hat{\tau}_i$. In column 1, we find that arrest declines are more negative in cities with smaller population and cities with residents who have higher levels of education, while the relationship to city crime rate is not significant. These patterns are consistent with the officers in these cities facing a bigger perceived shock from a peer death since total officer employment and networks are likely smaller in small cities.

In column 2, we regress the arrest effects on characteristics of the officer death event. We find no relation between the arrest impact and the deceased officer's race or gender, or whether the incident was during a traffic stop. We find larger reductions in arrests in cases when the suspect in the officer death was not apprehended within 48 hours. Note that the average arrest reduction in these cases versus the 48 hour apprehensions are both substantial, -0.18 and -0.07, respectively. Column 3 combines the city and incident characteristics. The coefficients on log population and share with less than a high school education, have a similar magnitude as in Column 1, but only the education variable is still statistically significant.

Do cities that we predict to have larger arrest declines also experience increases in crime? We next use the agency and event controls to construct a *predicted* version of our treatment-specific estimates, $E(\hat{\tau}|X)$.[28] The benefit of this approach is that it splits the sample into groups with different sizes of arrest decline leveraging only variation in pre-treatment characteristics, X. We split treatments into three groups based on $E(\hat{\tau}|X)$: the top quartile of predicted arrest declines, the interquartile range, and the bottom quartile.

Figure 2.5 plots the arrest and crime changes over time for these three separate groups, where arrest and crime changes are individual treatment-specific synthetic control estimates ($\hat{\tau}_i$) and groups are defined by quartiles of the predicted arrest effect, $E(\hat{\tau}_i|X)$.[29] The left panel plots the pattern for officer deaths with the largest predicted arrest declines, while the right panel plots the pattern for the smallest predicted arrest declines. Despite the substantial

[28] This prediction is constructed using a "leave-out" version of Column 3 of Table B.3, where each estimate of $E(\hat{\tau}_i|X)$ is produced with coefficients estimated from a regression using all treatments other than i.

[29] Table B.4 shows the average agency and death characteristic covariates for each binned group. In each figure, we present the interquartile range of estimated $\hat{\tau}_i$ effects in the dashed gray lines around the median treatment effect.

variation in arrest declines, there is no systematic increase in crime across any group. In particular, we do not identify an increase in crime for treatments in the top quartile of arrest declines, where the median arrest decline is approximately 15%.

While the patterns above suggest a null crime effect even for agencies with large arrest declines, this analysis is limited to the predictable variation in arrest effects based on city and incident characteristics. We next investigate variation in effects across cities based directly on the estimated magnitude of arrest decline. To do so, we return to our baseline estimation strategy from Section 2.3. We first estimate residuals of arrests and crimes conditional on the fixed effects and controls in the model but excluding the treatment indicators, D_{it}. We then calculate the difference between residuals in the months following an officer death versus the residual for the month prior to the officer death, $t = -1$, for both the crime and arrest outcomes. These differences in residuals approximate the effect of an officer death on both arrests and crime rates in each city. We estimate a local linear regression between these two residuals, and we construct our 95% confidence intervals using a bootstrap procedure.[30][31]

Figure B.8 plots the residual change in arrest against the residual change in crime, allowing us to trace an "arrest to crime curve." The top figure presents the crime residuals for the first month and shows a flat relationship with the size of an arrest decline. Within a range of a 20% arrest decline to no change in arrests, the standard errors of the local linear regression reject crime increases of more than 3.4% with 95% confidence. In Panels B and C, we plot the crime residuals for the entire post-treatment year, and we similarly find a flat relationship with no evidence of crime increases for any magnitude of an arrest decline.

[30] Standard errors (dashed lines) are produced by reproducing the results through block bootstrapping (re-sampling police department panels) 200 times and plotting the 5th and 95th percentile of the local linear regression lines from these iterations.

[31] We find similar results when conducting this exercise using synthetic control estimates of arrest and crime effects for each treatment, using the procedures described above. We present the simpler version using constructed residuals here, as we cannot replicate the standard error bootstrap for the synthetic control version, given computational demands.

2.5.2 Length of Arrest Decline and Crime Effect

To examine heterogeneity by duration of arrest decline, we take the residuals estimated above (Section 2.5.1) and calculate for each city the number of consecutive months after an officer death where the residual is lower than the residual for the month prior to the death. We bin arrest decline durations into groups from 0 months to > 5 months. We then plot the post-fatality crime residuals, separately by length of the arrest reduction in Figure B.9.[32]

The top panel presents the crime impact for the first month. Perhaps unsurprisingly, the average residual crime effect is close to zero for all time horizons, since a sustained arrest decline is not likely to lead to a markedly different impact in the first month. This provides a placebo test that agencies with different durations of decline are not experiencing different short-term crime responses. In the bottom panel, we plot the crime residuals averaged over the entire year after the officer death. Over this longer time horizon, we continue to find average effects that are small and statistically insignificant for all durations of arrest decline.

In this exercise, we stratify our sample by an outcome of the treatment rather than using pre-treatment experimental variation in the duration of arrest decline. As a result, we do not claim to have identified the causal impact of arrest declines at various durations. Similar caution is needed in interpreting our second analysis in Section 2.5.1, which stratifies effects by magnitude of arrest decline using estimated residuals from our model. Note, however, that this issue is not a concern for the above analysis that stratifies estimates by predictions of treatment-specific synthetic control estimates. Across these three tests, the results are remarkably consistent; they imply that there is not a threshold magnitude or duration of arrest decline within our sample for which crime increases.

[32] Similar to our arrest-to-crime curve estimation, we estimate confidence intervals using a block bootstrap, re-sampling police department panels in 200 iterations. In each iteration, we re-calculate the number of months with residuals lower than the pre-period month and re-group departments into duration bins. We then calculate the average crime residual for each group, $\hat{\mu}^b$. We use quantiles of $\hat{\mu}^b$ to determine the 95% confidence interval (Efron, 1982).

2.6 Mechanisms

2.6.1 Is the Decline in Arrests a Behavioral Response?

We argue that the arrest decline after an officer death is a behavioral response by fellow officers caused by heightened fear. Here, we consider the alternative explanation that there is a reduction in effective manpower, from the deceased officer or from their peers.

Quantitatively, our observed arrest declines are too large to be solely due to a reduction in effective manpower. Under the conservative assumption that half of a department's officers are patrol officers who regularly make arrests, the average officer in our treated cities makes 4 arrests per month. In contrast, the first month coefficient in our models implies an average decline of 92 arrests. Under the additional conservative assumption that officers are given ten days of bereavement leave,[33] this decline would correspond to 68 officers taking leave, or a quarter of the average treated department's patrol force. Even if the officer who died was exceptionally active, it is very unlikely that the loss of the deceased officer is driving the results or that one in four officers would stop making arrests after a peer death.

Alternatively, effective manpower could be impacted if officers are diverted from normal activities to investigate the death of their peer. As a direct test of whether the arrest declines are due to officers investigating their colleague's death, we revisit our analysis of event-specific impacts in Section 2.5. When we restrict attention to officer death events where the suspect is apprehended within 48 hours, the average arrest decline in month 0 is -0.073, which is indistinguishable from our sample-wide effect and provides validation that officer incapacitation is unlikely to be driving the arrest declines that we observe.

We can further validate a behavioral interpretation by examining officer deaths that are caused by accidents rather than felony homicides, events that likewise incapacitate a deceased officer. Table B.5 estimates the arrest and crime results for these events, which are nearly

[33] Ten days leave is higher than what we observe anecdotally, and we pick this number to be conservative. Our online searching indicates that three days leave is a common amount offered, e.g.: https://www.tdcj.texas.gov/divisions/hr/benefits/leave-paid.html.

all a result of car accidents. Here, officer fatalities are not counted as murders given their accidental nature. Officers do not respond to these events by reducing the number of arrests that they make and there is also no change in crime rates. This exercise shows that on-the-job fatalities caused by felony incidents are more impactful in changing behavior.

Lastly, we investigate this question through our rich 911 data. Our raw call data include some officer-initiated interactions (e.g. traffic stops, on sight investigations), which are discretionary actions that often lead to an arrest. Table B.5 shows that, in the month of an officer death, the volume of officer-initiated interactions declines by 4.7%. In addition, we can construct a measure of changes to officer presence in the 911 data. In 51 of the 56 cities in this sample, we geocode calls to Census block groups. In each city-day, we calculate the share of Census block groups with a 911 call or officer-initiated interaction, which we average to the city-by-month level. This measure reflects how much of a city is visited in the average day. Table B.5 shows null impacts on officer presence after a peer death. Since civilian contacts to the police and officer presence do not change, the decline in officer-initiated incidents is further evidence more consistent with a behavioral response than incapacitation.

Employment Outcomes – Beyond temporary leave, officers may choose to exit the police profession entirely following the death of a peer. We investigate this possibility by linking records of officer deaths to data on employment spells for police in Florida from the Florida Department of Law Enforcement. Results of this analysis are presented in Panel A of Table B.5. We are able to confirm the officer death effect in this data but fail to find robust evidence of any behavioral responses on officer employment. On net, the number of full-time equivalent officers is unchanged, and there is no systematic change in quits, firings or hirings. If anything, officer quitting appears to slightly decline in the long-term period (effect size is equivalent to 1 additional officer employed off of a base of 513 during this period). Collectively, this evidence shows that peer deaths do not motivate officers to quit policing.

2.6.2 Do Police Change other aspects of their Behavior?

Police Discretion in Recording Crimes – One alternative explanation for why we find no increase in crime after an officer death is that police not only reduce the number of arrests that they make but potentially also reduce the number of crime reports that they choose to file. In several cases, police have some discretion over which victim complaints are officially filed as criminal incidents. If officers are less likely to file criminal reports after a peer officer death, the estimates of changes to reported crime could be biased downward. Indeed, a large literature in criminology has highlighted concerns about the potential for crime reports to be manipulated by changes in officer reporting standards (Bayley, 1983; Levitt, 1997; Marvell and Moody, 1996; Mosher et al., 2010). Within our 911 data, we are able to measure changes in officer reporting among cities that record whether a call results in a criminal incident report being written. This metric allows us to directly test whether the treatment of an officer death systematically changes the likelihood that police officers choose to report crimes, conditional on a 911 call response. In Table 2.2 and Figure 2.4, we find that this conversion rate is unaltered by an officer death on average, suggesting that officers do not respond to these events by reporting fewer criminal incidents. Our estimates are quite precise and can rule out a greater than 1.4% decrease in the reporting rate in the month of an officer fatality, off a base of 26%. This test provides greater confidence in the null effects we identify for reported index crimes using the FBI UCR data.

Police Use of Force – It could be the case that officers may not only reduce arrests but also increase use of force following a line-of-duty death, consistent with research conducted in individual jurisdictions (Holz et al., 2023; Legewie, 2016). While these single-city studies examined the full range of use of force in response to a peer injury or death, there are no nationwide databases available for measuring non-fatal use of force. Instead, we examine this question using national data on civilians killed by police from the UCR Supplemental

Homicide Report and the crowd-sourced data resource, *Fatal Encounters*, in Table B.5.[34] For both outcomes, we find a small and statistically insignificant coefficient for the first-month effect of an officer death, implying no change in use of force. In the long-run, we find a marginally significant increase in only the Fatal Encounters measure. Both data sources are known to suffer from significant under-reporting and to have varying quality over time (Gonçalves, 2021; Loftin et al., 2017; Renner, 2019), so we consider these results to be suggestive evidence that there is no use-of-force response to an officer fatality.

2.6.3 Do Officer Deaths Impact Civilian Behavior?

We are also interested in whether an officer death itself directly causes civilian criminal activity or victim reporting behavior to change. In particular, it might be the case that civilians fear that they will face a stronger punitive response after an officer death and are consequently deterred from offending. Any decline in offending resulting directly from the reaction to an officer death could mask an increase in crime resulting from the reduction of arrests. To address this question, we ask whether cities with officer fatalities that have no arrest declines actually experience a *reduction* in crime, as the above story would suggest. In Section 2.5.1 above, we split the sample by the size of arrest declines in treated cities. We observe a flat relationship between the magnitude of arrest decline and level of crime change, and we do not see any declines in crime for departments with no arrest declines. This pattern supports a story where officer deaths generate fear and behavioral responses among peer officers but do not directly impact civilian offending behavior.

A second concern relates to whether we might be missing changes in crime that occur for categories outside of the most serious UCR Index I offenses. Here, we can examine effects from our 911 data collection. These data cover a larger range of crimes than the UCR crime

[34] This analysis excludes treatment events where the suspect of an officer fatality is shot and killed in the event to avoid a mechanical effect of the treatment on the outcome. The regressions include a panel for each treatment event in the data. Fatal Encounters was established in 2013 and includes back-filled data for earlier years; we restrict attention to records from 2010-2018 to address data quality issues in the data.

reports. The fact that we continue to find no impact on this broader indicator of crime indicates that we are not missing impacts on lower-level offending.

One way to further probe the question of whether an officer death affects civilian behavior is to ask whether officer deaths are actually salient to civilians. Figure B.10 plots the Google Trends search intensity of 71 officers killed in the field, which we compare to 137 high-profile deaths of *civilians* at the hands of police since 2010 using searches from the U.S. state where each event occurred.[35] Google trends provides a metric of *relative search* volume that is normalized between 0 and 100 and is a function of terms entered in a search (selected by the user). We include topical searches for heart attacks as a benchmark (as heart disease is the leading cause of death in the U.S.), which is searched relatively frequently and is not seasonal in search volume. This benchmark allows us to view a perceptible increase in searches at the time of the events and to compare the relative effect of events across time and space as well as between line-of-duty deaths and officer-use-of-force killings.[36]

In relative terms, the public is far less aware of the officer deaths than civilian deaths at the hands of police, with the average civilian death having a search popularity metric that is over three times the size of the average officer death. Search intensity for a civilian death persists to some degree in the weeks following a death, with subsequent spikes that may be associated with protests of the incident or an announcement of whether the involved officers will be charged. In contrast, the public awareness of an officer death is quite small and quickly levels to zero after these events. This evidence implies that the awareness of these deaths

[35] Information on high-profile deaths of civilians is taken from "Black Lives Matter 805 Resource and Action Guide." Information on officer line-of-duty deaths is acquired from the *Officer Down Memorial Page* and is described in more detail in Appendix B.5. The sample frame begins in 2010 to match the coverage of this list. We search each civilian and officer death separately within the state where the event occurred and plot the average within-state search intensities alongside the benchmark search term.

[36] All quantities are reported relative to the time period and search term with highest search volume, which is given a value of 100. We include topical searches for heart attacks as a benchmark (as heart disease is the leading cause of death in the U.S.), which is searched relatively frequently and is not seasonal in search volume. Given this type of output, the choice of an appropriate benchmark search term is critical, as a benchmark that is too popular would completely dwarf any evidence of search volume for officer death events. For example, benchmarks that are sufficiently more popular, such as "Google" or "Youtube", would negate any perception of relative search volume for both civilian and officer deaths. We purposefully select our benchmark to show that there is evidence of some salience of officer deaths in our data.

among community members is relatively minimal and short-lived. We hypothesize that officer deaths are thus unlikely to spark a change in criminal activity or civilian behavior.

If it is the case that an officer death has limited direct effects on civilian or offender behavior, it is also possible to connect our findings to the open and unresolved question of whether and how changes in marginal arrest enforcement may impact crime. In our setting, the declines in arrests we observe are driven by large reductions in low-level crimes, which could have limited public safety value. Viewed through this lens, our work provides new insights about the importance of changes in arrest enforcement to public safety, and it is useful to benchmark our estimates to the prior work on the impact of police manpower or presence on crime. To do so, we convert our estimates into an crime-to-*total arrest* elasticity by dividing our violent and property crime coefficients by the total arrest coefficient for period 0.[37] Our property and violent crime elasticity estimates are not significantly negative, -0.10 for property crime and 0.38 for violent crime, and do not statistically differ from 0. Figure B.11 shows that these crime-to-*arrest* elasticities are notably less negative when compared to the elasticity estimates of *police manpower* on crime, which has generally found large and significant reductions in crime from increased police employment (e.g. Chalfin and McCrary, 2018; Chalfin et al., 2022; Evans and Owens, 2007; Mello, 2019; Weisburst, 2019). These elasticity comparisons serve to emphasize that our null results for crime given a change in *arrests* are small relative to the crime increases we would expect from a comparable percent decline in manpower. In this way, our results are consistent with the view that police deterrence operates primarily through officer presence rather than arrest activity (Owens, 2013). To put our magnitudes in context, we calculate that if all U.S. departments reduced their arrests for only two months per year by the average impact we observe after a line-of-duty death, this decline would translate to about 116,000 arrests foregone annually and a statistically insignificant effect of 13,000 more crimes.

[37] The associated standard errors are constructed with the delta method: $var(Elasticity) = var(\beta_{crime})/\beta^2_{arrest} + var(\beta_{arrest}) * \beta^2_{crime}/\beta^4_{arrest}$.

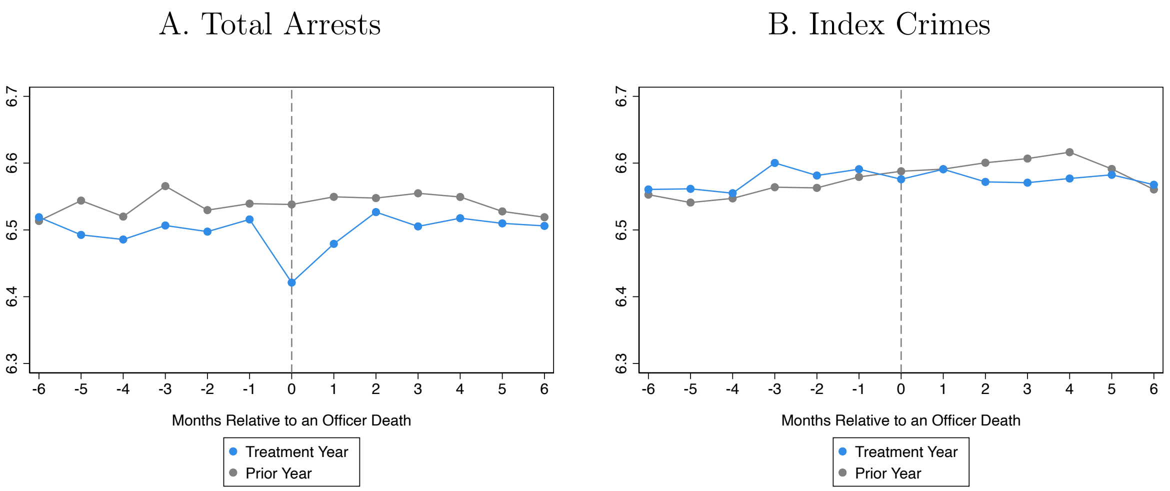
2.7 Conclusion

How does fear affect officer behavior and police efficacy? Policing is a dangerous and high-stakes profession where the undercurrent of fear has the capacity to influence officer actions, with potential adverse consequences for public safety. We find that police respond to an officer fatality by substantially reducing the number of arrests they make, with the largest effects for low-level arrests. When an officer chooses not to engage with a suspect and make an arrest, the officer is minimizing their likelihood of interacting with an individual who could cause that officer physical harm; thus, the arrest reductions we observe are consistent with risk mitigation due to heightened fear. While we observe a sharp 10% decline in arrests in the one to two months following an officer death, we fail to find evidence that this shock reduces public safety. Further, we do not find that crime increases in settings where officers reduce arrests by larger amounts or longer durations. Collectively, the results imply that fear reduces enforcement but is unlikely to contribute to higher crime.

We find limited evidence that officers change their behavior in dimensions other than arrest enforcement or that civilian or offender behavior is directly impacted by an officer death. Our work may thus offer suggestive insights about the impact of marginal arrest reductions on crime. Such questions are critical to the broader debate about law enforcement's heavy reliance on policing low-level offenses, an approach popularized since the 1980s as part of a "broken windows" policing philosophy (Bratton and Knobler, 2009; Kohler-Hausmann, 2018; Riley, 2020; Silva, 2020; Speri, 2020; Zimring, 2011). Related work on policies that affect arrest enforcement, such as changes in the felony classification (Dominguez et al., 2019) or the decriminalization of marijuana (e.g. Adda et al., 2014; Mark Anderson et al., 2013) have shown limited or mixed evidence of crime increases. Alternatively, some researchers have found crime-reducing benefits of particular types of enforcement, such as "hot spots" policing (e.g. Blattman et al., 2017) and forms of "focused" deterrence that target small groups of frequent offenders (Braga et al., 2018; Chalfin et al., 2021a). More research is needed to understand which forms of arrests and sanctions provide crime-reducing benefits.

2.8 Figures & Tables

Figure 2.1: Unadjusted Data Around Events, Log Outcomes

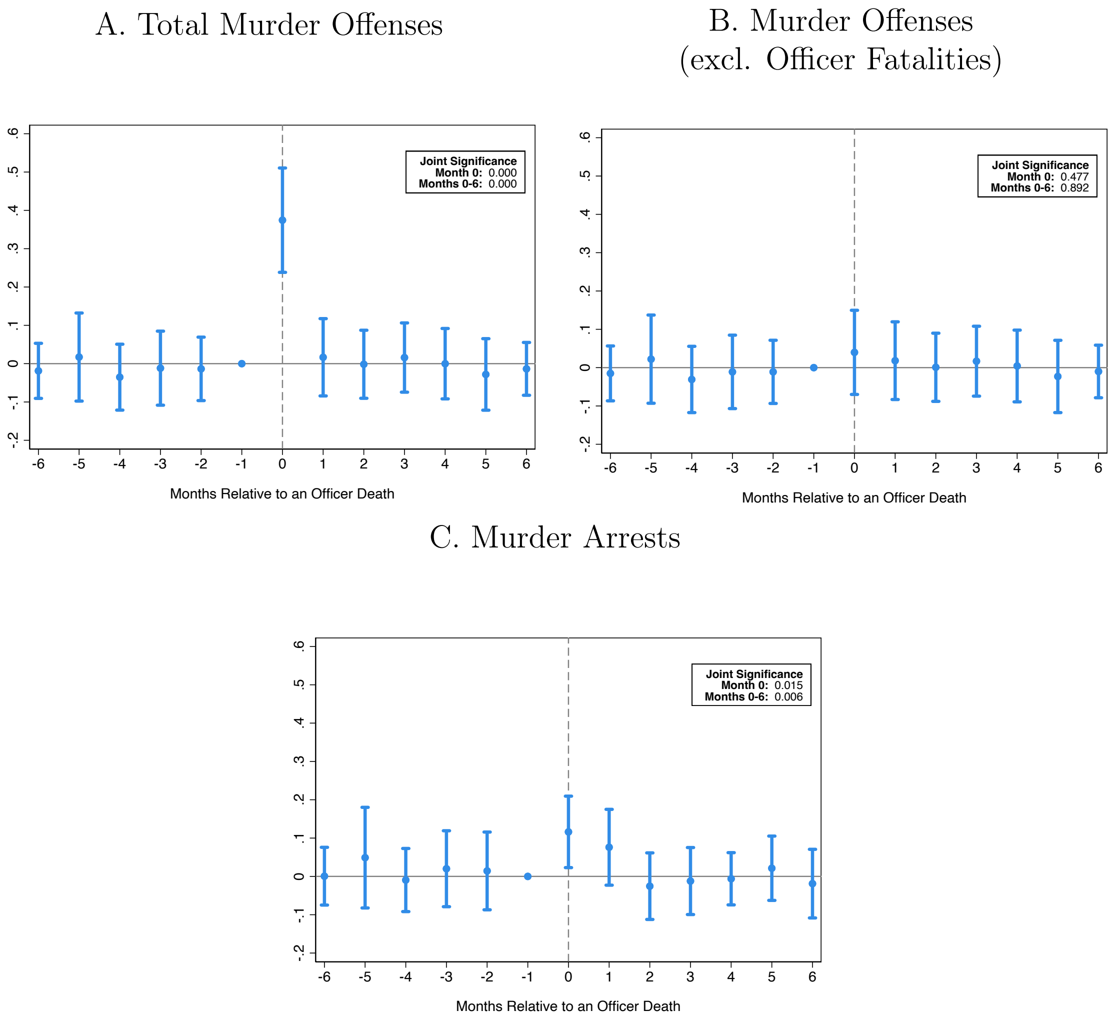


A. Total Arrests

B. Index Crimes

Notes: This figure plots the unadjusted data around the officer death events. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$. There are 125 officer death events in 76 agencies after excluding events that do not have enough periods before and after the event. Index crimes include rape, robbery, aggravated assault, burglary, theft, and motor vehicle theft.

Figure 2.2: Event-Study: Murder Outcomes



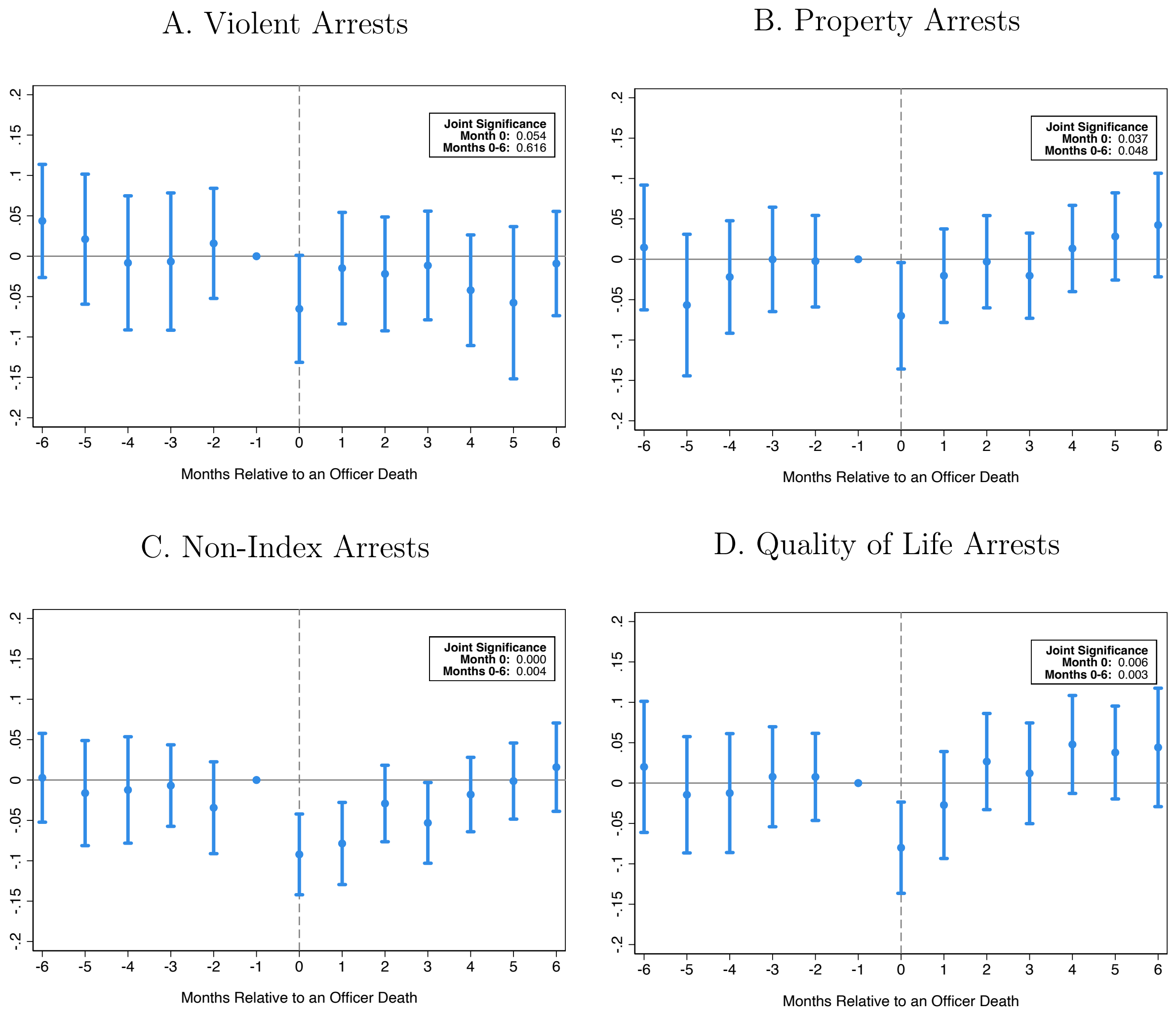
A. Total Murder Offenses

B. Murder Offenses (excl. Officer Fatalities)

C. Murder Arrests

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Months -6 and 6 include all months before month -6 and all months after month 6, respectively. Standard errors are clustered at the department level.

Figure 2.3: Event-Study: Arrests



A. Violent Arrests

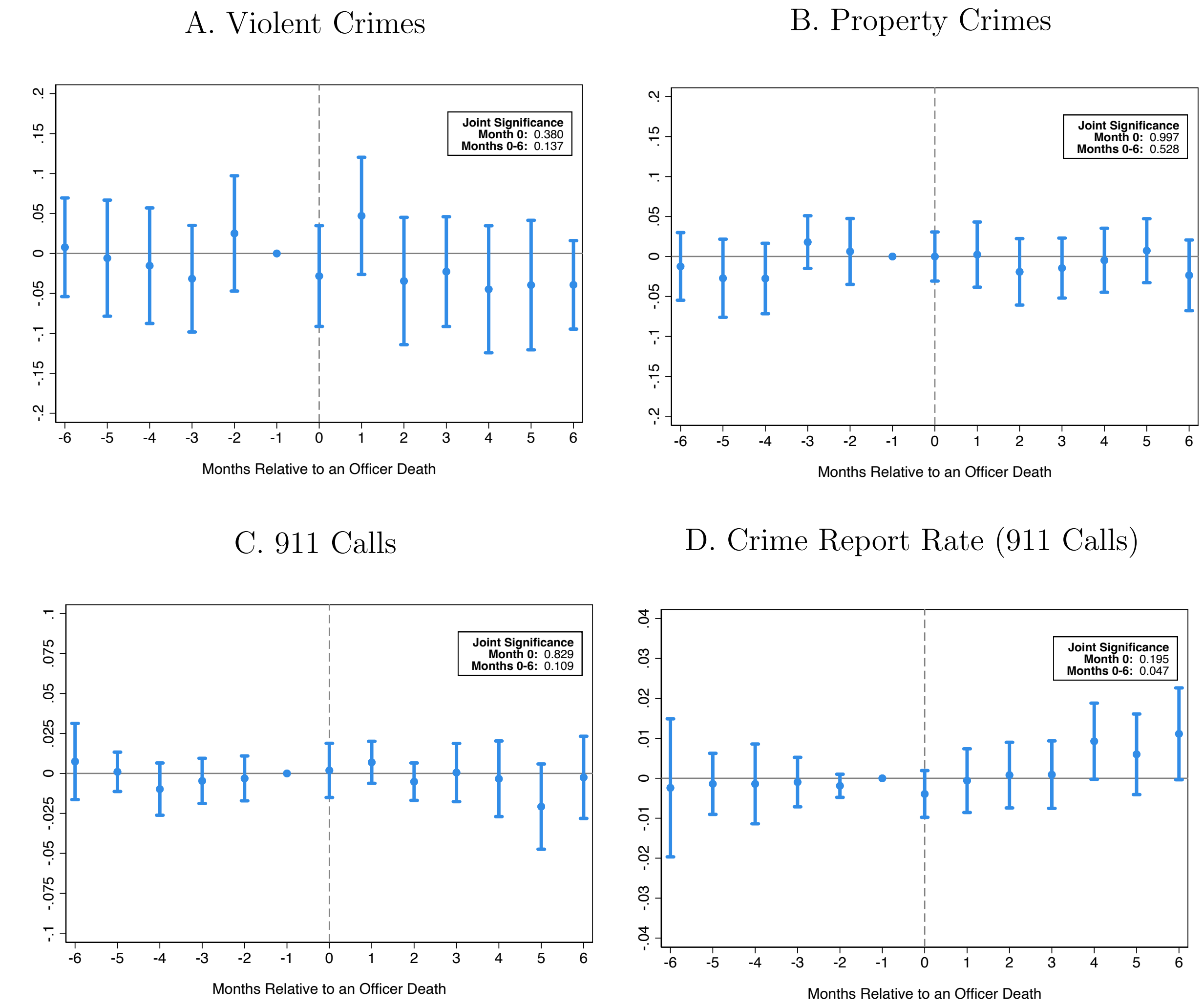
B. Property Arrests

C. Non-Index Arrests

D. Quality of Life Arrests

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Months -6 and 6 include all months before month -6 and all months after month 6, respectively. Standard errors are clustered at the department level. See Table B.7 for the list of arrest sub-types. Violent arrests include rape, robbery and aggravated assault. Property arrests include burglary, theft and motor vehicle theft.

Figure 2.4: Event-Study: Crimes and 911 Calls


A. Violent Crimes

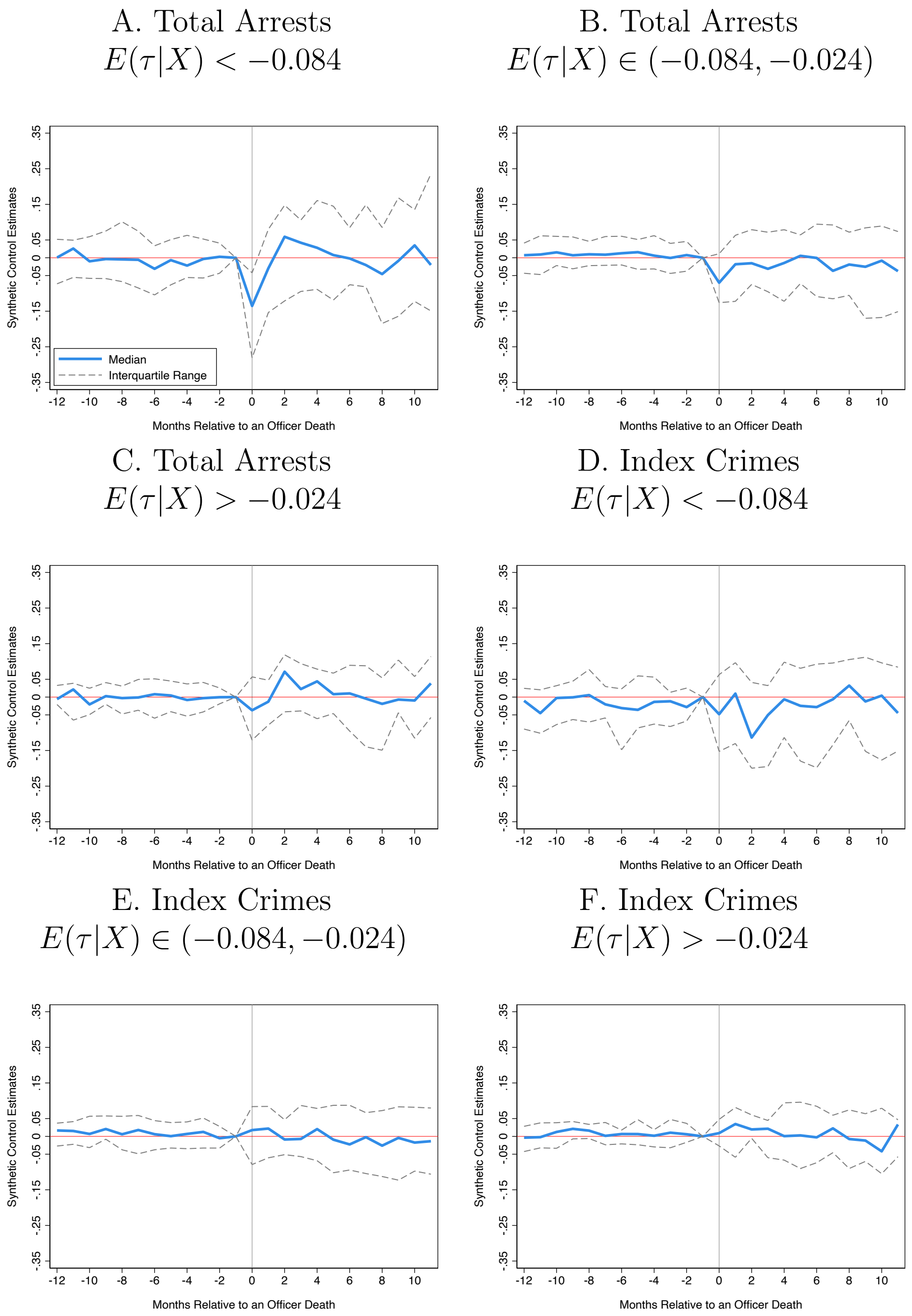
B. Property Crimes

C. 911 Calls

D. Crime Report Rate (911 Calls)

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Months -6 and 6 include all months before month -6 and all months after month 6, respectively. Standard errors are clustered at the department level. Violent crimes include rape, robbery, and aggravated assault. Property crimes include burglary, theft, and motor vehicle theft. "911 Calls" are records of all police calls for service at the city by monthly level. "Crime Report Rate (911 Calls)" is the share of calls that result in a crime being recorded by police.

Figure 2.5: Plotting Treatment Effects by Predicted Arrest Decline Quartiles



A. Total Arrests
$E(\tau|X) < -0.084$

B. Total Arrests
$E(\tau|X) \in (-0.084, -0.024)$

C. Total Arrests
$E(\tau|X) > -0.024$

D. Index Crimes
$E(\tau|X) < -0.084$

E. Index Crimes
$E(\tau|X) \in (-0.084, -0.024)$

F. Index Crimes
$E(\tau|X) > -0.024$

Notes: A set of 100 nearest-neighbor agencies that do not experience officer death within a year of treatment agency's officer death event is generated by matching on demographic characteristics in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment. Then, from this set, a synthetic control agency is created by matching on demographic characteristics in the treatment year. There are 120 matched pairs. The synthetic difference-in-differences is estimated and post-period treatment effects are obtained. Panels A, B and C show the treatment effect for total arrests, separately by predicted arrest effect quartiles. Panels D, E and F show the treatment effect for index crimes, separately by predicted arrest effect quartiles.

Table 2.1: Summary Statistics

		Full Sample			Treated Agencies	
	Mean	S.D.	N	Mean	S.D	N
Murder Outcomes						
Murder Offenses	0.221	(1.617)	354504	2.350	(6.357)	18510
Murder Arrests	0.165	(1.266)	354507	1.574	(4.890)	18510
Policing Activity						
Arrests	151.9	(479.4)	354507	964.5	(1716.5)	18510
Index Arrests	28.4	(94.2)	354507	177.0	(339.0)	18510
Violent Arrests	8.4	(41.1)	354507	62.0	(157.7)	18510
Property Arrests	20.0	(58.2)	354507	115.1	(200.4)	18510
Non-Index Arrests	40.9	(136.9)	354507	268.2	(505.4)	18510
Quality of Life Arrests	82.6	(263.9)	354507	519.2	(931.9)	18510
Traffic Stops	6200.8	(9489.0)	1491	9130.5	(11114.0)	423
Crime and Community Activity						
Index Crimes	140.0	(549.6)	354507	1023.5	(2032.5)	18510
Violent Crimes	18.3	(105.0)	354507	165.8	(412.0)	18510
Property Crimes	121.6	(452.9)	354507	857.7	(1654.9)	18510
911 Calls for Service	9488.5	(13397.3)	5904	20283.0	(19083.3)	1374
Crime Report Rate (911 Calls)	0.26	(0.14)	5151	0.28	(0.11)	1221
Fatal Traffic Accidents	0.26	(1.09)	283906	1.60	(3.61)	17040
Number of Agencies	1578					
Number of Treated Agencies	82					
Total Officer Death Events	135					
Treatments Per City (Treated)	1.65					
Officer Characteristics						
Cause of Death	*Gunfire*: 136		*Vehicular Assault*: 11		*Other*: 4	
Race	*White*: 115		*Black*: 20		*Other*: 16	
Gender	*Male*: 141		*Female*: 10			
Age	36.86	(9.16)				
Experience	11.14	(8.41)				

Notes: The number of agencies, number of treated agencies and total officer death events are from the data with crime and arrest activity outcomes. For the traffic stop outcomes, they are 18, 3, and 5. For the traffic accident outcome, they are 1252, 33, and 74. For 911 call outcomes, they are 56, 9, and 14. All arrest and crime subcategories exclude murder outcomes. Violent crimes and arrests include rape, robbery and aggravated assault. Property crimes and arrests include burglary, theft and motor vehicle theft. See Table B.6 and Table B.7 for the list of crime and arrest sub-types. "Crime Report Rate (911 Calls)" is the share of calls that result in an officer writing a crime incident report. The officer characteristics are from the *Officer Down Memorial Page*. Other causes of death include assault and stabbed.

Table 2.2: Impact of an Officer Death on Policing and Crime

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
Murder Outcomes									
Murder Offenses	0.391***	(0.058)	0.033	(0.039)	0.015	(0.013)	0.22	2.35	354504
Murder Arrests	0.111**	(0.044)	0.071	(0.043)	-0.000	(0.023)	0.17	1.57	354507
Policing Activity									
Arrests	-0.095***	(0.026)	-0.044*	(0.023)	-0.001	(0.023)	151.9	964.5	354507
Index Arrests	-0.083**	(0.033)	-0.024	(0.031)	-0.012	(0.027)	28.4	177.0	354507
Violent Arrests	-0.105***	(0.035)	-0.054**	(0.027)	-0.050**	(0.023)	8.4	62.0	354507
Property Arrests	-0.075**	(0.036)	-0.026	(0.037)	-0.009	(0.031)	20.0	115.1	354507
Non-Index Arrests	-0.089***	(0.024)	-0.076***	(0.026)	-0.013	(0.022)	40.9	268.2	354507
Quality of Life Arrests	-0.094***	(0.037)	-0.042	(0.032)	0.007	(0.030)	82.6	519.2	354507
Traffic Stops	-0.068	(0.107)	-0.146	(0.122)	-0.021	(0.094)	6201.7	9130.5	1477
Crime and Community Activity									
Index Crimes	0.003	(0.017)	0.015	(0.016)	0.000	(0.013)	140.0	1023.5	354507
Violent Crimes	-0.036	(0.027)	0.039	(0.029)	-0.034*	(0.018)	18.3	165.8	354507
Property Crimes	0.010	(0.018)	0.012	(0.016)	0.002	(0.014)	121.6	857.7	354507
911 Calls for Service	-0.005	(0.016)	0.000	(0.012)	-0.006	(0.012)	9473.7	20229.0	5873
Crime Report Rate (911 Calls)	-0.002	(0.006)	0.002	(0.008)	0.011	(0.008)	0.26	0.28	5127
Fatal Traffic Accidents	-0.023	(0.045)	-0.016	(0.031)	-0.025*	(0.013)	0.26	1.60	283906

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$ and outcome means are given in levels. Standard errors are clustered at the department level. The number of agencies, number of treated agencies, and total officer death events for crime and arrest outcomes are 1578, 82, and 135, respectively. For the traffic stop outcomes, they are 18, 3, and 5. For the traffic accident outcome, they are 1252, 33, and 74. For 911 call outcomes, they are 56, 9, and 14. All arrest and crime subcategories exclude murder outcomes. Violent crimes and arrests include rape, robbery and aggravated assault. Property crimes and arrests include burglary, theft and motor vehicle theft. See Table B.6 and Table B.7 for the list of crime and arrest sub-types. "Crime Report Rate (911 Calls)" is the share of calls that result in an officer writing a crime incident report. * p<0.1,** p<0.05, *** p<0.01.

Chapter 3

The Impact of Cash Transfers to Poor Mothers on Family Structure and Maternal Well-Being

(with Anna Aizer, Shari Eli and Adriana Lleras-Muney)

3.1 Introduction

Since the implementation of the first means-tested cash transfer program in the US in 1911, the Mothers' Pension (MP) program, critics have argued that welfare leads to the erosion of families by incentivizing mothers to remain single and have children out of wedlock, thus trapping women and their children in a cycle of poverty (Skocpol, 1995; Chappell, 2011). Because of their high poverty rates, single-headed households with children have always been the main target of means tested transfer programs. Indeed today, 87% of adult welfare recipients are unmarried (ACF, 2021), suggesting to some that welfare disrupts family formation. However, there is little empirical evidence to support this claim. Existing research on the short run effects of welfare on marriage and fertility is ambiguous, and there is a lack of empirical research on the lifetime effects of welfare on family structure and maternal well-being.

In this paper, we construct a new dataset and exploit a novel identification strategy to estimate the lifetime effects of the MP program on family structure and maternal lifetime well-being. The MP program was first implemented in 1911 in Illinois, adopted across most states by 1920, and finally replaced in 1935 by the federal Aid to Dependent Children (ADC) program, the precursor to Temporary Aid to Needy Families (TANF), today's welfare program. Before 1911, mothers who could not care for their children were forced to place them in orphanages or training schools. In response to reports of high death rates and poor outcomes of children in these institutions, states established MP programs to provide cash transfers to poor mothers so that they could care for their children at home. Since the inception of MP, commentators have been concerned about the unintended effects of the program with respect to family structure (Leff, 1973). Indeed, MP recipients who remarried lost transfers, encouraging women to remain unmarried; abandoned women could receive transfers, encouraging men to abandon their families; finally, the transfer was an increasing function of the number of children, encouraging out-of-wedlock fertility.

In order to estimate the short- and long-term impact of welfare receipt on marriage and fertility, we construct a novel dataset of over 16,000 women who applied for the program between 1911 and 1930, and follow them from the time of application until their death. It is challenging to track women over their lifetimes due to name changes, and for this reason historical work tends to focus on men (see Abramitzky et al., 2021 for discussion). To overcome this, we match data from the program's administrative records to family trees from FamilySearch.org, federal census records and vital statistics. The family trees represent a new source of data that aids in the tracking of changes in marital status (and names) as well as fertility.

The MP records include mothers who applied to the program and were accepted as well as those who passed an initial eligibility screening, but were ultimately rejected. This allows us to implement a novel identification strategy: comparing accepted and rejected applicants to estimate the causal impact of welfare receipt on family structure using OLS and machine learning approaches. Most existing work in this area leverages changes in state laws or policies over time that modify benefits or eligibility, an intensive margin where we might expect more limited effects (Bitler et al., 2004; Hoynes, 1996; Blank, 2002; Blundell et al., 2016; Grogger and Karoly, 2005). Moreover, the women are typically followed for only a short period of time. Our data allow us to estimate how the receipt of cash transfers at the extensive margin affected marriage market outcomes (remarriage, duration to remarriage, and characteristics of the new husband), and fertility over the mother's lifetime.

We find no difference in the lifetime remarriage rates of women who received transfers and those who did not: 47 percent of the women remarried, regardless of welfare receipt – our point estimates are small and we can rule out decreases in marriage rates above 10%. We also fail to find any effects on fertility: Although women who received transfers had more children *before* the application, they did not have more children *after* the application. We can rule out increases in fertility greater than 5%. However, among those who remarried, those with transfers took about a year longer (14 months) to do so.

Could delays in remarriage be welfare-enhancing? To answer this, we develop a model of welfare participation and search in the marriage market, similar to models of unemployment insurance in the market for labor and as suggested by Hutchens (1979). In the model, women search for husbands who are heterogeneous in quality. Receiving welfare benefits, like unemployment insurance, may cause a woman to be more selective when remarrying (her reservation "husband quality" increases) as it enables her to wait longer for the arrival of a preferred partner. Thus, the model predicts that receiving cash transfers can lead to delays in marriage but also increases in the quality of the husband, with implications for maternal and child welfare.

However, similar to the empirical literature on unemployment insurance (UI), which largely finds that UI results in longer unemployment durations but no improvement in the quality of the next job (Card et al., 2007; Lalive, 2007; Van Ours and Vodopivec, 2008; Schmieder et al., 2016), we find that welfare receipt results in longer time to remarriage but does not affect the characteristics of the new husbands. We provide two explanations for this. The first is that the marriageability of women declines with the duration of search. This is analogous to the explanation offered in the UI literature: as workers spend more time out of the labor force, their productivity declines, reducing the potential quality of their next match. Indeed, in our data, remarriage rates fall rather dramatically with maternal age, consistent with marriageability declining with age. The second possible reason is that welfare receipt is stigmatizing and also reduces marriageability, consistent with Moffitt et al. (1983). Although we cannot empirically test for stigma in our setting, we show that theoretical predictions regarding the impact of welfare on husband quality become ambiguous once stigma is introduced.

To estimate the impact of cash receipt on women's overall welfare, we compare the longevity and household income of accepted and rejected mothers. Longevity is an important determinant of overall lifetime welfare (Jones and Klenow, 2016). We find no significant differences in the overall longevity of accepted and rejected mothers. These results are consistent with the results of Price and Song (2018) who find no effects on adult longevity among participants of a negative income tax experiment. We also find no changes in household income of accepted mothers in 1940 (another important determinant of well-being), at least a decade or more after the mother's application. Thus, fears regarding the negative influence of welfare on mothers do not appear to be borne out in the data: Welfare does not impoverish women in the long run. However, welfare does not lift them out of poverty either. Overall, welfare appears to have little impact (either positive or negative) on mothers in the long run. This contrasts with evidence of long term benefits to children of welfare participation (Aizer et al., 2016; Hoynes et al. 2016).

There is a large literature in economics that investigates the effects of welfare on marriage and fertility (reviewed in detail later). The early work, reviewed by Moffitt (1992), found small effects but was not well identified, typically comparing marriage rates and fertility across states with different benefit levels. The second generation of papers concentrated on the effects of the 1996 welfare reform, which increased work incentives and reduced incentives to remain single and to have more children. This work is better identified, though it shifts the focus away from estimating the effects of receiving welfare to estimating the impact of changes in the design of cash welfare benefits. Welfare reform has been found to reduce marriage rates in some studies (e.g. Low et al., 2018) though not others (Grogger and Bronars, 2001), with little evidence of any impact on fertility (Kearney, 2004). One factor that complicates identification is that welfare reform in the US occurred amidst the backdrop of a strong economy and significant demographic changes (increases in non-martial fertility). Moreover, all these studies focus on short run effects and none consider either lifetime impacts of how welfare might affect the qualities of the new spouse.

We make several contributions to the literature. We are the first to follow a large sample of women over their lifespan. Our results show this matters: For marriage, the negative short-term effects of the program fade significantly over time resulting in small and insignificant lifetime effects. Second, our data also allow us to study the quality of the new match, which has not previously been considered or estimated. Here, our results present a puzzle to be resolved which mirrors the identical puzzle in the UI literature. Third, because our data include the longevity of welfare applicants and family incomes in 1940, we can estimate the impact of welfare on two major determinants of lifetime well-being. Taken together, our results show that welfare receipt did not create perverse incentives as was feared by U.S. policymakers in the early twentieth century. These same concerns played a large role in the dismantling of cash welfare beginning in 1997.[1] Today, these same fears are expressed by policymakers during debates over expansions of non-work and non-means tested transfers such as the Child Tax Credit or Universal Basic Income. Concerns regarding incentives embedded in such program to remain unmarried or have more children are voiced by the

[1] Welfare reform in 1996 significantly reduced the availability of cash assistance in the US and added significant conditions. In 1996, the year before welfare reform, 68% of poor mothers received welfare assistance compared with 22% in 2018. Moreover, the form of assistance has transformed from unconditional cash, to other, mainly non-cash, forms of assistance all of which included conditions.

general public as well.[2] Our results suggest that such concerns may be unfounded, that unconditional cash transfers would likely not generate perverse incentive effects with respect to family structure, nor consign parents to a lifetime of poverty. While family structure has evolved since the early-mid twentieth century, which might raise concerns about applying such lessons to the current setting, the short run evidence on the impact of welfare reform on fertility, from a more recent period, is certainly consistent with our findings (Kearney, 2004).

Finally, we can incorporate the impact of maternal behaviors and outcomes in the evaluation of the MP program by computing the Marginal Value of Public Funds (MVPF) using the methodology of Hendren and Sprung-Keyser (2020). We show that the MVPF is less than 1 when we consider effects on maternal behaviors and outcomes only. However, if there are even modest benefits to the children in terms of longevity or income, as found in Aizer et al. (2016), the program pays for itself. This suggests that the overall evaluation of the program depends crucially on children's outcomes and less so on the outcomes or behaviors of mothers.

3.2 Background on the MP Program and Existing Literature

In this section we explain the structure of the MP program, including a description of eligibility and benefit determination to clarify the incentives of the program with respect to family structure (marriage and fertility). Then, we situate our contributions in the context of the existing literature.

[2] In "Working Class Americans' Views on Family Policy," participants in focus groups expressed differing opinions regarding the role of policy in shaping family structure, with older participants expressing more concern about incentives that reduce marriage and increase fertility than younger participants (Brown, 2021).

3.2.1 Structure and Incentives of the Mothers' Pension Program[3]

In the early twentieth century, widowed or abandoned mothers had few options to earn a living and support their children. Marriage was by far the most common way for these women to address their economic needs.[4] With limited ability to provide for their children, many poor, single mothers were forced to place their children in orphanages, the main form of poverty relief for children provided by local governments (see Skocpol, 1995, p. 425). In response to poor outcomes for institutionalized children, states embraced cash transfers to poor mothers so they could care for their children at home. Illinois was the first state to do so in 1911, and by 1920, most states had followed suit. By 1931, every state in the continental US had an MP program with the exceptions of South Carolina and Georgia. The MP program inspired the eventual implementation of Aid to Dependent Children (ADC) with the passage of the Social Security Act in 1935.

The MP program was funded and administered by individual counties, after states passed the authorizing legislation. There was variation in how MP programs were administered from county to county within a state, and variation in the program's implementation across states. For instance, in most states the program was administered through the county's juvenile court or county clerk's office, but in some states, separate bureaus of child welfare were opened to specifically adjudicate the applications of poor mothers to the MP program.

Eligibility criteria for aid differed across states.[5] Widows, women with husbands in jail or in an asylum, and women with disabled husbands were almost always eligible.[6] However, women who had been deserted or divorced were eligible in some states but not others. Some states required periodic reapplication, while others granted payments until the child turned 14 or 16 years of age. In all states, income and property thresholds were not provided in

[3] We provide a brief description of the program here. More details are provided in Aizer et al. (2016).

[4] In the 1910 Federal Census, the vast majority of white women with children were married (92%) and very few of them worked (4.7%) (Table C.1).

[5] The details for the states we study are given in Appendix Table S1 of Aizer et al. (2016).

[6] In three out of the ten states that we study, only widows were eligible.

explanations of eligibility requirements. Rather, need was determined by local administrators at the time of application who would also determine the amounts granted to each applicant. These amounts were capped at the state level but were otherwise discretionary. The pensions provided about one-third of family income at the time, and the median duration of transfers was about 3 years (Aizer et al., 2016).

Women would apply to the program without knowing the income thresholds for eligibility. They would then undergo an initial review that was usually conducted by a social worker. After the review, a judge or adjudication panel would make a final determination regarding the application and the amount of pension to be granted. The data we have are based on the judge or adjudication panel decisions. That is, our sample consists not of all women who applied, but all women who "passed" the initial eligibility screening and whose eligibility was determined by the judge or panel.[7] Women were denied pensions for many reasons. Most commonly mothers were rejected because their income or wealth levels were deemed to be too high.[8] In Iowa, rejected applicants had a 35 percent higher predicted family income related to accepted applicants prior to receipt of MP.[9] Other reasons for rejection include the following: 1) ineligible (which may also include income in excess of the standard of need); 2) married or husband returns; 3) moved from county; 4) no children eligible; or 5) not a citizen.[10] For one county, Clay County, Minnesota, we have detailed information collected by a nurse who, through home visits, reviewed the living conditions and needs of all MP recipients in the county. The records from the nurse home visits are largely consistent with the above evidence on reason for termination: Families that appeared to have other sources of income were removed from the rolls whereas those that appeared to be in significant need remained on the rolls.

The program design created disincentives to remarry or move residences, as either would

[7] For some counties, not all years are available, but for years in which records are available, we believe we have the universe of records for this second stage of the application process.

[8] See Table 2 of Aizer et al. (2016)

[9] See Aizer et al. (2016), p. 10-11.

[10] See Table 2 in Aizer et al. (2016) which shows the frequency of rejection by reason.

immediately disqualify mothers. Transfers increased with the number of children, creating incentives to have more children. Maximum transfers ranged from \$9 to \$15 per month for the first child and \$4 to \$10 for each additional eligible child depending on the state. Incentives with respect to work were less uniform across the states. This may be in part because maternal work outside the home was relatively rare at the time particularly among white women (Goldin, 2006). In several states (6 out of the 10 that we study), women were required to stay home as a condition of the transfers, since the cash transfer was given in exchange for looking after the children. Other states limited the hours women could work; still others enacted a 100% benefit reduction rate on earnings. More generally, working women were by definition less likely to be deemed eligible since they had a source of income.[11] Given the variation across programs, for our analysis we include county of application FE to address any heterogeneity across counties in the administration of the program.

3.2.2 Existing Literature on Cash Transfers and Maternal Behavior

Moffitt (1992) reviews the existing research on the theoretical underpinnings and empirical evidence regarding the incentive effects of the US welfare system with respect to labor supply, family structure, and migration. With respect to family structure, because welfare benefits have historically been paid only to single mothers with dependent children, Moffitt writes, "the program provides an obvious incentive to delay marriage, increase rates of marital dissolution, delay remarriage and have children outside of a marital union" (page 27). Empirically, women on welfare are indeed less likely to marry and have more children (Hutchens, 1979; Teitler et al., 2009).

[11] While the MP program has many similarities to modern day welfare, there are important differences. Both are means-tested programs that offer unconstrained, but limited, cash transfers. The MP program terminated eligibility upon remarriage (to any man), creating strong disincentives to remarry. The modern-day welfare program terminates benefits upon marriage or cohabitation with the child's father, not necessarily any man. The MP program discouraged work — several states required women to stay home as a condition for the transfer, although some regulated the amount of work or simply lowered the transfers when women brought income home. This continued to be the case in most states until the 1996 welfare reform which capped lifetime benefits and required recipients to work.

However, there is little evidence that these effects are causal. Women on welfare differ in important ways from women who are not on welfare. These differences may explain their lower marriage rates and their higher fertility rates. In fact, though the cross-sectional comparisons across states suggest a positive relationship between welfare generosity and single motherhood, the time series evidence does not. While welfare benefit levels were increasing between 1960 and 1976, so were rates of single motherhood. However, when benefit levels started to fall from 1976 to 1984, the share of single-parent headed families continued to rise, inconsistent with welfare benefits lowering marriage rates. More detailed analyses based on comparisons within states over time confirm this finding: Changes in benefit levels within a state are not accompanied by changes in single motherhood (see Moffitt, 1992).

More recent work has focused on the effects of welfare reform efforts, namely the use of sanctions, time limits, and work requirements, on maternal behavior. This literature has found larger impacts of welfare on family structure. Moffitt et al. (2015) find that welfare reforms did increase the probability of cohabitation with a biological father, but not other males. Bitler et al. (2004), Bitler et al. (2006), and Low et al. (2018) find that time limits on beneficiaries imposed by the 1996 welfare reform reduced divorce rates and increased the likelihood that children live with unmarried parents. Kearney (2004) finds that welfare reform efforts that reduced financial incentives to have more children did not increase fertility. Work estimating the impact of welfare reform on intergenerational correlations in welfare receipt suggest that welfare reform did reduce daughter's reliance on AFDC/TANF, but not other safety net programs such as SNAP or SSI (Hartley et al., 2022.)

There is a related literature investigating the effects of other redistributive programs on marriage. The earned income tax credit (EITC) also contains disincentives to marry (Hotz and Scholz, 2007), but empirical work has found these effects to be economically small or insignificant (Dickert-Conlin and Houser, 2002, Herbst, 2011, Michelmore, 2016).

Several papers investigate how the marital status requirements embedded in women's eligibility for pensions upon the death of their husbands affect remarriage. These papers find larger effects on marriage rates (Salisbury, 2017; Brien et al., 2004; Persson, 2017). The subjects in these studies however tend to be older and richer than the average welfare-recipient. As a result, the opportunities for and benefits of remarriage may be lower.[12]

We make several contributions to this literature. First, we use a credible identification strategy at the individual level to investigate the effect of welfare receipt on remarriage decisions, examining the extensive margin where one might expect to see larger effects. Second, we follow women over their lifetime and establish not only whether they marry, but when they marry and who they marry, as well as whether they have more children. Third, we estimate the impact of welfare on a lifetime summary measure of well-being: longevity. Finally, we calculate the MVPF of the MP program taking into account both the behavioral impacts on mothers and the benefits of cash transfers to children.

3.3 Model of Welfare Receipt and Search in the Marriage Market

We adapt the canonical model of search in the labor market with unemployment insurance, first developed by McCall (1970), to model search in the marriage market with cash transfers.

In McCall's original model, an unemployed worker searches for employment. Offers of employment vary in quality, as measured by the wage, with a known distribution. Unemployed workers receive offers, which arrive at a given rate, and accept an offer if the offered wage exceeds the worker's reservation wage. If the worker rejects the offer and remains unemployed, they retain the option of waiting for another potentially better offer in the next period. In this model, unemployment insurance increases the value of remaining unemployed, thereby increasing the reservation wage. The model yields two predictions. First, workers with unemployment insurance will remain unemployed for longer than those without. Second,

[12] Dillinder (2016) also considers effects of Social Security receipt. Additionally, Fox (2017) investigates the effect of tax incentives on marriage as do Whittington and Alm (1997) and Fitzgerald and Ribar (2004).

when workers with unemployment insurance do accept an offer, the wage will be higher.[13]

We adapt this model to the marriage market where women are searching for husbands and offers of marriage arrive at an expected rate. Like offers of employment, offers of marriage also vary in quality. Cash transfers (welfare) have the same effects on the marriage market that unemployment insurance has in the labor market: It increases the woman's outside option and therefore the "reservation quality of the match," extending her duration of search (the time to marriage), and resulting in a higher quality husband when she does remarry. After describing the model, we discuss how to test its predictions with respect to the quality of the spouse in the data.

3.3.1 A Basic Model of Search in the Marriage Market

A single woman must decide every period whether to marry or to stay single. If she stays single, she has the option to marry the next period. If she marries, she will stay married forever.[14] Her patience level is given by her discount rate β. She searches for partners, and prospects arrive at a Poisson rate λ. Each prospect has a value of q, which summarizes his quality as a husband. This value has an unknown distribution in the population, $q \sim F(q)$ with support $\left[\underline{q}, \bar{q}\right]$ and $\bar{q} > b$. While she is single she receives a cash transfer of value b every period, but this transfer is lost upon remarriage.

The value of being single is given by

$$V_s = b + \beta\big(\lambda \int_{q=\underline{q}}^{\bar{q}} \max\left\{V_m(q), V_s\right\} \, \mathrm{d}F(q) + (1-\lambda)V_s\big).$$

and the value of being married to prospect q is given by:

$$V_m(q) = q + \beta V_m(q) = \frac{q}{1-\beta}.$$

[13] Other features have since been added to this model, such as simultaneous offers (Burdett and Judd, 1983).

[14] This is a simplifying assumption, but it is well supported by the data. Most women in our sample marry only once (only 5.6% married more than once after the transfer).

In this set-up, the agent accepts an offer to marry prospect q if $V_m > V_s$. Since the value of marriage is strictly increasing in q, the agent will follow a cut-off rule. There is a q^* such that she will accept all prospects with $q > q^*$. The cut-off rule is implicitly defined as

$$V_m(q^*) = V_s.$$

Considering that, and rearranging the definition of V_s, we can write

$$(1-\beta)V_s = b + \frac{\beta\lambda}{1-\beta}\int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\, \mathrm{d}q,$$

This function is continuous and positive at $q^* = b$ and negative at $q^* = \bar{q}$, so there exists a solution, and because it is strictly decreasing, the solution is unique. Intuitively, this equation states that the value of the minimum acceptable marriage, q^*, should be equal to the benefit, b, plus the option value of holding out for a good match. Given a reservation quality q^*, the probability of marriage is $\lambda\,(1 - F(q^*))$ and the average match quality is $\mathbb{E}[q|q > q^*]$. The duration until marriage is given by $D = 1/\lambda\,(1 - F(q^*))$. Duration is decreasing in the arrival rate and increasing in reservation quality.

3.3.2 Model predictions and testable implications

The following propositions are derived from this model. All proofs are provided in the Appendix.

Proposition 1. $\partial D/\partial b > 0$ and $\partial \mathbb{E}[q|q > q^*]/\partial b > 0$: *An increase in benefits b increases the number of periods the woman stays single and the average quality of the marriage.*

It is straightforward to test whether receiving a transfer leads to longer durations until re-marriage. Testing whether the quality of the match increases among those who marry is more difficult because there is no single indicator of the quality of a match. Suppose instead that there are many traits X that matter but that prospects can be ranked using a single index function $q(X)$ as in Becker (1973). If this function is known, then we can test the predictions in Proposition 1 by constructing this index function. Alternatively, if the function is not known, then we can investigate how transfers affect each trait X. The following proposition holds under the assumption that q is increasing in all its arguments X:

Proposition 2. *Without further assumptions about the joint distribution of X and the production function $q(X)$, the sign of $\partial\mathbb{E}[x_i|q > q^*]/\partial b$ is ambiguous for all i. However, the sign of $\partial\mathbb{E}[x_i|q > q^*, x_{-i}]/\partial b > 0$ for all x_i so long as all relevant X are observed.*

This proposition states that the theory does not provide any guidance about the effect of transfers on any one "input" into quality without knowing their joint distribution and how women trade-off these characteristics.[15] But the proposition also states that conditioning on one measure of quality, the other measure of quality will unambiguously increase with an increase in the transfer. If both measures of quality are observed, we can test this empirically by conditioning on one trait and estimating the impact of the transfer on another trait.

3.4 Data

3.4.1 Data Collection

Administrative data on MP applicants were collected directly from state and county archives in 14 states, 10 of which included rejected applicants in their records.[16] We limit the sample to mothers from the 10 states with rejected applicants, and to those who applied before 1930,

[15] In fact, we might observe that the average quality for any one trait (or for all traits) might decrease with the transfer even though the actual match is better. For example, consider a quality function $q(x_1, x_2) = x_1 x_2$. The joint distribution of the traits is uniformly distributed over three mass points $(1, 10); (10, 1); (4, 4)$. Suppose that, initially, the cutoff is $q^* = 10$. The average of each trait conditional on a match is equal to 5. Consider a small increase in the cutoff $(10 < q^* \leq 16)$. The new average of each trait is 4, lower than in the original situation, and suggesting that the average quality has gone down. However the quality of the match after the cutoff increase is 16, higher than the average quality before the cutoff increase.

[16] We study 10 states with early programs (dates of passage in parentheses) for which we obtained data: North Dakota (1915), Idaho (1913), Illinois (1911), Iowa (1913), Minnesota (1913), Ohio (1913), Oklahoma (1915), Oregon (1913), Washington (1913) and Wisconsin (1913). See Aizer et al. (2016) for details.

when most MP programs lost funding.[17] To track MP mothers and their children, we match these administrative data to family tree data available on FamilySearch.org, which includes more than 1.2 billion people.[18] The mother's name, combined with the names and dates of birth of her children, enables us to locate the mother on a family tree. Once a mother has been found, we observe her maiden name, her date of birth, her date of death, and the names, dates of birth and dates of death of all her husbands and children. For all women in our sample, we employed researchers at the BYU Linking Lab to search for any evidence that she married after the MP program using information in the trees. This strategy of using families to create matches was pioneered by Joseph Price at the BYU Lab (see Price et al., 2019).[19]

In addition, we had the BYU Lab researchers hand match all other records available on Ancestry.com and FamilySearch.org (e.g. the Social Security Death Master File, other state death records, cemetery records, birth certificates and marriage certificates). Therefore, we do not only rely on the family trees that were available. We improve on strategies in the automated linking literature as well as the tree-matching literature by individually searching for each mother's marriage information in available records on the genealogy sites. Finally, research assistants also manually linked mothers and their new husbands to 1910, 1920, 1930 and 1940 Census Records if these links are not already made in the family tree. We observe several measures of the characteristics of the new husband: his education, longevity, age, and occupation, as reported in various census years. We describe our matching methodology in more detail in the Online Appendix.

[17] We also drop a small number of mothers who applied multiple times and those who did not appear to be mothers (grandmothers, sisters and step-mothers). Sometimes a woman appears more than once in our records. In this case, we kept a single record using the following rules: (i) Keep only the observations of the first successful attempt. (ii) If applied successfully more than once the same year, keep the application with more children listed. (iii) Keep the smallest family ID if applied successfully more than once the same year, with the same number of children.

[18] Recent research (Kaplanis et al., 2018) suggests that data from the trees are quite accurate when validated using genetic information. The information also appears to be roughly representative of the population, as life expectancy and other summary measures derived from the trees reproduce population patterns.

[19] While our paper represents the first example of implementation of the technique of linking mothers, more recently others have followed suit. These include working papers by Craig et al. (2021), Marchingiglio and Poyker (2021) and Withrow (2021).

The resulting dataset allows us to determine if a woman in our MP sample ever remarried, the duration until the marriage and the characteristics of her new husband. They also allow us to track all her children (and when they were born) as well as her own longevity. Thus, we have lifetime measures of marriage, fertility and maternal longevity, as well as the quality of her new husband. To our knowledge, this type of data has never been collected for a sample of welfare recipients. We can also observe employment and occupation in each census year (1920, 1930 and 1940) and income in 1940. These labor market measures, in contrast to our measures of family structure, are only spot measures. Because these data are more limited (and less novel), they are not the main focus of our work.

3.4.2 Summary Statistics

Our sample includes 16,228 applicants in 132 counties across 10 states. Summary statistics are presented in Table C.2 for the full sample, and for the subsample of unmarried women at the time of MP application (13,383 mothers, or 82% of the full sample). About 53% of the applicants were widowed at the time of application and about 21% were married.[20] The husbands of married mothers were either disabled or in jail, mental institutions or sanatoriums. Very few (3%) were divorced. About 10% of the applicants are rejected. The average woman in our sample was 37 years old at the time she applied and listed 2.6 children under the age of 14 in the application. About 98% are white, and 17% are foreign born.[21]

Forty-eight percent of unmarried MP mothers eventually remarried, and they waited an average of 6.4 years to do so. Only 15% of all unmarried mothers married within 3 years of applying for welfare. When they remarried, they married men who lived almost as long as they did (71 years for men and 74 years for women) but who were less educated than them on average (the education gap is -0.23 years). Post-application fertility was low with only

[20] The rest do not have marital status, in many cases this is because only widows are eligible.

[21] We have data on the duration of the transfer or reason for termination for only a small subset of the sample – for this reason they are not included here. Therefore, we cannot perform "common" tests in the UI literature such as testing whether people marry just before the end of the transfer.

0.26 children born on average after applying for welfare, already suggesting that any fertility effects are likely to be small.

The information on maternal work, income and location comes from decennial census data so we cannot observe the entire history of employment, income and location. Only 12% of MP mothers were in the labor force in 1910. Women's labor force participation remained low despite their high poverty rates: rising to a max of 37% in 1930 and falling to 26% by 1940. Women's wages and occupational scores were low, as were their incomes (Figure C.1).

3.4.3 Data Quality and Limitations

Historical administrative data have several advantages for this analysis. They allow a long follow-up period and have lower attrition than modern survey data. We discuss these aspects now.

Data Quality. Of the sample of 13,383 mothers who were unmarried at the time of application, we found remarriage data for 84% of the sample. Among those who remarry (5,435), we have the exact date of marriage for 70% of the sample. With respect to measures of new husband quality, we measure longevity for the entire sample, but for other measures such as his wage income from the 1940 census, we find only 52% (see Table C.2). For the mothers, we determined maternal longevity for 80% of the sample and found maternal education for 85% of those who were alive in 1940.[22]

These match rates compare favorably with recent work using US census data from the early part of the twentieth century which hover around 10 to 30%.[23] These rates are also higher than follow-up rates in modern survey data tracking women on welfare. For example,

[22] There are several reasons we might not find a person. Many of our outcomes come from censuses, which undercount the population particularly in the past (Hogan and Robinson, 1993). We might also fail to find them due to spelling errors or other inaccuracies in the data. Finally birth, marriage and death records are not always available.

[23] For example, Abramitzky et al. (2014) estimating the impact of migration on earnings trajectories achieve match rates of 16% for the native born and 12% for foreign born men. On the higher end, the Life-M Project matches about 30% of birth certificates to death certificates in the states of Ohio and North Carolina (Bailey et al., 2022).

the follow up rate in the SIPP is about 63% over 12 waves/years (Zabel, 1998). In the PSID, the follow-up rate for mothers collecting welfare is lower than 40% over 35 years.

All of our data were hand-matched across multiple sources and all data entry were double checked. A validation exercise showed the accuracy of the matches to the tree, the death certificates and the 1940 census to be very high (above 97% in all three cases). We discuss strategies to address missing data and data quality below.

Limitations. We are unable to generalize our results to African American mothers as they accounted for only 1.3% of the population in the counties we study and 2% of applicants in our data.[24] Because of the small number of women who were rejected (only 10% of the sample), we cannot conduct heterogeneity analysis with any precision, though we do present results in an appendix and investigate it using random forest approaches. Last, as previously mentioned, the data on women's labor market outcomes are limited.

3.5 Empirical Strategy and Identification

3.5.1 Empirical Strategy

We test the model's predictions using the following equation:

$$y_{ict} = \beta_0 + \beta_1 Accepted_{ict} + \theta X_{ict} + \gamma_c + \gamma_t + \varepsilon_{ict}$$

where y_{ict} is an outcome for woman i applying to the program in county c in year t. *Accepted* is an indicator equal to one if the mother was given a cash transfer and it is equal to zero if she applied for the transfer but was denied after investigation. We also include county and year of application fixed effects γ_c and γ_t in all baseline specifications to account for the fact each county had a different program that varied over time, and to

[24] States and counties with large black populations often did not implement the Mothers' Pension program (Eli et al., 2022), and when they did, they appear to have systematically discriminated against them as many were never deemed eligible (Eli and Salisbury, 2016; Roberts, 1993; Ward, 2009).

account for secular trends in outcomes over this period. We can also include a vector of controls (X_{ict}) that includes the characteristics of the mother and family at the time of application: the number of children, age of the oldest and youngest, her marital status at application (widowed, divorced or missing), maternal age at application, and county-level and state-level time varying covariates.[25][26] We report standard errors clustered at the county level. We also estimate standard errors just correcting for heteroskedasticity or clustering at the county-by-year level. The results are robust to these alternatives. Finally, although our main model is linear for many outcomes, we consider alternative function forms.

Our main coefficient of interest is β_1, which represents the impact of welfare receipt on the outcome. Thus, our strategy consists of comparing the mean outcomes of accepted and rejected mothers who applied in the same county and year and were similar on observables. For rejected mothers to be an appropriate counterfactual, it must be the case that they are not otherwise different than mothers who were accepted, as discussed below.

In addition to estimating standard OLS models with and without covariates, we also estimate the average treatment effect (ATE) of the cash transfer on outcomes using the causal random forests methods recently developed by Wager and Athey (2018) and Athey et al. (2019). This approach has several advantages over OLS. First it is a matching approach which provides consistent estimates of the ATE under the standard assumption of unconfoundedness. Like

[25] A difference-in-differences analysis using variation across counties or states over time in the creation of a MP program cannot be conducted given likely violation of identifying assumptions for the following reasons. First, eligibility and generosity varied considerably across states and counties, complicating our ability to use other states or counties as "counterfactuals" in our specification. Second, we do not know for all counties whether/how quickly after a state authorized MP programs the counties developed their MP programs. It could be that a state authorized an MP program but it took years for most counties to develop their programs. As a result, it's not clear how much of a state is actually treated. Third, we cannot identify likely eligible mothers in counties/states before the MP program from available data (e.g., marital status alone does not determine eligibility). Given this, the strongest identification strategy involves comparison of outcomes for mothers who applied in the same county, under the same eligibility rules.

[26] County controls include: sex ratio (M/F) aged 18-55, share females in the labor force aged 18-55, share Black aged 18-55, share rural aged 18-55. County controls match linear interpolated information from the 1910, 1920 and 1930 census with the year of application to the program. State-varying controls include: manufacturing wages, education/labor laws (age must enter school, age can obtain a work permit, and whether a continuation school law is in place), state expenditures in logs (education, charity, and total expenditure in social programs), state laws concerning MP transfers (work required, reapplication required, the maximum legislated amount for the first child, and the legislated amount for each additional child).

other matching estimators it assumes that untreated observations with similar propensities to be treated as treated observations provide appropriate counterfactuals. This method leverages machine learning, specifically random forests, to find the "nearest neighbors" and computes treatment effects for each treated unit using these untreated but similar observations.[27]

Since we obtain individual level treatment effects, this method allows to investigate if treatment effects are heterogeneous, a second major advantage of this alternative approach. Sloczynski (2022) shows that in the presence of heterogeneity, OLS estimates a weighted average of the treatment effects across groups, where larger groups get smaller (rather than larger) weights. In our case the rejected group is substantially smaller than the treated group, and thus if the treatment effect is heterogeneous, it is possible that the OLS differs from the ATE, the Average Treatment on the Treated (ATT) and the Average Treatment on the Untreated (ATU). The random forest approach allows us to investigate this possibility without imposing any specific functional form in the estimation of the propensity score (and without imposing linearity in the treatment as OLS does). We report the ATE and the ATT that results from this approach. Details of the implementation of this procedure are in the Appendix.

3.5.2 Identification: Comparing Accepted and Rejected Mothers

Three pieces of evidence presented in Aizer et al. (2016) showed that rejected mothers were slightly better off. We summarize these here and also offer additional evidence based on our new data.

First, investigating the basis for rejection (when available), we found the most common reason (35%) was "other means of support," suggesting rejected mothers had greater incomes. Second, comparing accepted and rejected mothers, we found that the rejected had on average fewer children and that their children were older. We used these characteristics and marital status to predict family income using the 1915 Iowa State Census—the only income data

[27] In this method the nearest neighbors are the observations that fall under the same leaves of a given tree.

available in the US prior to 1940. Women who were rejected from the program have higher predicted income than those who were accepted, consistent with the evidence on reasons for rejection. A third piece of evidence comes from a comparison of the *pre-application* characteristics of accepted and rejected mothers whom we can find in either the Iowa State Census of 1915 (for the Iowa sample of mothers) or in the 1900-1920 US Federal Census for the Ohio sample of mothers.[28] In both cases, we find that for the majority of the variables we observe, accepted applicants were worse off (Aizer et al., 2016).

We use our newly collected data to further assess the pre-determined differences between the two groups. Specifically, we now have information on the mother's educational attainment (from 1940 census records), her date of birth, place of birth, race and ethnicity, the longevity of her first husband, and information on all her children, including those who died prior to applying for the pension, and those who were too old to be eligible (and were therefore not listed in the MP records) but could potentially provide income or other resources to their mothers. We also observe the number of siblings the mother had who could also serve as alternative means of support.

We continue to find that rejected mothers were slightly better off than accepted mothers when comparing them on these newly collected predetermined characteristics (Table C.3).[29] To assess the magnitude of the observed differences between accepted and rejected mothers, we repeat our previous analysis and predict maternal income again but include these newly collected measures. Accepted mothers are more likely to be at the lower end of the distribution of predicted income (Figure 3.1), but these differences are modest. The predicted income of accepted mothers is about 50 dollars (6 percent) lower than that of rejected mothers (Table C.3). Thus with the newly collected data on mothers, we confirm our previous findings

[28] We focused on Ohio because a large portion of our records come from Ohio.

[29] Controlling for county and year of application fixed effects, accepted mothers had more children who died before the application (which is significant for the sample of unmarried mothers) and fewer children over the age of 14. They were also younger, and had husbands who died more recently and at a younger age. All other differences (number of siblings, race, foreign born status, work and occupation in 1910 or education levels in 1940) are not statistically significant in the full sample or in the sample of unmarried mothers.

that, on average, accepted mothers appear to be slightly poorer than rejected mothers.[30]

Based on this finding, we may be biased towards finding more harmful effects of welfare on maternal outcomes. For example, this slightly negative selection into MP receipt would likely bias downwards any positive impact of cash transfers on maternal longevity, and lead to overestimates of the impact of welfare receipt on marriage delay and fertility. We conduct two exercises to assess the extent of omitted variable bias. First, we report bounds for β_1 using the Oster (2019) proposed correction to assess the extent to which our assumptions about unobservables affect the coefficient estimates.[31] Second we estimate causal random forest treatment effects, which as explained above flexibly account for observables to construct counterfactual groups, in the spirit of matching methods.

3.5.3 Assessing the Impact of Missing or Low Quality Data on our Estimates

Missing data. Although attrition in our data is low, missing data can bias our results if the data are missing differentially for accepted and rejected mothers. We investigate whether accepted mothers are differentially missing outcomes by regressing an indicator for missing on the indicator for accepted (Table C.4, Panel A). We find no differential attrition in our data for all outcomes related to family structure (remarriage, duration, husband quality and fertility).[32]

We do however find evidence of differential attrition for our labor market outcomes in

[30] The mean predicted income of the accepted and rejected groups using the Iowa samples are both higher than in Aizer et al. (2016). The main reason is that we can now observe the age of the mother and use this age in the prediction. This results in significantly higher predicted family incomes. We have predicted incomes using many different specifications and control variables and we find very similar results across all of these: Although the means vary, the accepted group is always slightly poorer than the rejected group.

[31] To compute these bounds we assume that the R-max is 1.3 times greater than the R-squared that is estimated in the regression with controls, as suggested by Oster. We assume that δ = (-1, 1) for lower and upper bounds to capture that the omitted variables are positively or negatively correlated with the regressor of interest.

[32] Accepted status predicts only one marriage related outcome at the 5% level (whether the new husband's age at death is observed).

1930 and 1940. Labor force participation, occupation scores and family income in 1940 are all less likely to be missing for accepted mothers (Table C.4, Panel B). Conditional on controls, the differences are about 10%. The same is true for location and family income in 1940 (Table C.4, Panel B).

To address this issue we take two approaches. First, we estimate OLS models that account for attrition using the semi-parametric two-stage approach proposed by Newey (2009). In the first stage we predict attrition, including a predictor (an instrument) that is not part of the main equation of interest. Our instrument for selection is research assistant (RA) finding rates. RAs are assigned arbitrarily to the mothers in our data. RA quality affects the likelihood of finding a match. Thus differences in finding rates reflect RA ability rather than underlying likelihood that the record can be matched based on observables. In the second stage, we estimate a linear regression of the outcome on controls and a fourth degree polynomial of predicted values from the first stage, i.e. a semi-parametrical selection correction term. Second we estimate Inverse Propensity Weight (IPW) OLS models that use the estimated probability of a match as an (inverse) weight in the regressions, as recommended by Bailey et al. (2020) when matching historical data sets. We report both of these alternative estimates in the tables.[33]

Mismatched data. There is considerable debate among economic historians regarding the quality of linked data and how it varies based on various matching methods (Bailey et al., 2017; Abramitzky et al., 2019). We test whether the quality of the match influences our results. To do this, we compute measures of the quality of matches and re-estimate results using only high quality matches.[34] We also present results using data from multiple sources – for example we can compare our marriage information from the trees to the information that

[33] We do not implement these robustness checks for the random forest estimators for which these adjustments have not been developed. Since we find little evidence of heterogeneity (see results section), we view the OLS adjustments to be informative.

[34] A high quality match is a match with quality above the median. The quality measure is a weighted sum of Jaro-Winkler distance assessing the similarity of the name, place of birth and age match between the different datasets. The data codebook details how we compute each quality measure.

is derived from the census. If the results are similar across different data sets, this reduces concerns that matches to one source of information may be incorrect.

3.6 The Effects of Welfare on Marriage and Fertility

3.6.1 How does Welfare Affect Marriage Decisions?

Unmarried mothers on welfare are not less likely to remarry over their lifetime (Table 3.1, Column 1). Accepted mothers are slightly (1.4 percentage points) less likely to remarry than rejected mothers (conditional on controls), but the difference is not statistically significant and it is small relative to the average remarriage rate for rejected mothers (47 percent). This effect is not sensitive to how we estimate the standard errors, correct for missing data or whether drop the lowest quality matches. The causal random forest ATE and ATT are somewhat more negative (-0.02 and -0.026 respectively) but they are also statistically insignificant and small in magnitude. Interestingly the ATE, ATT and OLS are similar – in fact we cannot reject the null that they are the same, i.e. that there is no heterogeneity in the treatment effects.[35]

How large are these effects? Using the largest Oster bound, being accepted lowered the probability of remarriage by 0.02 percentage points, an economically small effect. If we use the confidence interval from our main OLS specification or the one from the random forest, we can reject declines in marriage rates larger than 10%. If we compare the estimated impact of MP receipt on remarriage to the impact of age at MP application, we find that the impact of MP receipt on marriage is roughly equivalent to a one year increase in maternal age.

Next we investigate the impact of welfare receipt on duration until remarriage. A histogram of the duration to remarriage suggests that rejected mothers were more likely to marry soon (within two years) after applying (Figure 3.2, Panel A). Kaplan-Meier survival estimates of

[35] A test that OLS is equal to ATE cannot be rejected. We also test for heterogeneity as suggest by Athey et al. (2019) and find no evidence of significant heterogeneity.

the probability of remaining single, where the clock starts the day of the MP receipt and ends at death, show a similar pattern: accepted mothers remain single for longer and are more likely to remarry later (Figure 3.2, Panel B). While women on welfare are not less likely to ever remarry, they wait longer to do so.[36]

How much longer? A regression of time to remarriage on accepted status suggests 1.3 years longer, which is identical to the causal random forest ATE and ATT estimates (Table 3.1, Column 2). The coefficient is similar with the Oster bound (1.4) but smaller (0.9) if we drop low quality matches or use IPW. Relative to the duration of 5.47 years to remarriage for rejected mothers, this represents an increase of 20-24 percent relative to the mean. Estimates from an Accelerated Failure Time model (AFT), using the log of the duration as the outcome, are very similar around 24% (column 3).

To further explore timing, we estimate regressions where the dependent variable is whether the mother remarries within a year, two years, five years, etc. For these regressions, mothers who did not ever remarry are coded as zero. Mothers whose marital status could not be defined, or who are missing marriage dates are excluded. We find a marginally significant effect of receiving welfare on short durations but no significant differences on longer durations, consistent with Figure 3.2 (Table 3.1, last 5 columns). The coefficient estimate suggests that remarriage within one year is 2.4 percentage points lower for mothers on welfare. Because the baseline is low in the first year (0.04), the relative effects are large: Welfare receipt lowers the likelihood of remarriage by 60% within a year. This falls to 15% within 5 years and is small and insignificant after five years.[37] The year-by-year estimates are presented in Figure 3.2, Panel C and show that the effects, as a percentage of the baseline, are large but decline and become insignificant. The short run effects are larger if we drop low quality matches, but lower if we use the IPW. Overall they are still small and insignificant after the fifth year. The

[36] We corroborate these findings using another source of data on marriage in the Census. While there are no differences in marriage rates in 1930 or 1940, there is a statistically significant 25% decline in the likelihood of being married in 1920 (See Table C.5, Columns 2-4). These results also suggest that cash transfers increased the duration until marriage in the short run, but not the in medium or long run.

[37] We also estimate Logit models. The results are very similar to those reported here.

ATE and ATT estimates are somewhat larger but they also decline magnitude relative to the mean. By the 10th year the effect is about a 10% decline in the probability of remarriage.

In sum, duration to remarriage increases between 0.9 and 1.4 years with welfare receipt. This effect is accounted for by short run behavior: women are less likely to remarry but only in the short run. After five years, there are no large differences in marriage rates. Over the lifetime we can reject declines in marriage rates greater than 10%. Our results are consistent with previous research finding of immediate effects of welfare reform on remarriage (Low et al., 2018), but we are the first to show that over a longer follow-up period, the difference falls to zero. Overall, we conclude that the effects of welfare on marriage are modest, and not as large as short-term estimates imply.

3.6.2 Who Do Mothers on Welfare Remarry?

Were these marriage delays associated with increases in the quality of the husband and match as theory predicts? In this section, we describe how we construct our measures of husband and match quality. We follow this with an analysis of whether waiting does increase quality and conclude with an analysis of whether welfare receipt, which leads to delays in remarriage, results in a higher quality husband or match.

Measuring Husband and Match Quality

We calculate five measures of the quality of the new match: three characteristics of the husband and two of the match. The former includes his longevity, his education and his predicted income based on occupation score. Longevity is an excellent measure of health and also an indirect measure of his lifetime resources, as it partly reflects the socioeconomic conditions he experienced as a child and as an adult.[38] Education is a good predictor of

[38] Many papers document that conditions in utero affect health and longevity (for a review see Almond and Currie, 2011). Another extensive literature shows that individuals nutrition as well as their parents' income and education while growing up predict health (Case et al., 2002; Hayward and Gorman, 2004, see Almond et al., 2017 for a review). Finally, socio-economic status (education, occupation and income) in adulthood are very large predictors of longevity (Cutler et al., 2006; Chetty et al., 2016).

permanent income and is also associated with marital stability (Lundberg et al., 2016), but it can only be observed in the 1940 Census and therefore not observed for all.[39] Finally, we predict the husband's lifetime income (in 1950 dollars) using the latest pre-marriage occupation observed in census data.[40]

We construct two measures of the quality of the match: the age and education gaps between spouses. We assume that the optimal age gap is 2.5 years based on previous work.[41] For the second measure of match quality, the education gap, a more equal distribution is preferred (Doss, 2013; Hitsch et al., 2010). We can only compute the latter for couples in which neither has died prior to 1940.

Finally, we combine these measures of husband and match quality into a single index, using two methods. In the first, we standardize all the measures and sum them, giving each equal weight.[42] In the second, we combine them into an index using the model calibrated by Grow and Van Bavel (2015) which is based on marriage patterns in contemporary Europe.[43] This index corresponds to the utility associated with a given match, which is a function of both the woman's and the man's traits.

[39] Because 18% of remarried husbands died prior to 1940, it is not observed for all men.

[40] We use the IPUMS constructed "occscore." This measure assigns income to individuals based on their occupation, imputing income in that occupation in 1950. We assign each man the occupation score we observe in the latest census where he is observed before marriage under the assumption that this is the most likely occupation that the MP woman would have observed at the time of her marriage decision.

[41] Empirically, small age gaps predict greater satisfaction (Lee and McKinnish, 2018) and lower divorce rates (Lillard et al., 1995), and they are preferred in online dating (Hitsch et al., 2010). The optimal gap of 2.5 is based on work by Grow and Van Bavel (2015).

[42] To do this, we first normalize each measure (subtracting the mean and dividing by the standard deviation) and then sum them together as in Kling et al. (2007). To maximize sample size we use any measure available, so the index is defined for those that have any measures.

[43] We use the utility function and the parameters defined and calibrated in Grow and Van Bavel (2015). The index is given by $v_{ij} = \left(\frac{S_{max} - |si - sj|}{S_{max}}\right)^{w_s} \left(\frac{y_i}{Y_{max}}\right)^{w_y} \left(\frac{A_{max} - |\alpha_i - \alpha_j|}{A_{max}}\right)^{w_a}$. The first term of the equation is the similarity of education, the second term is the earnings prospects and, the last term is the age gap. We follow the same categorization of variables as in the original paper, except for education, where we divide it in 4 quartile categories instead of the four categories in the paper (no schooling, primary, secondary and tertiary). The calibration parameters are given by $S_{max} = 4; Ymax = 5; A_{max} = 800; w_s = 0.385; w_y = 1.201; w_a = 10.833$. Note that this is **not** a sorting index like those used in Becker's assortative matching models.

Duration to Remarriage and the Quality of the Husband and Match

The basic model predicts that if women delay marriage they will marry more desirable husbands. Duration to remarriage is indeed positively and statistically significantly related to husbands' education, occupation, and longevity; duration is also statistically significantly associated with smaller education gaps and age gaps (Figure 3.3). To our knowledge this is the first paper documenting that there is a strong correlation between waiting to marry and the quality of the husband.

Welfare Receipt and Quality of the Husband and Match

Comparing the estimated densities of the quality measures for accepted and rejected mothers does not support the prediction that welfare receipt improves the quality of the husband or the quality of the match (Figure 3.4). The new husbands of welfare recipients do appear to live a bit longer, but they are not more educated or likely to be employed in higher paying occupations. The distribution of match quality (age and education gaps) is also very similar for both groups. We cannot reject the null that the distributions of any trait are identical for accepted and rejected.

Regression analysis yields similar findings. The results (Table 3.2, Panel A) suggest that mothers on welfare marry husbands who are roughly similar: Except for longevity, all the coefficients for accepted are statistically insignificant. The results without covariates are very similar (Table C.6). While some point estimates are positive (longevity), several are negative (predicted income and education). Estimates of the impact of welfare receipt on match quality (age and education gaps) are also insignificant and often of different signs. A joint test (Column 6) shows that we cannot reject the null that all coefficients are equal to zero at the 5 percent level.

Using the index based on equal weights (Table 3.2, Panel A, columns 7 and 8), we find a positive and significant effect of welfare receipt on husband and match quality, but this result is mostly driven by the positive impact on longevity and it is small, on the order of 10 percent of a standard deviation in the index. Using the index based on Grow and Van Bavel (2015), the coefficient is small and insignificant (Table 3.2, Panel A, Column 9). The results using causal random forests are very similar and suggest no overall effect on husband quality (Panel D).

However recall that this test may not be informative since the theory is ambiguous about the effect of transfers on any given trait and there is uncertainty about how to combine the traits into a single index. To address this we repeat the analysis but controlling for other husband traits, as proposition 2 of the model (Table 3.2, Panel B) suggests. The results are roughly similar and do not unambiguously point to an increase in quality, except for longevity. None of these results are affected by Oster corrections, corrections for missing data or quality of the data (Table C.6). We worried in particular about our use of occupation as a means to assess income since the mapping between occupation and wages/income varied overtime. Our results are very similar if we use two alternative measures of occupation-based income computed by Olivetti and Paserman (2015).[44]

We also rule out that the transfers affected assortative mating (Figure C.2). More educated women were more likely to marry more educated men as they do today. However, this is equally true among both accepted and rejected women. A final test of the hypothesis that quality of the match increased is to examine whether husband and wife live together in 1930 or 1940: these are indicators that the marriage was long-lasting and therefore a good match. We find that accepted mothers are less likely to be living with their spouses in 1930 and in 1940, suggesting that if anything these matches are of worse, not better quality (columns 9 and 10 of Table C.7). We conclude that the transfers did not meaningfully improve the quality of the matches.

[44] In Table C.6, we show results for these alternative occupation measures and also show additional specifications for the results. In Table C.7 we show results for several other traits of the new husband (1940 income or earnings score, foreign born status, farming status and number of children). The coefficients on accepted are never statistically significant and vary in their sign. In Table C.10, we show heterogeneity in results.

3.6.3 Why Does the Theory Fail?

We consider five possibilities. First, it may be that the attractiveness of women declines with age, just as the human capital of workers declines when they are unemployed. If so, waiting to marry a higher quality husband would result in a depreciation of the mother's own quality or attractiveness (her age and fertility). Figure C.3 shows that, as in other settings, women are much less likely to marry as they age. Theory (proposition 3 in the Appendix) suggests this should not affect the predictions of the effects of the cash transfer–those who receive the transfer should still find better men. But it does suggest that the effects might be small if waiting to marry reduces her attractiveness. If we control for the age at marriage of the mother, our conclusions are unchanged (Table 3.2, Panel C).

A second possibility is that there is negative selection into marriage among those who delay (the "best" women marry first). There is little evidence of this once age is accounted for. There are no predetermined characteristics that predict duration to remarriage, aside from her age and number of children (Table C.8), suggesting that negative selection likely does not explain this.

A third possibility is that stigma associated with welfare receipt reduces the quality of the husband. Proposition 6 in the Appendix states that if transfers lower the rate of arrival of prospects or worsens their quality, then the predictions of the model become ambiguous.[45] Thus once stigma is included in the model, the predictions with respect to partner quality can reverse, even if duration is increasing.[46] We cannot provide empirical evidence for stigma but historical accounts suggest that there has always been strong stigma from receiving charitable help, from private or public institutions, and this was also true during the period we study

[45] Cutoff quality moves in the same direction as the benefits, the change in the probability of proposals, λ, and the distribution of quality, $F(q)$. With stigma, the program increases b but lowers λ or the distribution of quality. The original effect increases the cutoff but the stigma effect lowers it. It is unclear which one we should expect to dominate.

[46] The predictions of the model with respect to quality are still ambiguous even though duration increases. This is because a duration increase is to be expected even if quality doesn't change. The only way duration could decrease is if the quality cutoff was substantially lower with the transfer. In other words, both an increase and a small decrease in quality are consistent with duration increasing.

(Skocpol, 1995).

Fourth, perhaps welfare did not affect marriage prospects because it created incentives not to move since mothers would loose the transfer upon moving. Though welfare would reduce incentives to move, mothers who do move, should move to “better” places (see Appendix), because location influences marriage prospects and determines long-term outcomes of children (Chyn and Katz, 2021). We do find women who receive welfare are about eight percent more likely to live in the same county where they applied for welfare compared those that were rejected (Table 3.3, Columns 1 and 2).[47] Thus welfare receipt significantly lowers geographic mobility. However accepted women move to similar places as rejected women. Moreover, neither group appears to move to better areas relative to where they applied, where “better” places are defined as having higher levels of education, or higher sex ratios. Thus while geographic mobility was affected by transfers, it would seem that marital prospects were not.

Finally, our insignificant results maybe be due to our large standard errors. While for some measures of husband quality we can rule out large differences in quality (e.g. education and occupation), for others we cannot (eg, longevity, the indices) due to lack of statistical precision. Specifically using the 95% confidence intervals, we can rule out increases in the education of the husband greater than 3%, or increases in his occupation-based income larger than 5%. For longevity we can only rule out increases larger than 4 years of life (a 6% increase). As a result, the indices that use weights could improve as much as 20% driven largely by the longevity gains.

The bulk of the evidence presented here suggest that even though the transfers did initially delay marriage, in the long run, women who received welfare married similar men and at similar rates relative to women who did not receive welfare. This is most consistent with aging and stigma effects, though our large standard errors do not allow us to completely rule out the possibility quality increased particularly in terms of the husband’s longevity. Because

[47] We find no effects on the likelihood of staying in the same state. Thus, the reduction in mobility is local. The Oster bounds are tight for these outcomes (Table 3.3, Panel B). The largest upper bound we estimate for the effect is 0.10 (from the CI of the Newey estimates), which is a 15 percent increase in the likelihood of remaining in the same county.

most women had lost their husbands, it's possible that they valued health and longevity highly when finding a new husband.

3.6.4 Effects on Fertility

The MP program incentivized fertility. We test empirically whether welfare recipients had more children after receiving welfare. Fertility post application to the MP program was modest: Only 14% of mothers had any children post welfare application and the differences across the two groups are very small (Figure 3.5). Women on welfare did have 0.421 more children on average, but this difference existed pre-welfare receipt (Table C.3). As Table 3.4 shows, there is no effect of getting welfare on post-welfare fertility, among all mothers or among unmarried mothers only. To rule out that this is due to the relatively old age of mothers in our sample (median age 37), we show that the results are identical if we look at only the youngest mothers in the sample (Table 3.4, last two columns).

These conclusions do not change when we correct for missing data, drop observations with low quality, or compute Oster bounds. The causal random forest ATE and ATT are very similar to the OLS estimates: negative, small and statistically insignificant. Nor are they changed when we estimate fertility from census data which only include the number of children in her household in the 1930 and 1940 census (Table C.9). These results, like the results for marriage are closer to precisely estimated zeroes: Among all mothers, we can rule increases in fertility larger than 0.01 children, a small number relative to the mean number of children of 4.5.

In sum, we find no significant effects on fertility post-welfare receipt, although there are significant differences before.

3.7 Lifetime Maternal Welfare and Implications for Program Evaluation

3.7.1 Overall Maternal Welfare

Critics of welfare often argue that welfare is harmful to women as it traps them in a cycle of poverty and dependence.[48] To shed light on this, we collected two measures of maternal long-run well-being: longevity and her household income in 1940. In Figure 3.6, we compare the distributions of longevity (Panel A) and 1940 household income (Panel B) of the mother by acceptance status. In both cases, we cannot reject that the distributions are identical (p-values reported in the figure). We confirm this in our regression analysis. There do not appear to be any large or significant effects of welfare receipt on long run maternal well-being (Table 3.5): Receiving welfare has a small and positive but insignificant effect of 0.25 years on maternal longevity (Column 1), and a small and negative but insignificant effect of roughly $60, -6% decline relative to the rejected applicants' meanincome (Column 2). A difference of 6% is almost identical to the difference in predicted income at the time of application (last column of Table C.3). These estimates are more sensitive to accounting for attrition: The IPW are positive and statistically significant for longevity (0.9) and essentially 0 for income, perhaps suggesting there are indeed some improvements.

However, these results are imprecise. The OLS CI for longevity ranges from -0.8 to 1.4 years . While not statistically different from the OLS point estimates, the causal random forest estimates are larger and the confidence interval for the ATE is [-0.35; 1.4], which includes small to moderate negative effects and large positive effects.

For income the OLS CI ranges from -12% to 1%, and the causal random forest CI is [-9%; 5%]. Given an initial gap of 6%, this suggest that welfare could have decreased income by 6% (3%) or increased by 7% (11%). While not insignificant, these magnitudes do not suggest large effects on economic outcomes. Thus mothers who applied for welfare were poor and

[48] For example, see Cato Handbook for Policymakers, 8th Edition (2017) chapter 41 Poverty and Welfare

remained so by 1940, regardless of welfare receipt, having roughly half of a typical household's income.

Consistent with these results for income, we find that cash transfers had no large disincentive effects on labor market outcomes (Table C.11), though the data on this are more limited in great part because we observe these only in 1920 and 1930.[49] To estimate these effects we restrict the samples only to women who applied for the MP transfer between 1918 and the census in 1920 (and similarly for 1930) and investigate whether their labor market outcomes in 1920 (1930) differ as a result of the transfer. Although we cannot confirm that all women in this sample are still receiving welfare, the median duration in our records is of 3 years so we expect most are still in the program. We find no statistically significant effects of receiving welfare on the likelihood that women were in the labor force or that they worked. In addition, we find no statistically significant effects of receiving welfare on their earned incomes or their occupational scores when they worked (Table C.11). The labor supply estimates are very noisy though — they are statistically insignificant and include both large positive and negative effects of the transfers. The lowest 95% CI for the estimates correspond to labor supply effects that range from -6% (a small response) to -30% (a more substantial response) associated with a 30% increase in income due to the MP transfer. For comparison, the estimated extensive margin elasticity in the EITC range from 0.7 to 1 (Bastian, 2020; Nichols and Rothstein, 2016). Our estimate is so imprecise likely because we are missing these data for much of our sample and the measure, when we do observe it, is a spot measure. Indeed, the inadequacies of these data are the main reason maternal labor supply response is not the focus of this paper. Interestingly by 1940 (when most applicants would no longer be on welfare) our estimates are positive and significant for labor force participation and work. Thus our estimates suggest that in the long run welfare moms returned to work. Given that welfare did not affect marriage rates and the type of husbands women married, and that it doesn't change labor market outcomes for women, it is not surprising that family

[49] Also because these data are missing at higher rates and differentially by accepted status, see the data section.

income is unchanged, and that ultimately health was not affected either. If mother's longevity did not change as a result of the transfer why did the longevity of her children increase? Together these findings are consistent with the idea that childhood is a critical period for physical development, and is in line with other research that finds that the returns to various government programs is largest for children and young adults (Hendren and Sprung-Keyser, 2020).

In sum we find very few significant changes in the economic and demographic circumstances of women associated with welfare receipt, explaining why we also find no significant improvements or declines on their long-term wellbeing. Thus, the long run evidence from the first welfare program in the US does not support the claim that welfare harms women.

3.7.2 Was the Program Worth it? Marginal Value of Public Funds Computations.

Our previous work documented large positive effects of welfare receipt on the education, income and longevity of their sons (Aizer et al., 2016). Here we find that cash transfers resulted in marriage delays of about a year and decreases in geographic mobility. But they otherwise had no statistically significant negative impacts on maternal behavior and no positive effects on maternal outcomes. We now compute the MVPF of the program using the methodology of Hendren and Sprung-Keyser (2020) to determine how these estimates change the overall evaluation of the program.

The computations are in Table 3.6 Panel A lists the dollar value (in 2019 dollars) of all the benefits and costs associated with the program, using the results from this paper and Aizer et al. (2016) documenting increases in the education, income and longevity of recipients' sons. The benefits of the program are given by the total willingness to pay of recipients. This includes the value of the transfer which lasted three years (about $20,000), plus the value of the spillovers to sons, minus the dollar value of the negative behavioral responses. In Column 1, we ignore spillovers to children. The negative behavioral response we estimate is a delay in remarriage of about a year, costing about $3,500. The total costs of the program are given by the size of the transfers ($20,715) plus or minus the changes in taxes received by the government. Since we estimate that labor supply increases (though this is statistically insignificant), the total cost of the program is a bit lower ($500) as a result. Considering only the benefits to mothers, once we include the dollar value of behavioral responses, the MVPF of the program is 0.84, below one.

However, a more realistic and comprehensive calculation would also consider whether the transfers benefitted children. Aizer et al. (2016) find that boys' longevity increase by about 1.5 years and that labor market income increased by 10% as a result of the transfer. We use their results on the effects of the transfer on the survival curves, along with estimates of the value of life to compute the present discounted value of children's longevity and earnings increases (using a 3% discount rate). These amount to about $61,000 which are added to the total willingness to pay estimates. More earnings also reduce the cost of the program through increased taxes, which, assuming at 10% tax rate, amounts to a savings of about $5,000. Once we incorporate these benefits to the sons, we find that the MVPF of the program is greater than 5, even with maternal behavioral responses (Column 2). The results are similar if we only include spillovers in the benefits and do not count the transfers itself (Column 3, MVPF 3.86).

These computations are subject to uncertainty. Aizer et al. (2016) only tracked the longevity of about 50% of the sons, and the incomes of roughly 15% of the sons in 1940. Additionally they could not track outcomes for daughters. Our computations so far include only benefits for sons and assume there are no benefits for daughters.To address this issue, we compute the smallest increases in income or longevity of the sons that would be needed for the MVPF to be larger than one. We find that if the sons' income over their lifetime increases by only 0.75% then the MVPF exceeds one. Alternatively if their longevity increases by 0.3 years of life, the MVPF would also exceed one. Thus, relatively small benefits for at least some children allows the program to pay for itself, in part because behavioral responses from the mother are relatively minor, and the benefits accrue to sons over a long time horizon.[50]

3.8 Conclusion

Tracking over 16,000 women who applied for the first welfare program in the US between 1911 to 1930, we establish that cash transfers to poor women had no effect on lifetime remarriage rates and fertility. Those with transfers were not less likely to remarry over their lifetime, and they delayed remarriage only in the short-term. The cumulative effect was to delay time to remarriage by about a year. These findings underscore the importance of conducting long-term evaluations, as short-term effects can be misleading.

Why were the effects of the program on marriage so modest? One reason is that the transfers were small relative to the lifetime income that a marriage would bring. The average woman that remarried in our sample was 39 years old and married a 43 year old man who died at age 71. Marriage would bring 22 years of income with relative certainty (assuming retirement at age 65). Cash transfers instead accounted for less than half of the income these women needed to live, and receipt was not guaranteed: Women had to reapply and could lose the transfer if they moved, for instance. The median duration of transfers was three years. Thus, a very rough back-of-the-envelope calculation shows that cash transfers represent only 7% of what a marriage would bring over a lifetime.[51]

We also find that women who received transfers did not marry different men. Although women who wait to remarry do marry better husbands in general, delays induced by welfare receipt are not associated with improvements in the quality of the matches. Thus our findings reject the predictions of a simple search model of welfare and the marriage markets. Other forces such as age and stigma may be more important determinants of marriage behaviors

[50] The table also shows alternative computations. For example in the benefits of the program we count the transfer as a benefit. If we do not count it, and instead only count the benefits for the children, then we require a 6% increase in child income or a 1.5 increase in longevity for the MVPF to be greater than one.

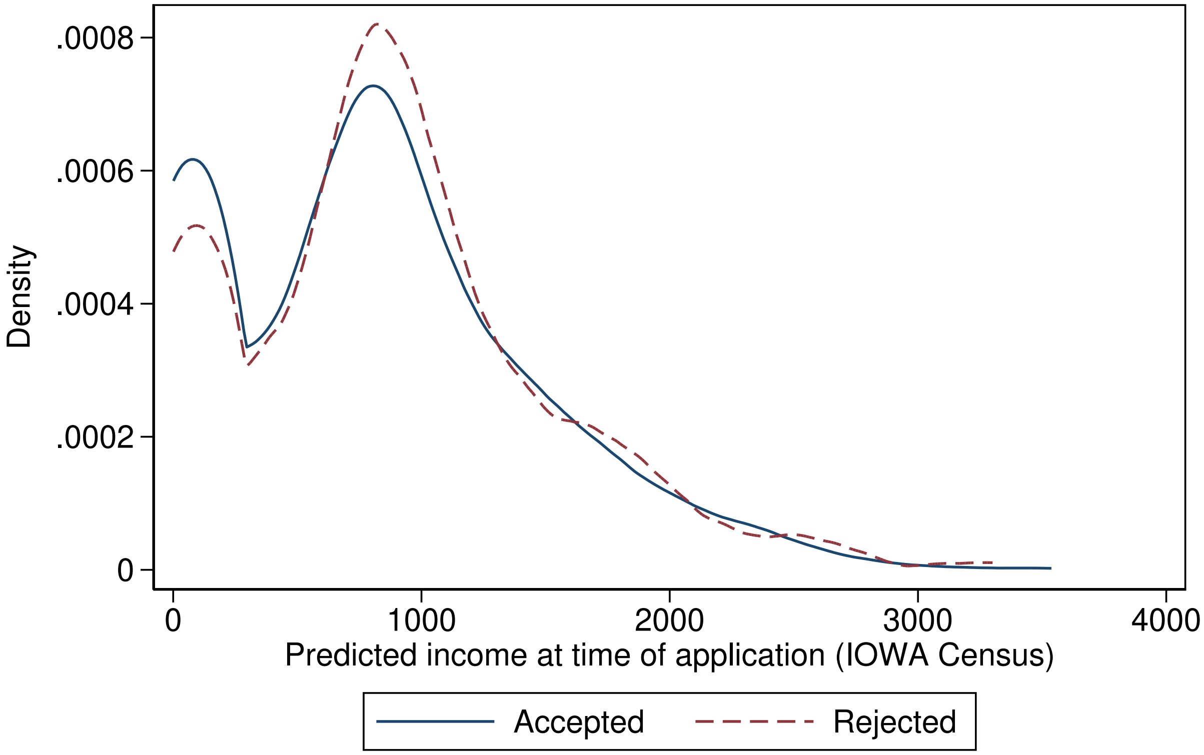
[51] Assuming that a marriage brought in 100% of family income and that the transfer brought in 50% of that income, we compute that the ratio of cash transfer income to marriage income is 3*50/22*100 = 7%)

than the monetary incentives embedded in these government programs. Incorporating these forces into standard models of behavior and further assessing their empirical importance is an important area for future research.

We conclude that the program did not generate large negative incentive effects as predicted by economic models and as feared by policy makers, nor did it help mothers escape poverty. It did, however, appear to help alleviate short-term cash constraints. Thus, ultimately the program should be judged largely by the impact it had on its intended beneficiaries — the children.

3.9 Figures & Tables

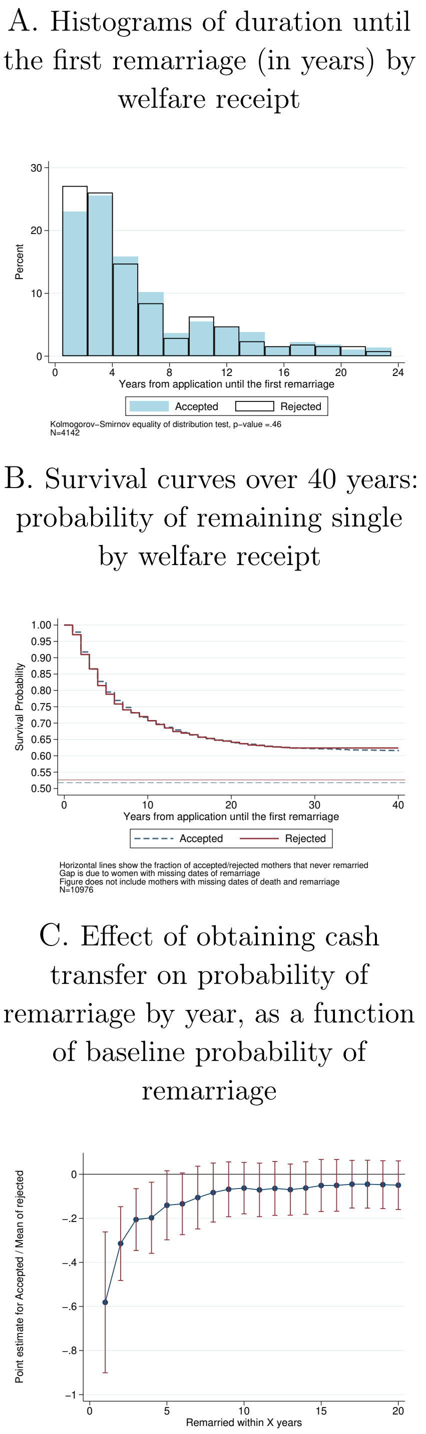
Figure 3.1: Welfare Recipients Have Lower Predicted Incomes Pre-Application



Epanechnikov kernel function, bw=130.99
Kolmogorov−Smirnov equality of distribution test, p−value =.03
N=5332

Notes: Data come from administrative data collected by the authors. Sample includes women with non missing predicted income. Income<1 set to =1. Sample includes 5332 individuals for whom we could compute predicted income using the Iowa Census. The predicted income was computed by running a regression of family income on covariates (widow, mother age at application, number of kids at each age (0-18), age of the youngest and oldest kid, number of kids over 14, mother is foreign, black, education and occupation score. We include interactions of the covariates with the variable widow, and some of the covariates are included in a dummy format.) in the Iowa Census and then using the estimated betas to predict income for all mothers in the MP sample. In the MP sample we use the 1910 census occupation scores and 1940 census education.

Figure 3.2: Welfare Recipients Delay Time to Remarriage



A. Histograms of duration until the first remarriage (in years) by welfare receipt

Kolmogorov–Smirnov equality of distribution test, p–value =.46
N=4142

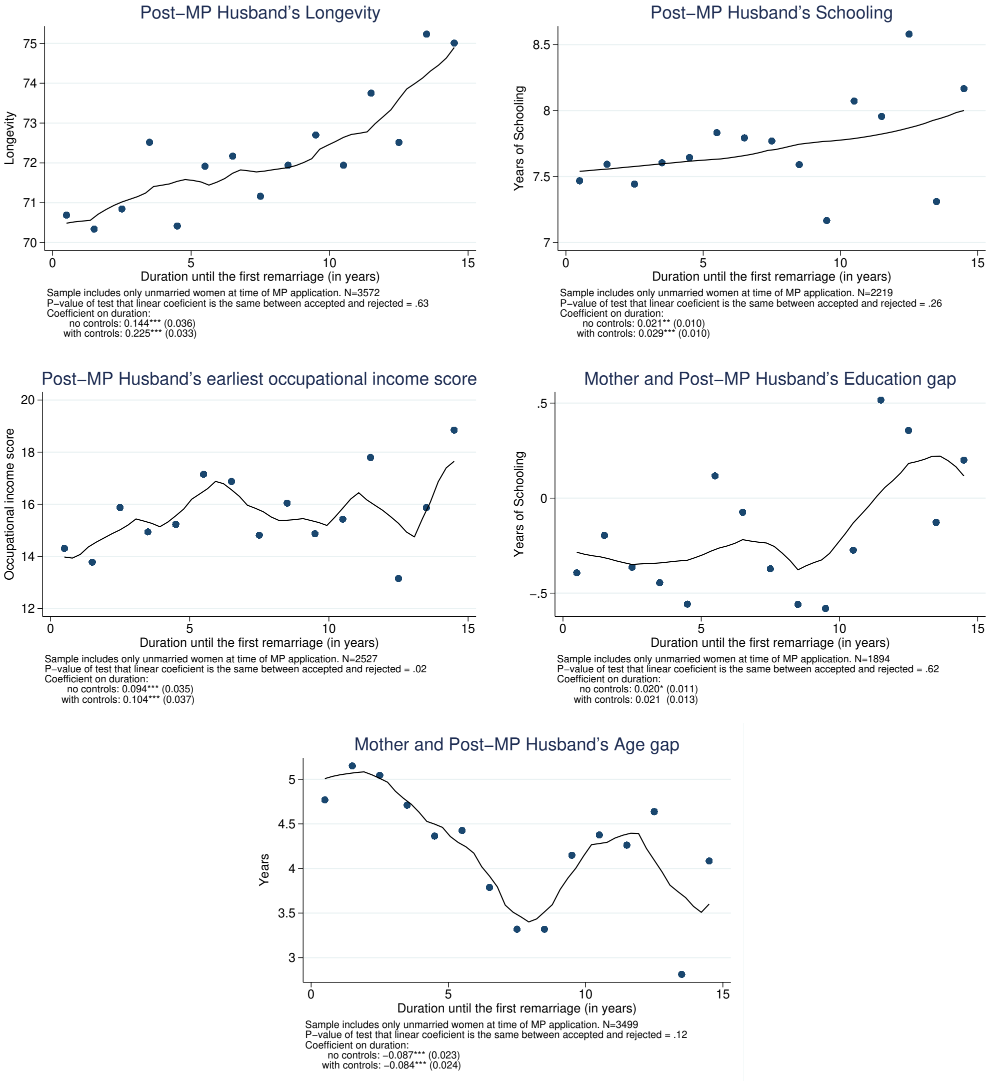
B. Survival curves over 40 years: probability of remaining single by welfare receipt

Horizontal lines show the fraction of accepted/rejected mothers that never remarried
Gap is due to women with missing dates of remarriage
Figure does not include mothers with missing dates of death and remarriage
N=10976

C. Effect of obtaining cash transfer on probability of remarriage by year, as a function of baseline probability of remarriage

Notes: Panel A: The figure plots the duration until the first remarriage by accepted among women who were not married at the time of the application. We cannot reject that the distributions are equal. Sample includes only women that remarried. Panel B: The figure plots the survival curves by accepted for the duration until the first remarriage. Panel C: The figure plots the estimated coefficients of "accepted" divided by the baseline probability of remarriage among rejected applications and 95% confident intervals. Coefficients come from regressions where we regress a dummy indicating that the mother remarried within x years on accepted status and all predetermined characteristics. Standard errors are clustered at the county level. See information in Table 3.1.

Figure 3.3: Delaying Remarriage Improves the Quality of the New Husband



Post–MP Husband's Longevity

Sample includes only unmarried women at time of MP application. N=3572
P–value of test that linear coeficient is the same between accepted and rejected = .63
Coefficient on duration:
no controls: 0.144*** (0.036)
with controls: 0.225*** (0.033)

Post–MP Husband's Schooling

Sample includes only unmarried women at time of MP application. N=2219
P–value of test that linear coeficient is the same between accepted and rejected = .26
Coefficient on duration:
no controls: 0.021** (0.010)
with controls: 0.029*** (0.010)

Post–MP Husband's earliest occupational income score

Occupational income score
20
18
16
14
12
0
5
10
15
Duration until the first remarriage (in years)

Sample includes only unmarried women at time of MP application. N=2527
P–value of test that linear coeficient is the same between accepted and rejected = .02
Coefficient on duration:
no controls: 0.094*** (0.035)
with controls: 0.104*** (0.037)

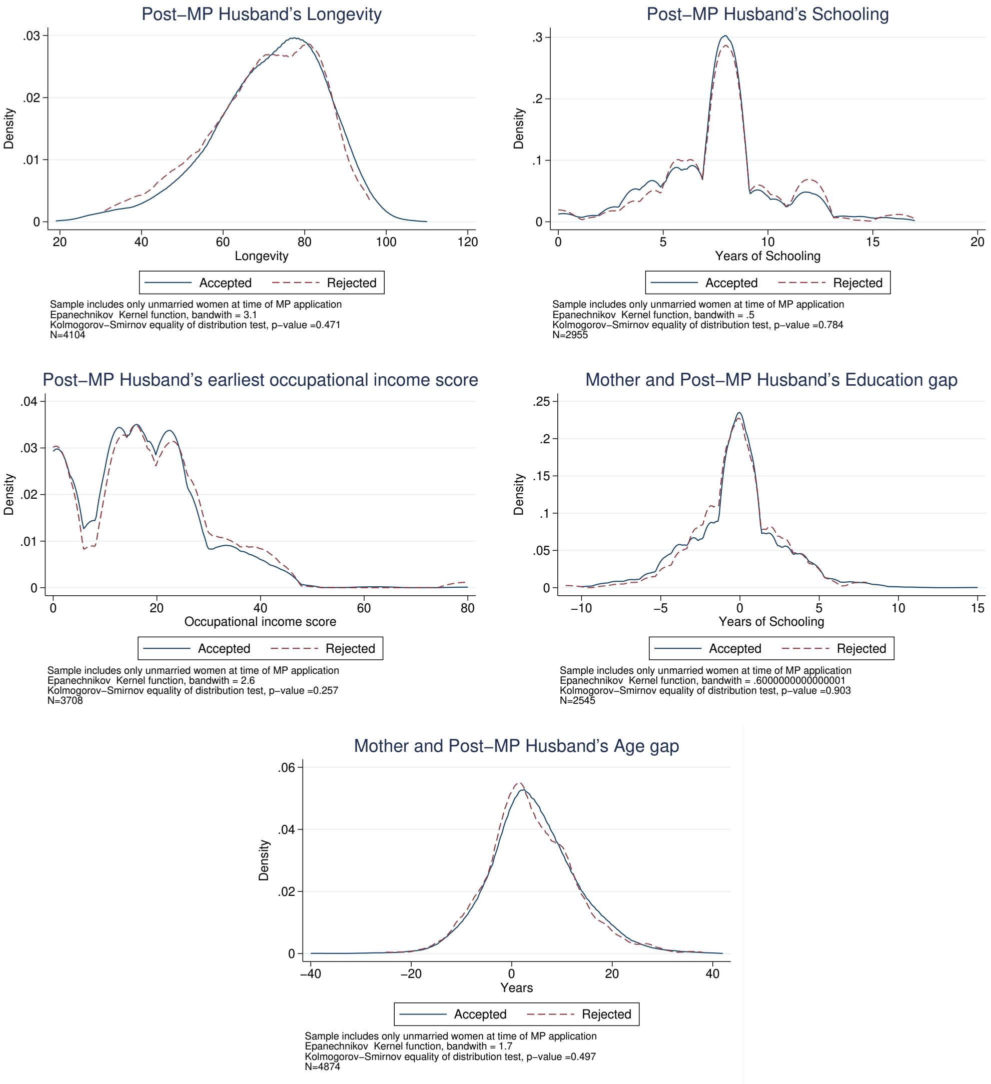
Mother and Post–MP Husband's Education gap

Sample includes only unmarried women at time of MP application. N=1894
P–value of test that linear coeficient is the same between accepted and rejected = .62
Coefficient on duration:
no controls: 0.020* (0.011)
with controls: 0.021 (0.013)

Mother and Post–MP Husband's Age gap

Sample includes only unmarried women at time of MP application. N=3499
P–value of test that linear coeficient is the same between accepted and rejected = .12
Coefficient on duration:
no controls: –0.087*** (0.023)
with controls: –0.084*** (0.024)

Figure 3.4: Welfare Recipients Do Not Marry Better Men



Post−MP Husband's Longevity

Sample includes only unmarried women at time of MP application
Epanechnikov Kernel function, bandwith = 3.1
Kolmogorov−Smirnov equality of distribution test, p−value =0.471
N=4104

Post−MP Husband's Schooling

Sample includes only unmarried women at time of MP application
Epanechnikov Kernel function, bandwith = .5
Kolmogorov−Smirnov equality of distribution test, p−value =0.784
N=2955

Post−MP Husband's earliest occupational income score

Sample includes only unmarried women at time of MP application
Epanechnikov Kernel function, bandwith = 2.6
Kolmogorov−Smirnov equality of distribution test, p−value =0.257
N=3708

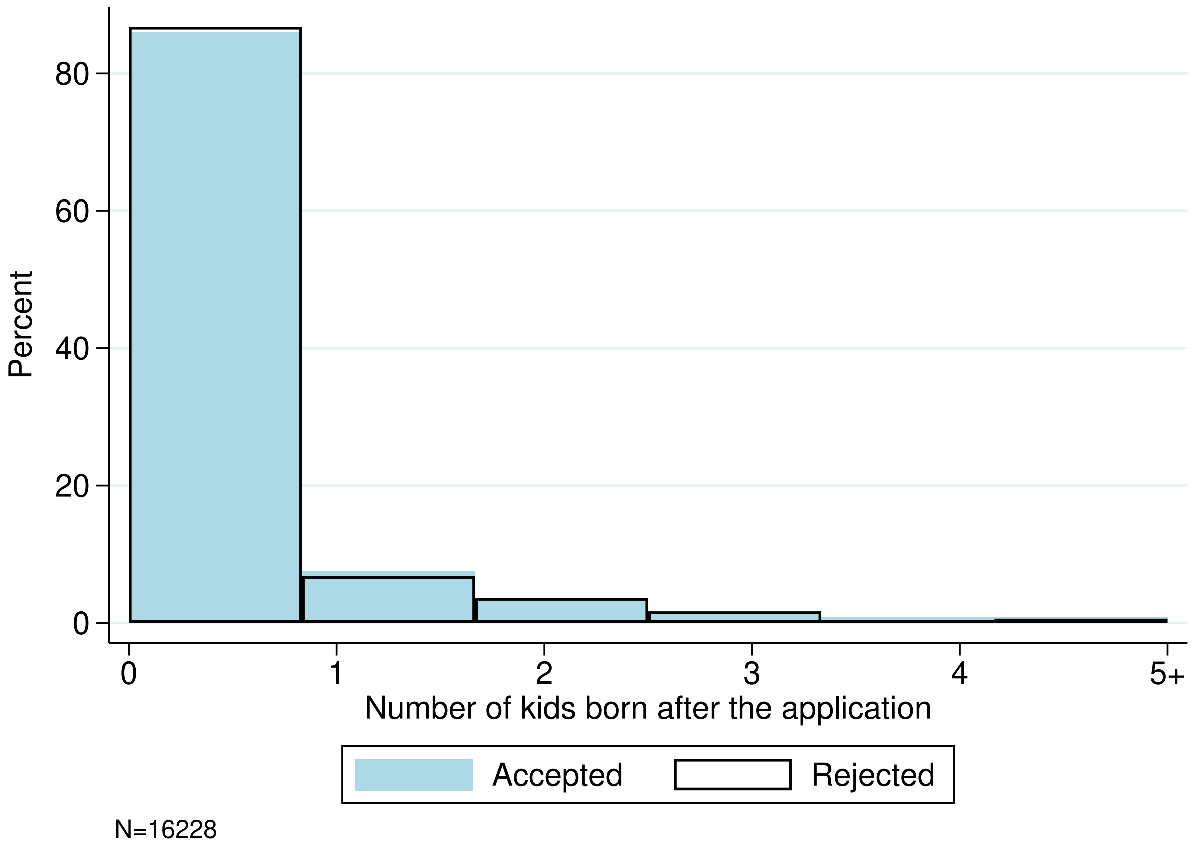
Mother and Post−MP Husband's Education gap

Sample includes only unmarried women at time of MP application
Epanechnikov Kernel function, bandwith = .6000000000000001
Kolmogorov−Smirnov equality of distribution test, p−value =0.903
N=2545

Mother and Post−MP Husband's Age gap

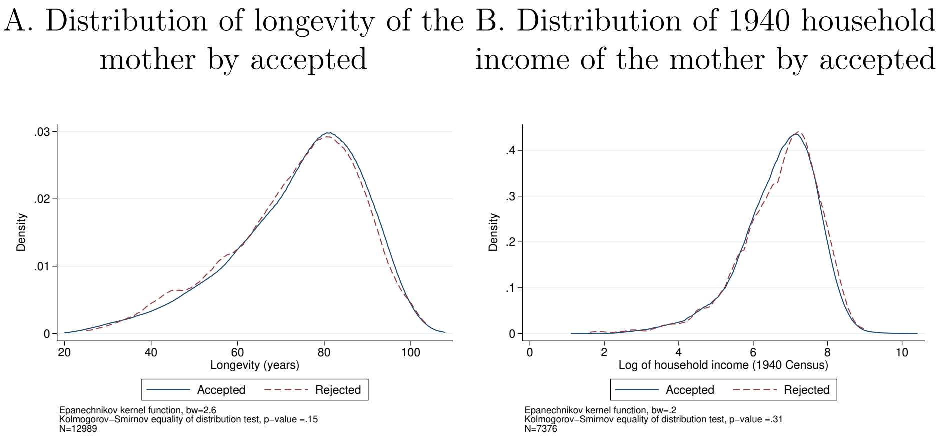
Sample includes only unmarried women at time of MP application
Epanechnikov Kernel function, bandwith = 1.7
Kolmogorov−Smirnov equality of distribution test, p−value =0.497
N=4874

Figure 3.5: Welfare recipients do not have more children after receiving welfare



Notes: The figure plots the distribution of kids born after application by accepted. The sample includes all women.

Figure 3.6: Welfare Recipients' Long-Term Well-Being Is Not Affected By Receiving Welfare



A. Distribution of longevity of the mother by accepted

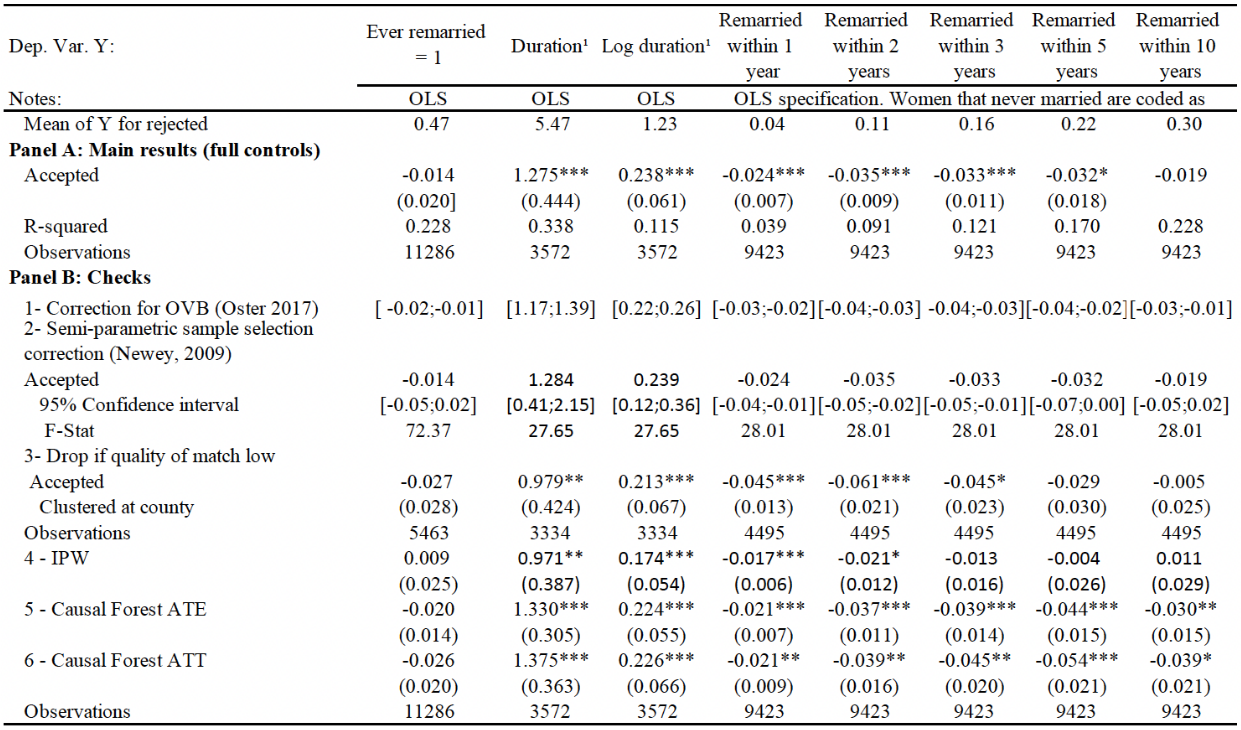
B. Distribution of 1940 household income of the mother by accepted

Epanechnikov kernel function, bw=2.6
Kolmogorov−Smirnov equality of distribution test, p−value =.15
N=12989

Epanechnikov kernel function, bw=.2
Kolmogorov−Smirnov equality of distribution test, p−value =.31
N=7376

Notes: Panel A: The figure plots the distribution of the longevity of the mother by accepted. We cannot reject that both distributions are equal. The sample includes all women with non-missing longevity. Panel B: The figure plots the distribution of 1940 household income by accepted. We cannot reject that both distributions are equal. The sample includes all women with non-missing and non-zero household income.

Table 3.1: Welfare recipients with cash transfers delay remarriage



Dep. Var. Y:	Ever remarried = 1	Duration[1]	Log duration[1]	Remarried within 1 year	Remarried within 2 years	Remarried within 3 years	Remarried within 5 years	Remarried within 10 years
Notes:	OLS	OLS	OLS	OLS specification. Women that never married are coded as				
Mean of Y for rejected	0.47	5.47	1.23	0.04	0.11	0.16	0.22	0.30
Panel A: Main results (full controls)								
Accepted	-0.014	1.275***	0.238***	-0.024***	-0.035***	-0.033***	-0.032*	-0.019
	(0.020]	(0.444)	(0.061)	(0.007)	(0.009)	(0.011)	(0.018)	
R-squared	0.228	0.338	0.115	0.039	0.091	0.121	0.170	0.228
Observations	11286	3572	3572	9423	9423	9423	9423	9423
Panel B: Checks								
1- Correction for OVB (Oster 2017)	[-0.02;-0.01]	[1.17;1.39]	[0.22;0.26]	[-0.03;-0.02]	[-0.04;-0.03]	-0.04;-0.03]	[-0.04;-0.02]	[-0.03;-0.01]
2- Semi-parametric sample selection correction (Newey, 2009)								
Accepted	-0.014	1.284	0.239	-0.024	-0.035	-0.033	-0.032	-0.019
95% Confidence interval	[-0.05;0.02]	[0.41;2.15]	[0.12;0.36]	[-0.04;-0.01]	[-0.05;-0.02]	[-0.05;-0.01]	[-0.07;0.00]	[-0.05;0.02]
F-Stat	72.37	27.65	27.65	28.01	28.01	28.01	28.01	28.01
3- Drop if quality of match low								
Accepted	-0.027	0.979**	0.213***	-0.045***	-0.061***	-0.045*	-0.029	-0.005
Clustered at county	(0.028)	(0.424)	(0.067)	(0.013)	(0.021)	(0.023)	(0.030)	(0.025)
Observations	5463	3334	3334	4495	4495	4495	4495	4495
4 - IPW	0.009	0.971**	0.174***	-0.017***	-0.021*	-0.013	-0.004	0.011
	(0.025)	(0.387)	(0.054)	(0.006)	(0.012)	(0.016)	(0.026)	(0.029)
5 - Causal Forest ATE	-0.020	1.330***	0.224***	-0.021***	-0.037***	-0.039***	-0.044***	-0.030**
	(0.014)	(0.305)	(0.055)	(0.007)	(0.011)	(0.014)	(0.015)	(0.015)
6 - Causal Forest ATT	-0.026	1.375***	0.226***	-0.021**	-0.039**	-0.045**	-0.054***	-0.039*
	(0.020)	(0.363)	(0.066)	(0.009)	(0.016)	(0.020)	(0.021)	(0.021)
Observations	11286	3572	3572	9423	9423	9423	9423	9423

Notes: Sample includes only women who were not married at the time of application. Standard errors clustered at the county level. Controls for county and year-of-application fixed effects and individual, county and state controls. Individual controls: Kids: MP age of the youngest and oldest, MP dummies for number, FS number older than 14, FS number that died before MP, FS number with dates missing. Mother: last name length, dummies for divorced, widowed and missing marital status, age at application, missing age, number of siblings, foreign, missing nativity, first husband's longevity, first husband's longevity is missing. County controls: for ages 18-55: sex ratio (M/F), shares of white married mothers in the labor force, black and rural. County controls match linear interpolated information from the 1910, 1920 and 1930 census with the year of MP application. State controls: manufacturing wages, education/labor laws (age must enter school, work permit age, and continuation school law in place), state expenditures in logs (education, charity, and social programs), state laws concerning MP transfers (work required, reapplication required, maximum amount for the first child and for each additional child). [1]The duration measure starts at 0.5 (the variable is duration + 0.5, so we assume that marriages occur uniformly within a year). We also assume that if women married the same year they applied for the pension (and the exact data of marriage is missing) that the marriage took place after the MP application. [2]Low quality of match is defined as observations with remarriage dates that do not include day, month and year of marriage. Omitted variable bounds: We use Oster (2017) to construct omitted variable bias (OVB) bounds. We assume that the R-max is 1.3 times greater than the R-squared from panel B. We assume delta = (-1, 1) for lower and upper bounds. Sample Selection Correction: We follow the two-step estimation suggested by Newey (2009) to correct for sample selection. First, we regress the dummy indicating whether the outcome is missing on RA fixed effects (73 dummies) and all other controls. We report the F-statistic of the test of relevance of these dummies. Second, we estimate a linear regression of the outcome on controls and on a fourth degree polynomial of predicted values from the first stage. We jointly bootstrap the two stages and report the 95% bias corrected confidence interval clustered at the county level, from 200 repetitions. Quality of match: Regressions that drop low quality matches (quality measure below its median) include all controls and cluster the standard errors at the county level. The quality of match between census, family search and administrative data is constructed as the weighted sum of variables that access the similarity between first name, last name, full name, age and place of birth in each dataset. IPW: We estimate the average treatment effect using the estimated probability weights to address for potential missing outcomes. The standard errors are clustered at the county level and a logit model is used to predict the accepted status. Causal Forest: We implement the generalized random forest algorithm proposed by Athey, Tibshirani, and Wager (2019). We estimate the average treatment effects using a doubly robust augmented-inverse-propensity weighting estimation method and report the ATE and ATT. See Appendix for more details.

Table 3.2: Welfare receipt does not increase quality of Post-MP husband

	New husband's traits			New match characteristics			Overall Index		
Outcome:	Post-MP Husband Longevity	Post-MP Husband Education	Occ Score[2]	Age gap (shifted by 2.5 years)[1]	Education gap[3]	P-value (H0: all = 0)	Equal weights [4]	Equal weights (no age, education gap)	Satisfaction weights [5]
	(1)	(4)	(3)	(2)	(5)	(6)	(7)	(8)	(9)
Mean of outcome for rejected	70.130	7.798	21.220	6.661	1.821		-0.0470	-0.0465	0.361
Panel A: Main results (full controls)									
Accepted	1.821**	-0.226	-0.828	0.275	-0.064	0.095	0.095**	0.087*	-0.006
	(0.903)	(0.228)	(0.574)	(0.289)	(0.185)		(0.046)	(0.044)	(0.021)
Observations	4,104	2,955	3,556	4,874	2,545		4,894	4,606	2,540
Panel B: Control for other traits (proposition 2)									
Mean of outcome for rejected	73.99	7.946	20.18	6.345	1.818				
Accepted	1.368	-0.334	-0.425	0.247	0.031	0.719			
	(1.309)	(0.279)	(0.749)	(0.599)	(0.239)				
Observations	1,887	1,887	1,887	1,887	1,887				
Panel C: Control for mom's age at marriage									
Mean of outcome for rejected	71.08	7.905	20.28	6.826	1.924		0.0214	0.0184	0.360
Accepted	0.906	-0.362	-1.293**	0.133	-0.044	0.115	0.103*	0.103*	-0.013
	(0.960)	(0.221)	(0.624)	(0.346)	(0.198)		(0.058)	(0.056)	(0.022)
Observations	3,116	2,218	2,424	3,499	1,893		3,505	3,333	1,889
Panel D: Main Results using Causal Forest									
Causal Forest ATE	1.430*	-0.313*	-0.714	0.092	0.019		0.063	0.072	0.001
	(0.732)	(0.168)	(0.609)	(0.272)	(0.123)		(0.043)	(0.044)	(0.015)
Causal Forest ATT	1.475	-0.348*	-0.659	0.110	0.010		0.059	0.070	-0.001
	(0.919)	(0.206)	(0.771)	(0.352)	(0.146)		(0.055)	(0.056)	(0.018)
Observations	4104	2955	3556	4874	2545		4894	4606	2540

Notes: Standard errors clustered at county level. See Table 3.1 for description of controls, restrictions and checks. Panel B includes the other inputs (Post-MP Husband longevity, age gap, Post-MP Husband latest occ. score, Post-MP Husband 1940 education and education gap) as controls (except if the input is the regression dep. var.). In column 6, we present the P-value of the test with null hypothesis that the estimates from columns 1 to 5 are jointly equal to zero. [1]Age gap is defined as the absolute value of the husband's age minus the mother's age minus 2.5. [2]Defined from pre-marriage data: uses 1940 if available, then 1930, then 1920, then 1910. Never uses a measure that is observed post-MP marriage. [3]Education gap is defined as the absolute value of difference in highest grade between the mother and the husband. [4]Equal Weights regressions give the same weight to each of the quality measures. Values are standardized to zero mean and variance equals one. [5]Satisfaction weights include husband's occ. score, education and longevity. We use the utility function and the parameters defined and calibrated in Grow and Van Bavel (2015) to construct the dependent variable. The equation below presents the utility function. The first term of the equation is the similarity of education, second term is the earnings prospect and, last term is the age gap. We follow the same categorization of variables as in the original paper, except for education, where we divide it in 4 quartile categories instead of the four categories in the paper (no schooling, primary, secondary and tertiary). $\alpha_i = a_i + 25$ To take into account that female agents prefer partners who are about 2.5 years older. The parameters are: $S_{max} = 4$; $Y_{max} = 5$; $A_{max} = 800$; $w_s = 0.385$; $w_y = 1.201$; $w_a = 10.833$. $v_{ij} = \left(\frac{S_{max}-|si-sj|}{S_{max}}\right)^{w_s}\left(\frac{y_i}{Y_{max}}\right)^{w_y}\left(\frac{A_{max}-|\alpha_i-\alpha_j|}{A_{max}}\right)^{w_a}$

Table 3.3: Welfare receipt lowered geographic mobility

Sample:	All mothers		All mothers who moved			
Outcome:	lives in MP county in 1930	lives in MP county in 1940	lives in more educated county in 1930	lives in higher sex ratio county in 1930	lives in more educated county in 1940	lives in higher sex ratio county in 1940
Mean of Y for rejected	0.65	0.59	0.50	0.50	0.51	0.54
Panel A: Main results (Full controls)						
Accepted	0.048**	0.063***	0.028	0.038	0.021	-0.004
	(0.024)	(0.018)	(0.024)	(0.031)	(0.024)	(0.032)
R-squared	0.176	0.114	0.399	0.333	0.405	0.289
Observations	11178	9358	3123	2009	3177	3136
Panel B: Checks						
1- Correction for OVB (Oster 2017)	[0.04;0.06]	[0.06;0.07]	[0.03;0.03]	[0.03;0.05]	[0.02;0.02]	[-0.02;0.01]
2- Semi-parametric sample selection correction (Newey, 2009)						
Accepted	0.049	0.063	0.028	0.037	0.022	-0.003
95% Confidence interval	[0.00;0.10]	[0.03;0.10]	[-0.02;0.08]	[-0.02;0.10]	[-0.03;0.07]	[-0.07;0.06]
F-Stat	25.57	116.82	22.13	27.58	29.86	27.52
3- Drop if quality of match low						
Accepted	0.069***	0.053*	-0.023	0.080	-0.016	0.049
Clustered at county	(0.026)	(0.028)	(0.042)	(0.062)	(0.042)	(0.041)
Observations	5589	4679	1249	775	1362	1352
4 - IPW	0.080***	0.078***	0.032	0.019	0.024	-0.047
	(0.027)	(0.028)	(0.048)	(0.053)	(0.044)	(0.049)
5 - Causal Forest ATE	0.081***	0.092***	-0.002	0.033	0.012	-0.018
	(0.016)	(0.018)	(0.024)	(0.033)	(0.025)	(0.026)
6 - Causal Forest ATT	0.093***	0.103***	-0.006	0.032	0.014	-0.019
	(0.023)	(0.025)	(0.029)	(0.036)	(0.030)	(0.030)
Observations	11178	9358	3123	2009	3177	3136

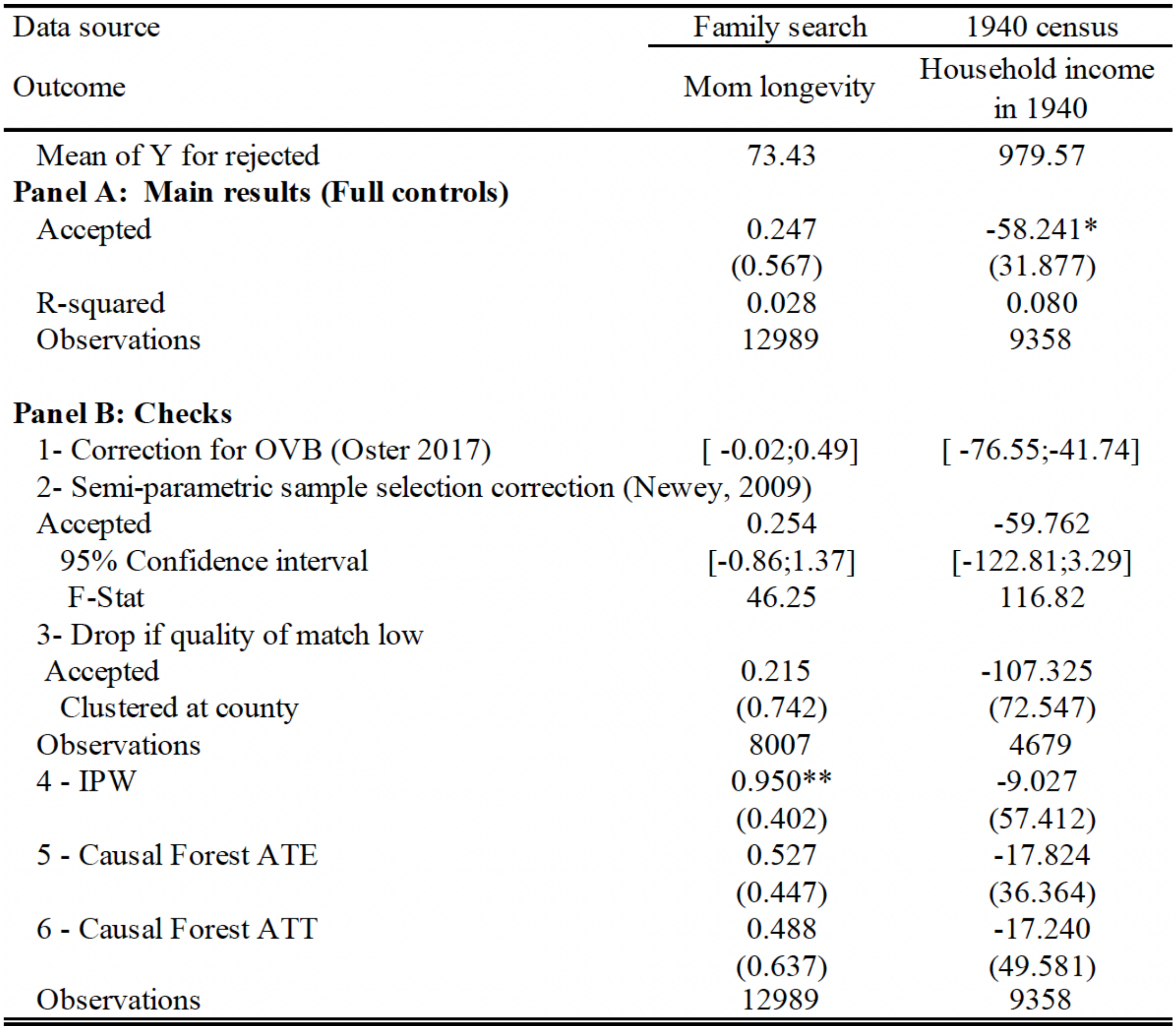
Notes: Sample: all mothers in application. Refer to Table 3.1 for a full description of the controls, restrictions and checks. Counties are ranked by the average schooling in the population between 18 and 55 years old in the 1940 census. Counties are ranked by the sex ratio at the year of application (interpolated between 1910, 1920 and 1930 censuses). We then estimate whether women moved to places above of below the median.

Table 3.4: Welfare recipients do not have more children

Outcome: Number of kids born after application for welfare	All ages		Women below median age (37)	
Sample:	All mothers	Mothers that were not married at time of application	All mothers	Mothers that were not married at time of application
Mean of Y for rejected	0.25	0.22	0.42	0.40
Panel A: Main results (Full controls)				
Accepted	-0.023	-0.009	-0.032	-0.005
	(0.018)	(0.021)	(0.036)	(0.045)
R-squared	0.160	0.162	0.157	0.160
Observations	16228	13383	9014	7168
Panel B: Checks				
1- Correction for OVB (Oster 2017)	[-0.04;-0.01]	[-0.03;0.01]	[-0.06;-0.01]	[-0.03;0.02]
2- Semi-parametric sample selection correction (Newey, 2009)				
Accepted	-0.023	-0.009	-0.030	-0.003
95% Confidence interval	[-0.06;0.01]	[-0.05;0.03]	[-0.10;0.04]	[-0.09;0.09]
F-Stat	.	75.57	42.64	16.13
3- Drop if quality of match low				
Accepted	-0.014	0.008	-0.048*	-0.003
Clustered at county	(0.028)	(0.031)	(0.026)	(0.034)
Observations	7577	6266	5738	4782
4 - IPW	0.022	0.038	0.032	0.057
	(0.025)	(0.025)	(0.040)	(0.043)
5- Causal Forest ATE	-0.010	0.009	-0.025	0.002
	(0.020)	(0.019)	(0.035)	(0.038)
6- Causal Forest ATT	-0.008	0.014	-0.028	0.002
	(0.029)	(0.028)	(0.050)	(0.052)
Observations	16228	13383	9014	7168

Notes: Standard errors clustered at the county level. Refer to Table 3.1 for a full description of the controls, restrictions and checks.

Table 3.5: Welfare receipt did not benefit or hurt mothers in the long run



Data source	Family search	1940 census
Outcome	Mom longevity	Household income in 1940
Mean of Y for rejected	73.43	979.57
Panel A: Main results (Full controls)		
Accepted	0.247	-58.241*
	(0.567)	(31.877)
R-squared	0.028	0.080
Observations	12989	9358
Panel B: Checks		
1- Correction for OVB (Oster 2017)	[-0.02;0.49]	[-76.55;-41.74]
2- Semi-parametric sample selection correction (Newey, 2009)		
Accepted	0.254	-59.762
95% Confidence interval	[-0.86;1.37]	[-122.81;3.29]
F-Stat	46.25	116.82
3- Drop if quality of match low		
Accepted	0.215	-107.325
Clustered at county	(0.742)	(72.547)
Observations	8007	4679
4 - IPW	0.950**	-9.027
	(0.402)	(57.412)
5 - Causal Forest ATE	0.527	-17.824
	(0.447)	(36.364)
6 - Causal Forest ATT	0.488	-17.240
	(0.637)	(49.581)
Observations	12989	9358

Notes: Sample includes all mothers regardless of marital status. Please refer to Table 3.1 for a full description of the controls, restrictions and checks. The quality measure uses the standardized Jaro-Winkler distance for longevity in column 1, and the standarized Jaro-Winkler distance for the 1940 census match in Column 2.

Table 3.6: Marginal Value of Public Funds for the Mothers' Pension Program

All values expressed in 2019 dollars

	Mothers	Including Spillovers on Boys and Assuming no Spillovers to Girls	
		Income and longevity benefits on kids	Transfer not counted as a benefit
Panel A: computations based on the results of this paper and of Aizer et al. (2016)			
Dollar value of maternal behavioral response (marriage delay and mobility decrease)	3,660.68	3,660.68	3,660.68
Dollar value of spillover for kids (mortality + income)	0	61,481	61,481
Dollar value of increased income taxes from kids (10% tax rate)	0	5,225	5,225
Dollar value of increased income taxes from mom (10% tax rate)	507.59	507.59	507.59
Total transfer	20,715	20,715	20,715
Total benefit or WTP (transfer + spillovers - cost of behavioral responses)	17,054	78,535	57,820
WTP excluding cost of behavioral responses	20,715	82,196	61,481
Total cost (transfer - taxes from increased earnings)	20,207	14,982	14,982
MVPF without behavioral responses from mother	**1.00**	**5.49**	**4.10**
MVPF including behavioral responses	**0.84**	**5.24**	**3.86**
Panel B: Minimum gains for children needed for an MVPF of 1			
Minimum change in kids' life expectancy[1] in years (for a MVPF=1)		0.34	1.45
Minimum percentatge change in kids' income[2] (for a MVPF=1)		0.75%	5.67%

Notes: This table computes the Marginal Value of Public Funds (MVPF) using the methodology of Hendren and Sprung-Keyser (2019). We correct for discounting using a 3% rate, and we do not consider the implications of life extensions on Medicare and SSA pensions. We ignore the effects of the pension on marriage rates, type of husband, and years of schooling of the children. These are treated as intermediate outcomes whose ultimate value is reflected in increases in income and longevity. The dollar value of maternal behavioral response includes the discounted effects on marriage delay and mobility decrease. The value of spillover for kids includes the discounted effects on mortality from age 10 to 85 and discounted income effects for the children's average working period, 45 years. We assume a 10% tax rate that is discounted for mothers and children average working periods (27 and 45 years, respectively). The total transfer takes into account that mothers are in the program, on average, for 3 years. [1] Assumes no change in kids income. [2] Assumes there is no change in kids longevity and takes into account the increase in income taxes from kids.

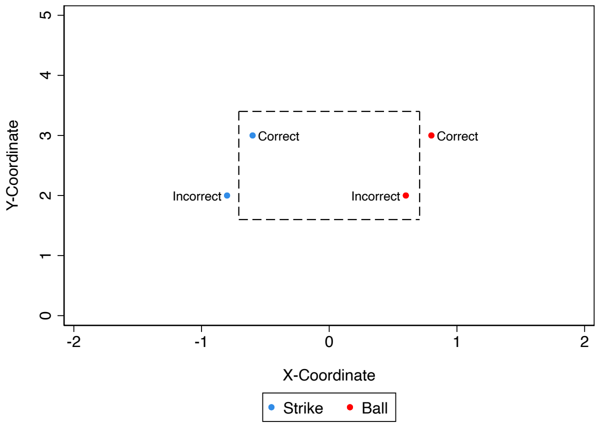
Appendices

Appendix A

Appendix to “The Effect of Robot Assistance on Skills”

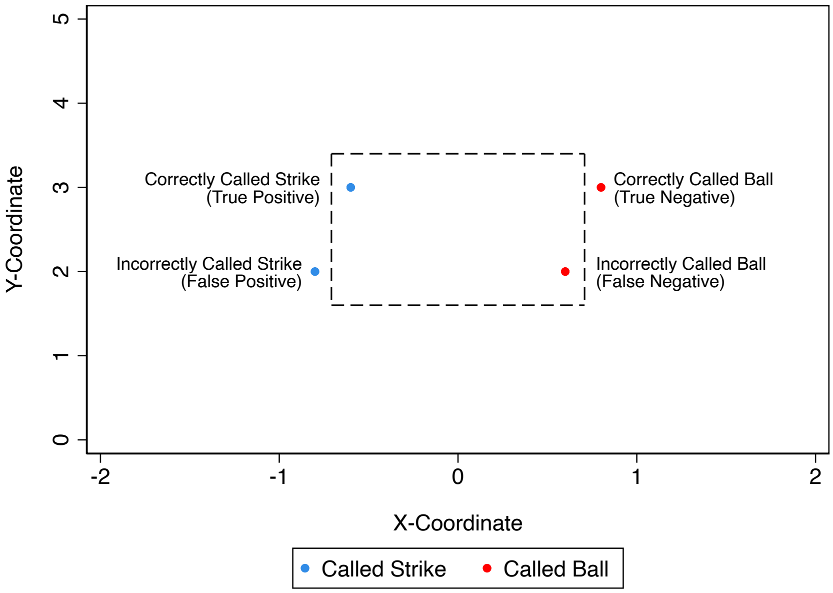
A.1 Figures & Tables

Figure A.1: Strike Zone



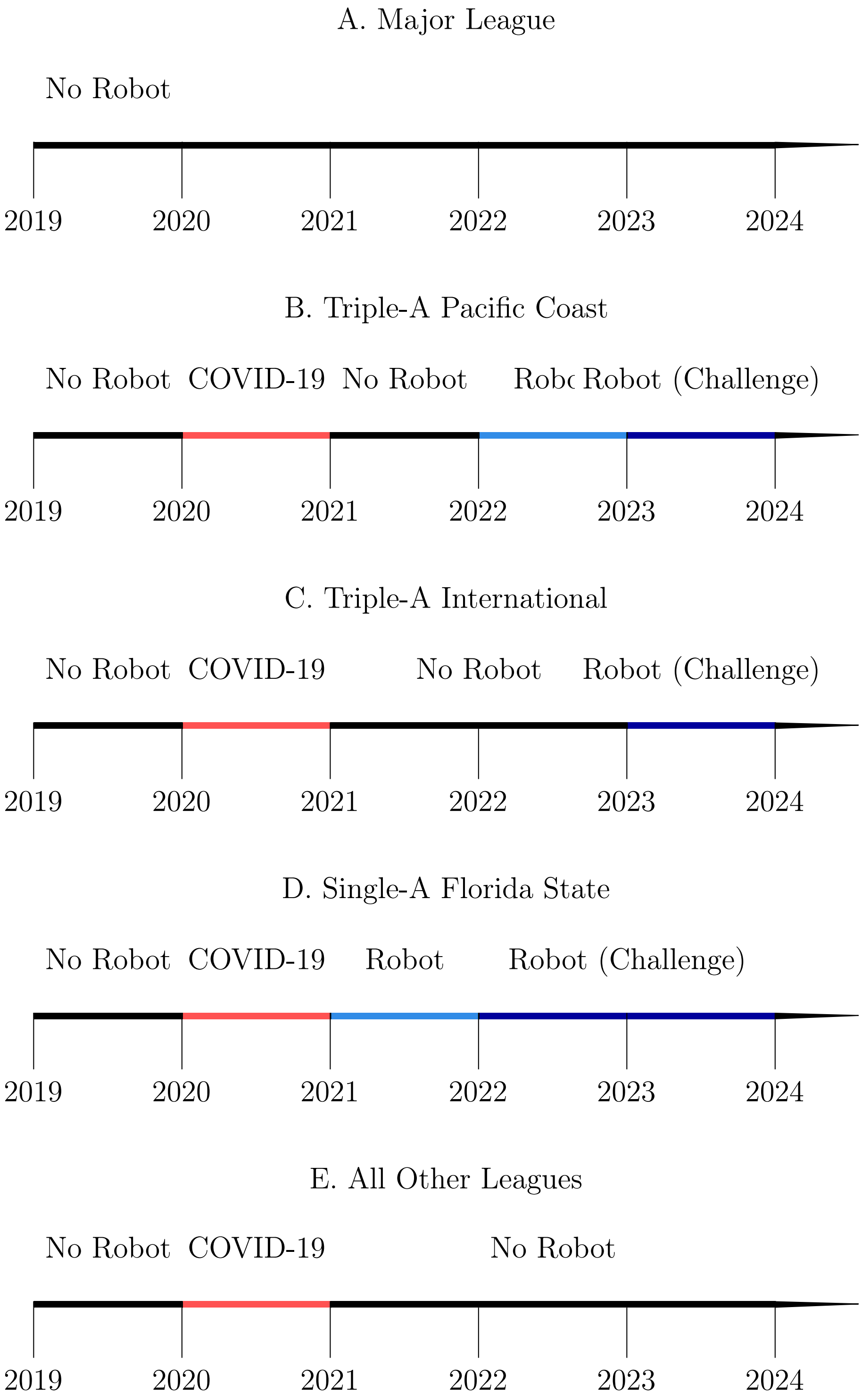
Notes: The figure describes the strike zone and the umpire’s decision classification. Pitches that are called strikes and ball are in blue and red, respectively. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. The strike zone is defined as an imaginary rectangular region over the home plate that extends roughly between the batter’s shoulders and kneecap and the dimension is about 20 by 25 inches. Axes are in feet.

Figure A.2: Strike Zone - Robot Implementation Adjustment



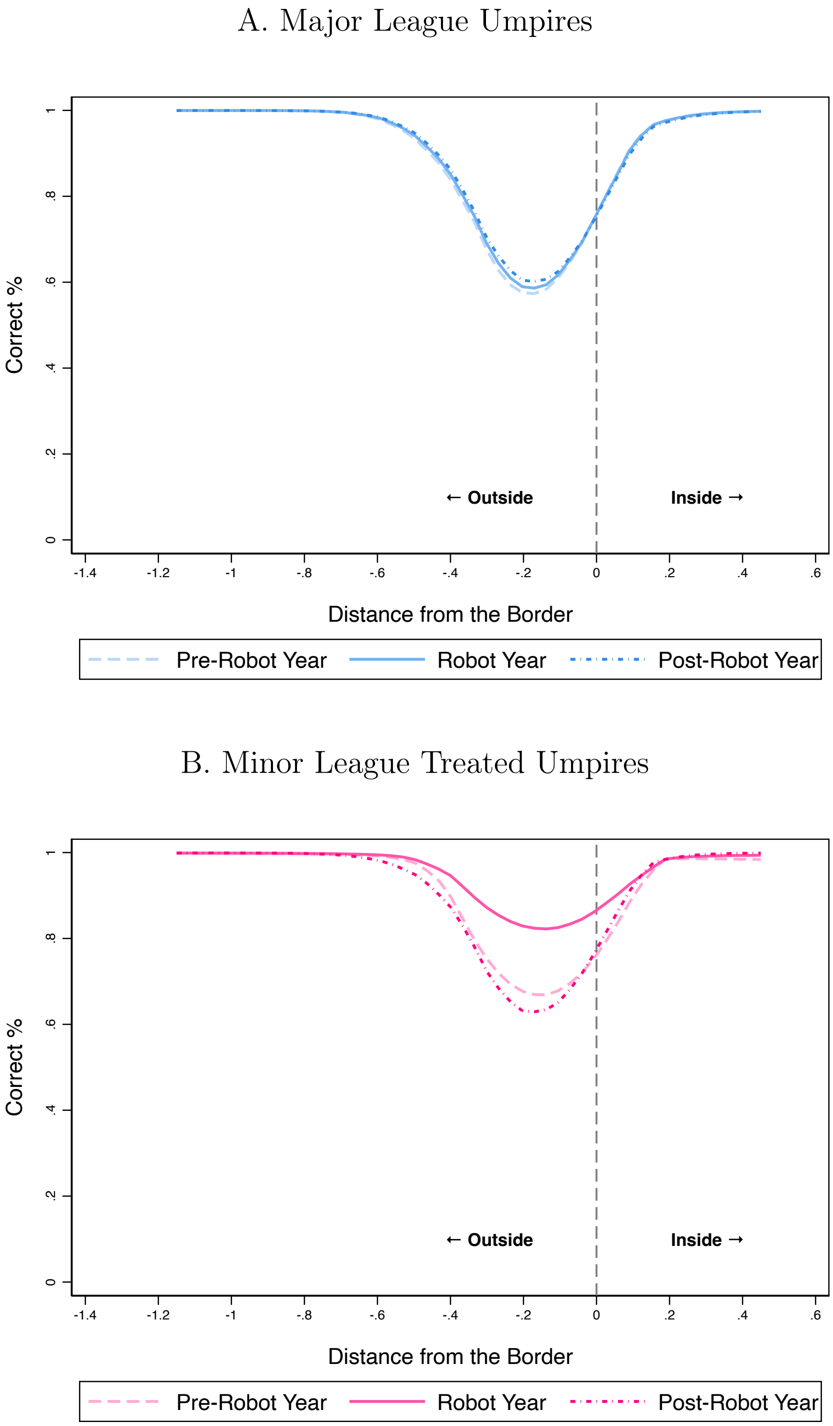
Notes: The figure describes the strike zones. The red-dashed lines represent the adjusted strike zone implemented for the robot since July 20, 2021 and the black-solid line represents the original strike zone. The strike zone is defined as an imaginary rectangular region over the home plate that extends roughly between the batter's shoulders and kneecap and the dimension is about 20 by 25 inches. Axes are in feet.

Figure A.3: Chronology of Robot Implementation



Notes: The figure shows timeline of the implementation of the robot in different leagues. COVID-19 pause is shown in red. “Robot” indicate that the league implemented the robot and “Robot (Challenge)” indicate that the challenge system described in Section A.3 is used.

Figure A.4: Pitch Distribution and Accuracy

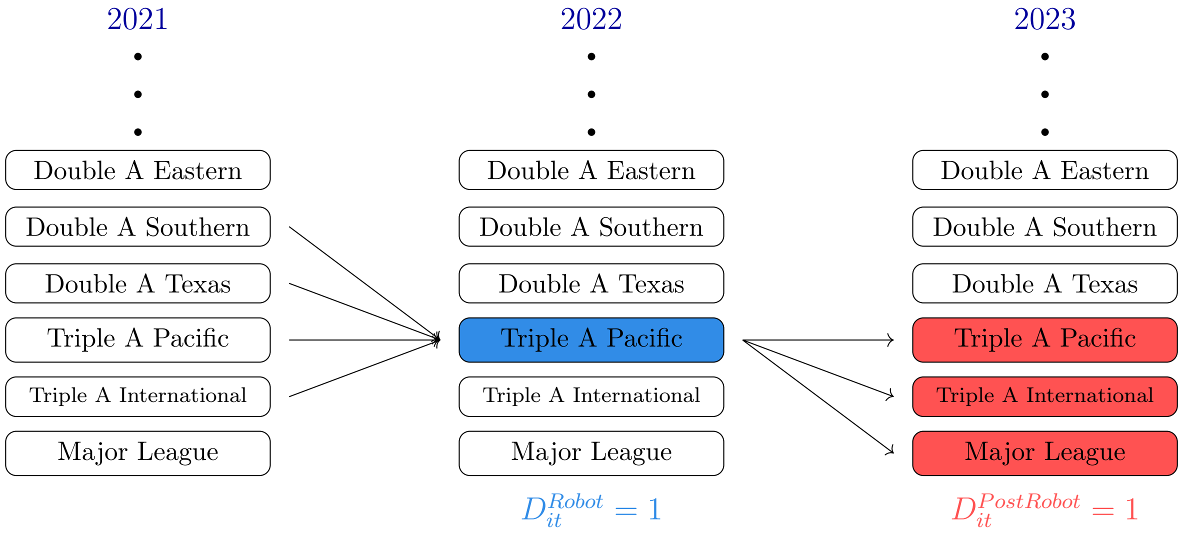


A. Major League Umpires

B. Minor League Treated Umpires

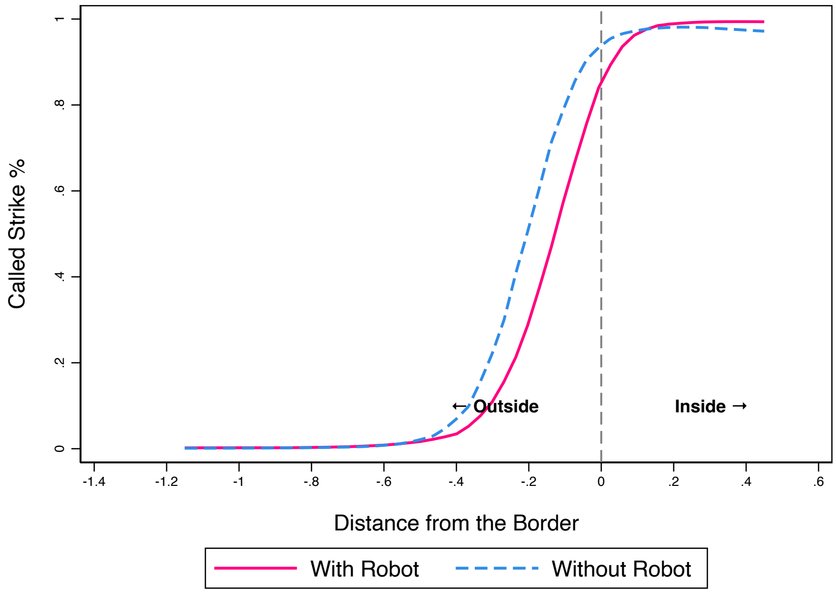
Notes: X-axis is the pitch distance from the nearest border of the strike zone in feet and y-axis shows the average accuracy rate. To the left of origin are pitches falling outside of the strike zone, and to the right are pitches falling inside the strike zone. Panel A shows the distribution of pitches for the umpires who worked in the Major League from 2021 to 2023 and Panel B shows for the umpires who worked in Triple-A Pacific in 2022 and games without robot in 2021 and 2023.

Figure A.5: Robot Implementation and Umpire Moves



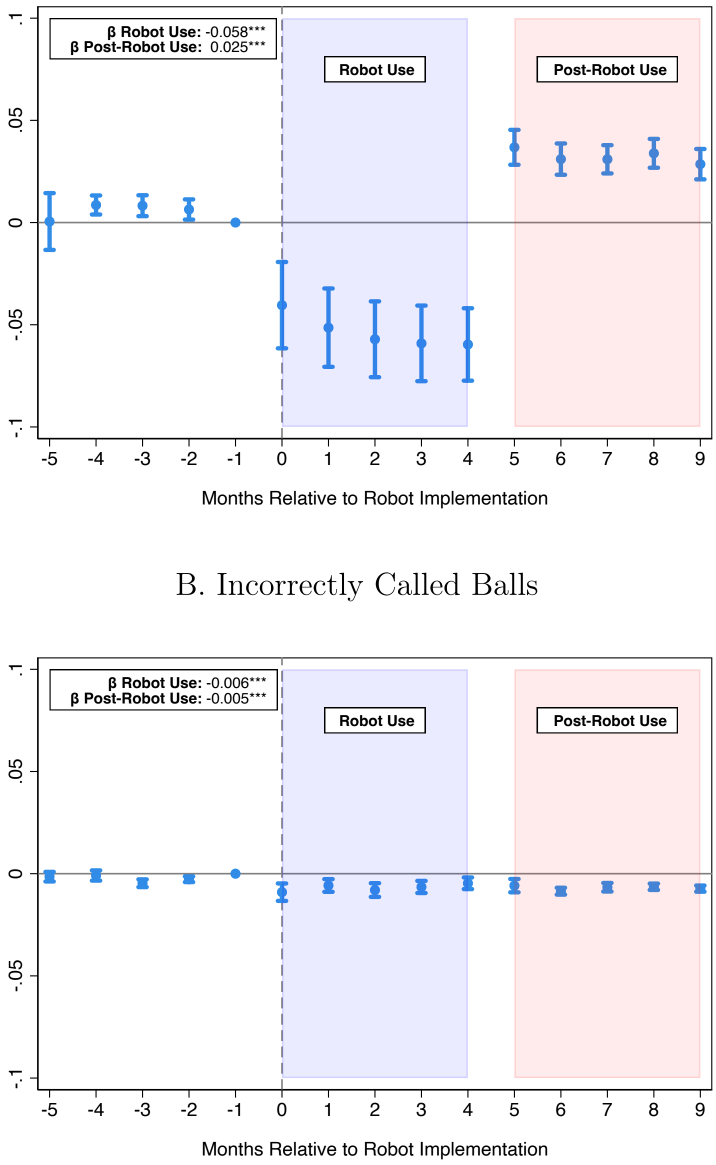
Notes: The figure describes potential lateral and vertical movements of umpires across leagues and robot implementation. In 2022, Triple-A Pacific Coast League implemented robot and 66 umpires (41 with more than 10 games with robot) are assisted by the robot. These umpires, in 2021, worked in Double-A leagues and Triple-A leagues. In the following year, in 2023, these umpires stayed in Triple-A and/or moved up to the Major League. In 2023, games held on Tuesdays to Thursdays used robot and games on Fridays to Sundays are called by umpires without robot. $D_{it}^{Robot} = 1$ indicates the games with robot and $D_{it}^{PostRobot} = 1$ indicates the games without robot in 2023.

Figure A.6: Pitch Distribution and Called Strike



Notes: X-axis is the pitch distance from the nearest border of the strike zone in feet and y-axis shows the average called strike rate. To the left of origin are pitches falling outside of the strike zone, and to the right are pitches falling inside the strike zone.

Figure A.7: Event-Study - Umpire Decisions

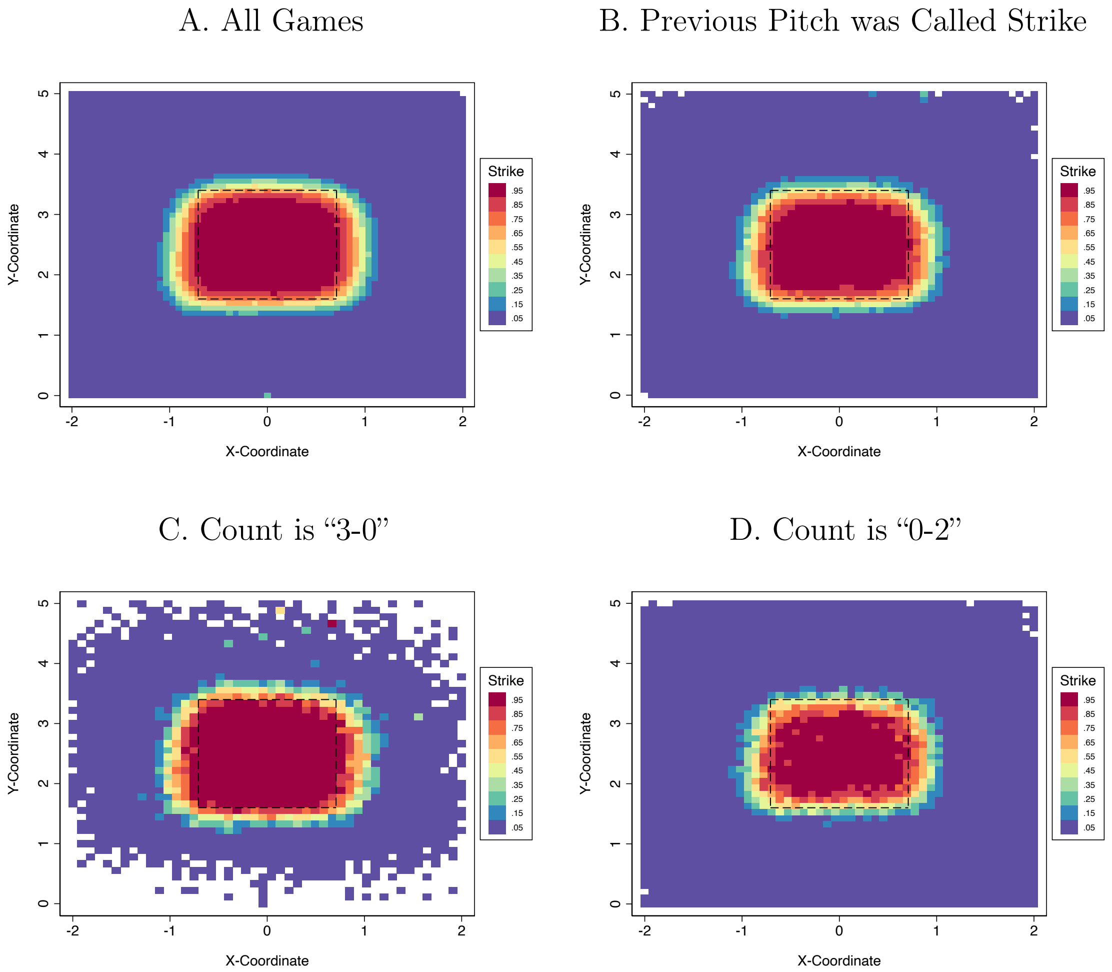


A. Incorrectly Called Strikes

B. Incorrectly Called Balls

Notes: All regressions include year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Use" indicate that robot is assisting umpires calling the game and "Post-Robot Use" indicate that the umpire returned following robot-assistance. X-axis is months relative to the first month of robot implementation. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. * p<0.1,** p<0.05, *** p<0.01.

Figure A.8: Called Strike Heatmaps - By Situations without Robot



A. All Games

B. Previous Pitch was Called Strike

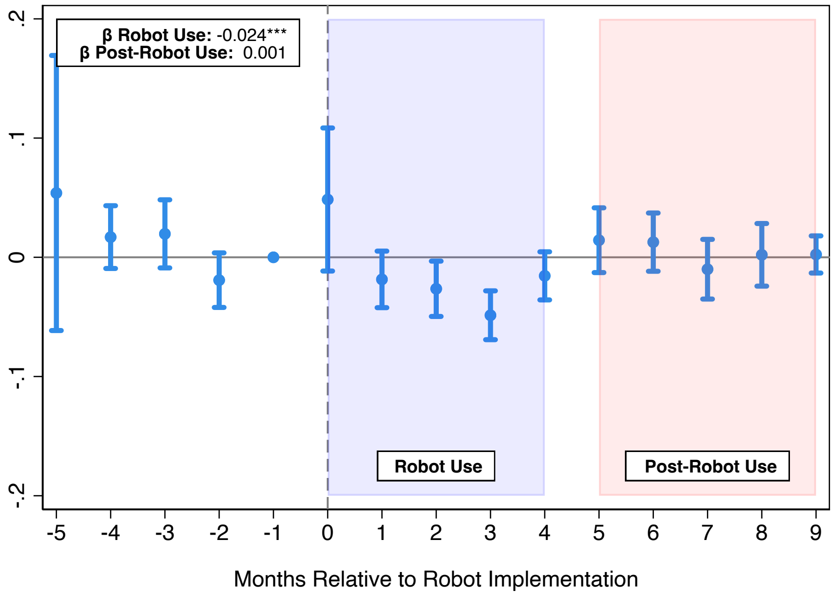
C. Count is “3-0”

D. Count is “0-2”

Notes: The figures plot the share of pitches that are called strike by pitch location and by game situations. The black dotted line shows the strike zone for an average batter.

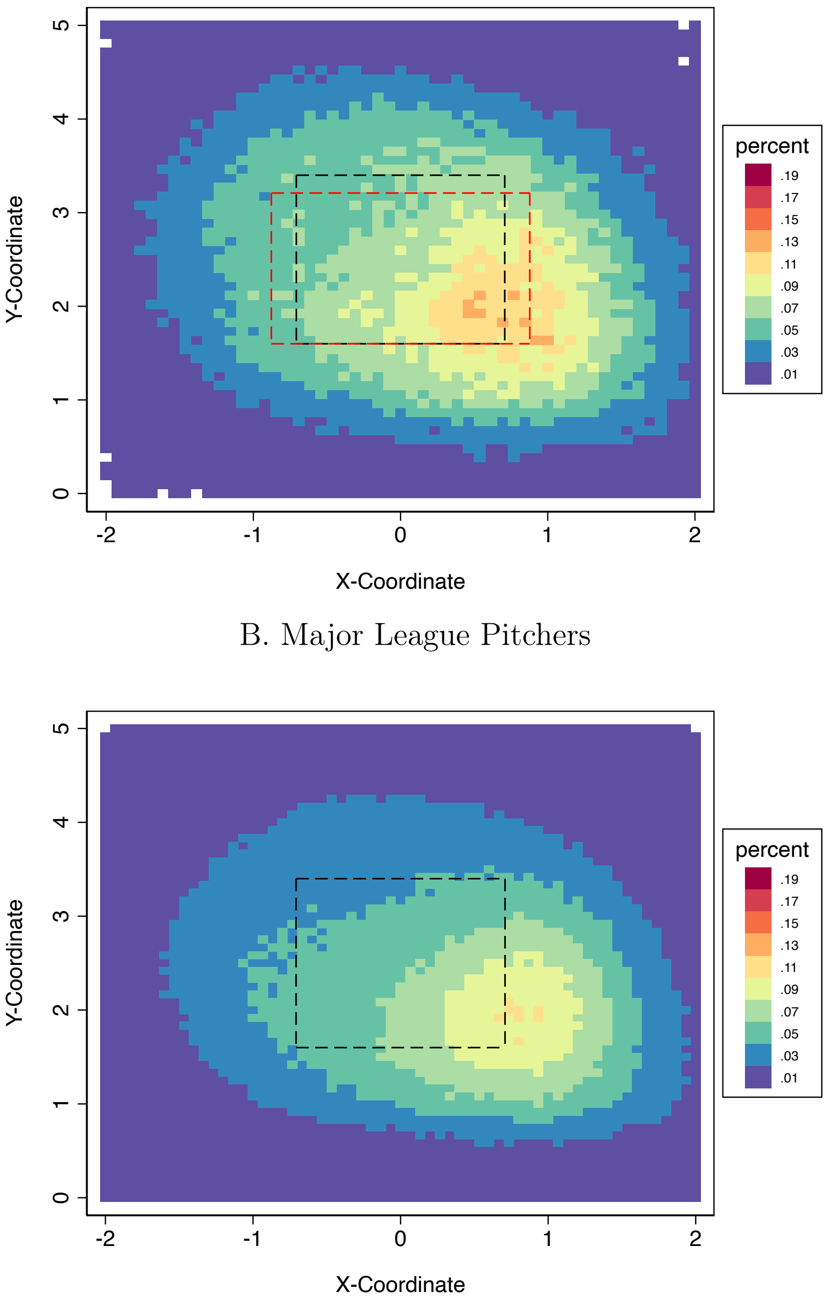
Figure A.9: Event-Study - Umpire Ejections

A. Ejections



Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Use" indicate that umpire is calling the game with robot-assistance and "Post-Robot Use" indicate that the human umpire returned following robot-assistance. X-axis is months relative to the first month of robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Figure A.10: Player Pitch Heatmaps

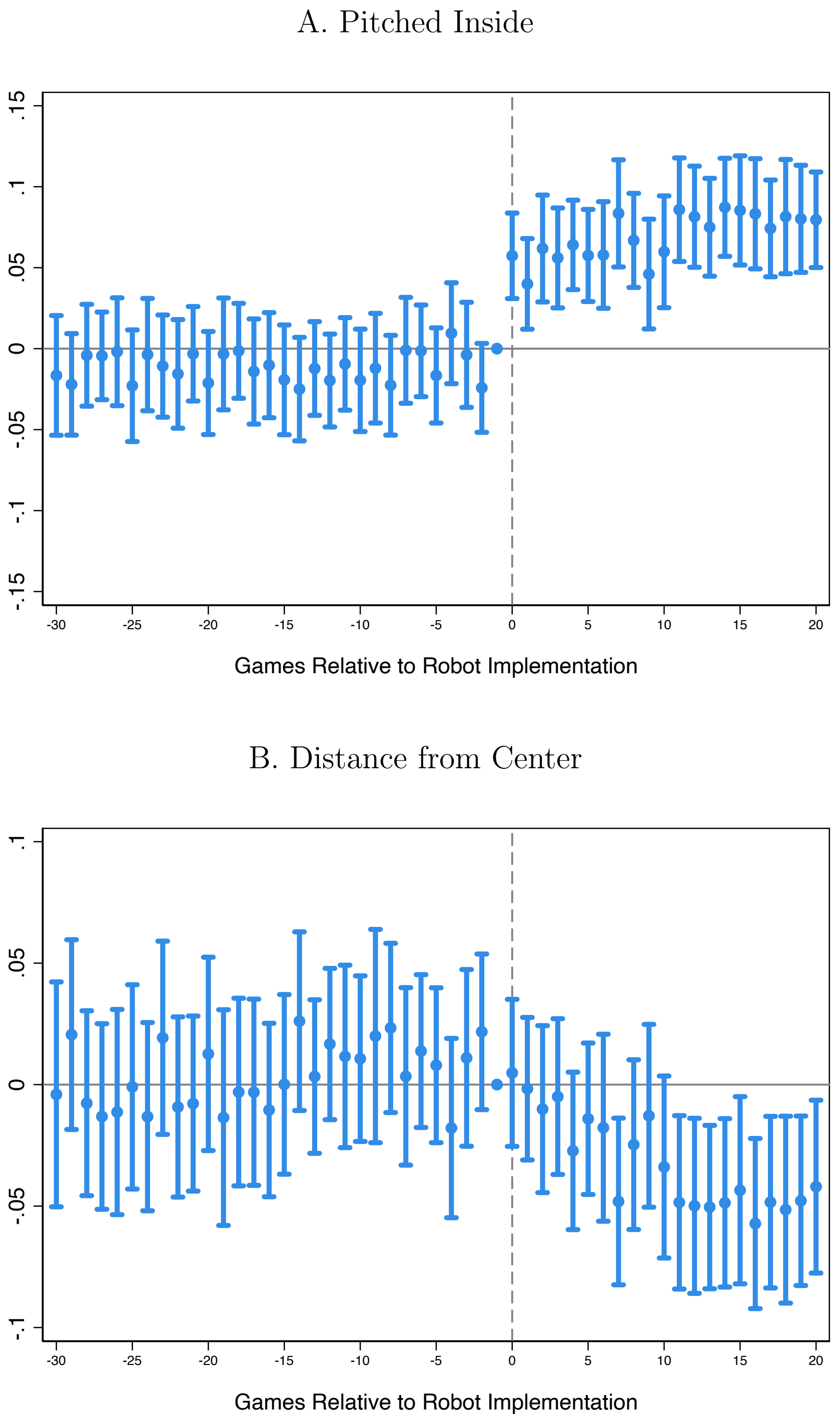


A. Pitchers Facing Umpire with Robot-Assistance

B. Major League Pitchers

Notes: The figures plot the share of pitches by pitch location of pitchers facing right-handed batters. The black dotted line shows the strike zone for an average batter and the red dotted line in Panel A shows the adjusted strike zone that was implemented on July 20, 2021 for games with robot.

Figure A.11: Event-Study - Pitchers Game-Level

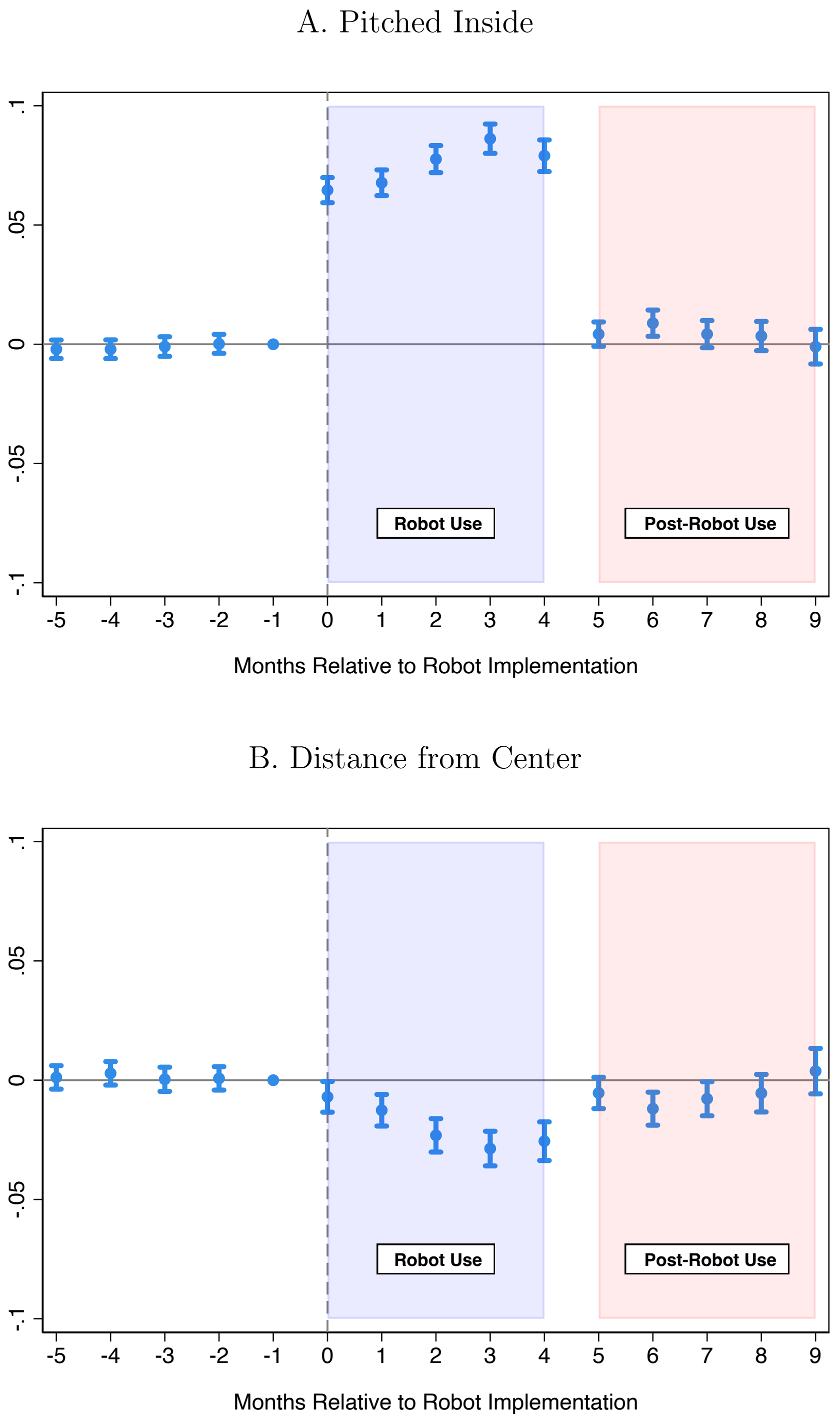


A. Pitched Inside

B. Distance from Center

Notes: All regressions include a vector of covariates at the pitch-level (excluding pitch location controls), year-by-month, umpire and player and team-by-year fixed effects. Standard errors are clustered at the player-level. "Pitched Inside" indicates that a pitch falls inside the strike zone. Distances are measured in feet. X-axis is number of games relative to the last game with robot.

Figure A.12: Event-Study - Pitchers Month-Level

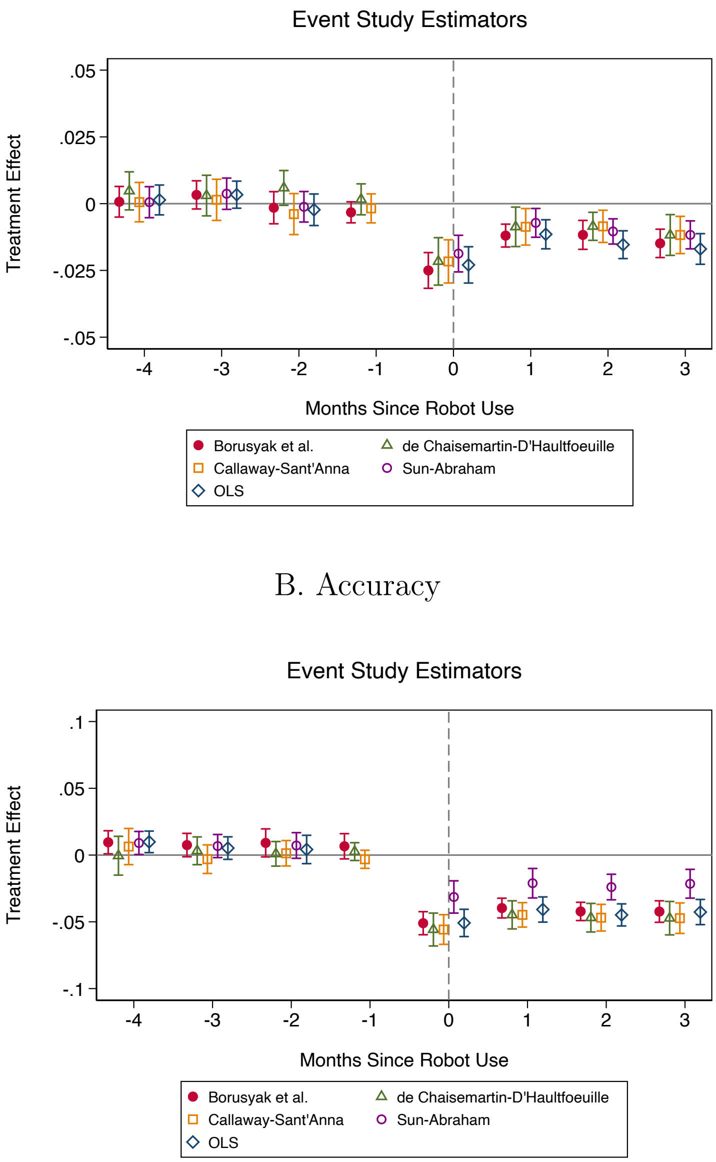


A. Pitched Inside

B. Distance from Center

Notes: All regressions include a vector of covariates at the pitch-level (excluding pitch location controls), year-by-month, umpire and player and team-by-year fixed effects. Standard errors are clustered at the player-level. “Pitched Inside” indicates that a pitch falls inside the strike zone. Distances are measured in feet. X-axis is months relative to the first month of robot implementation.

Figure A.13: Event-Study - Different Estimators

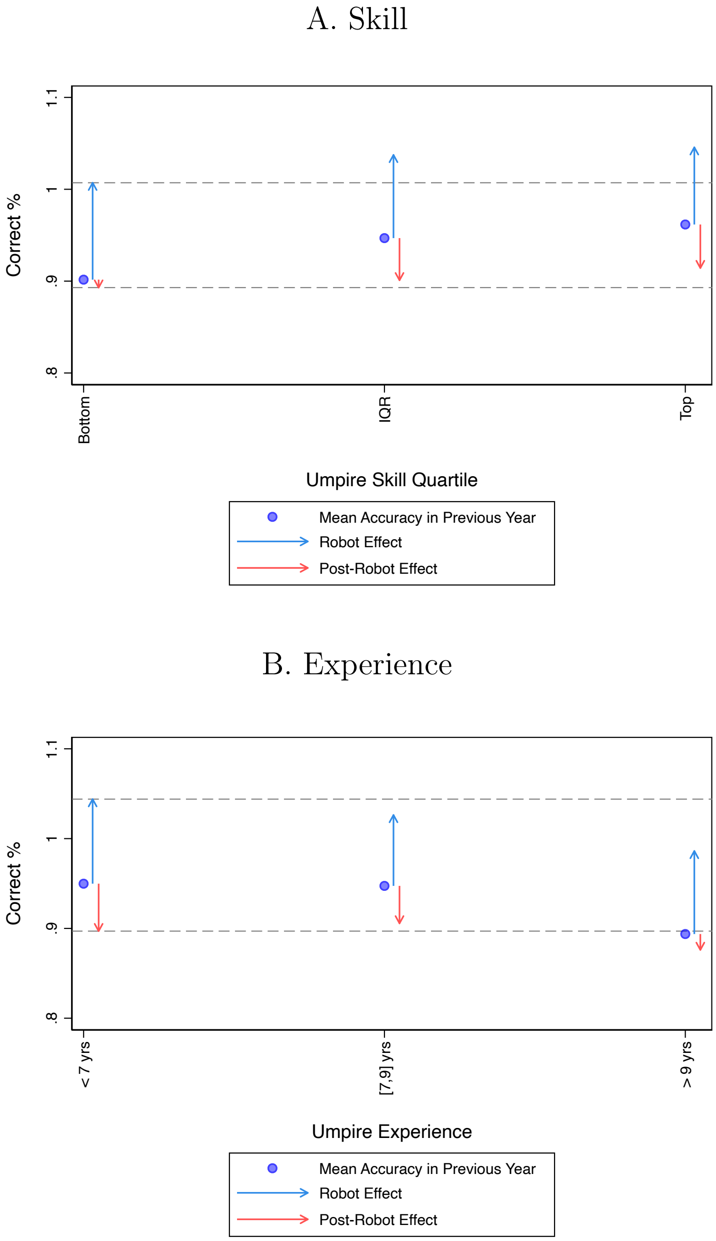


A. Residualized Accuracy

B. Accuracy

Notes: The figure plots coefficients from four different event study estimators and two-way fixed effects model (Borusyak et al., 2024; Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2021). X-axis is months relative to the last month of robot implementation. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. The outcomes in Panel A and B are decision accuracy residualized for pitch location and fixed-effects and raw decision accuracy. The outcomes are collapsed at the month-level.

Figure A.14: Does Skill Distribution Compress?

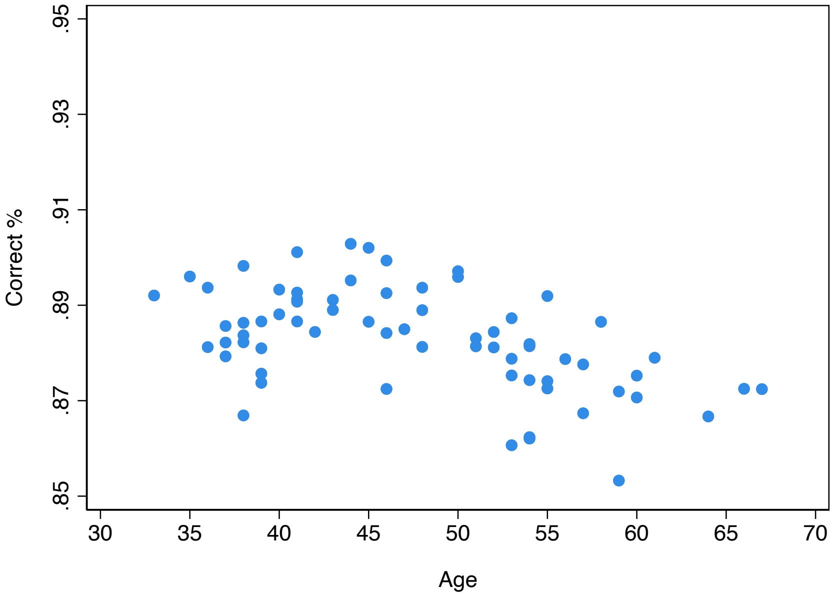

A. Skill

B. Experience

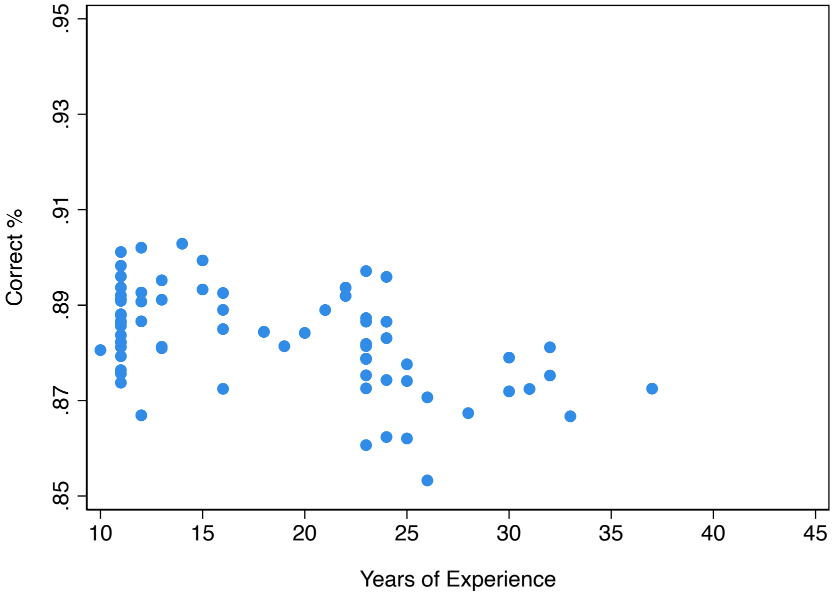
Notes: The figure plots the results in Table A.12 and Table A.14. All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. "Mean Accuracy" is measured in 2021. "Robot Effect" indicate the effect of robot assisting the umpire calling the game and "Post-Robot Effect." indicate that the effect of human umpire returning following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. I restrict the sample to never-treated umpires and umpires treated in Triple-A implementation in 2022. Umpires' skill are measured in 2021 and divided into quartiles. Umpires' years of experience are measured in 2021.

Figure A.15: Umpire Skill by Age and Experience

A. Age

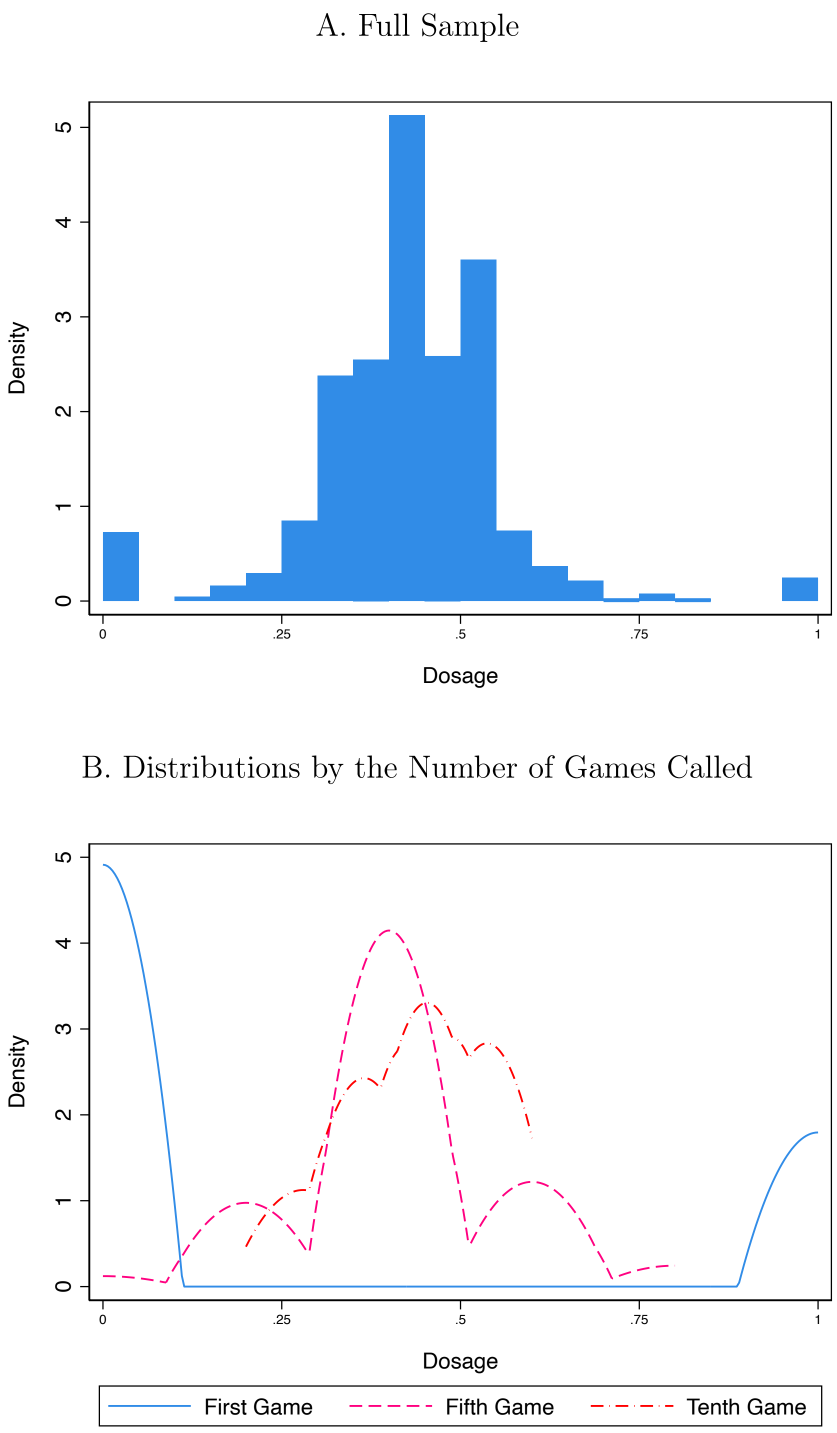


B. Experience



Notes: The figure plots the average accuracy of the Major League umpires in 2022 by age and years of experience.

Figure A.16: Treatment Intensity Distribution



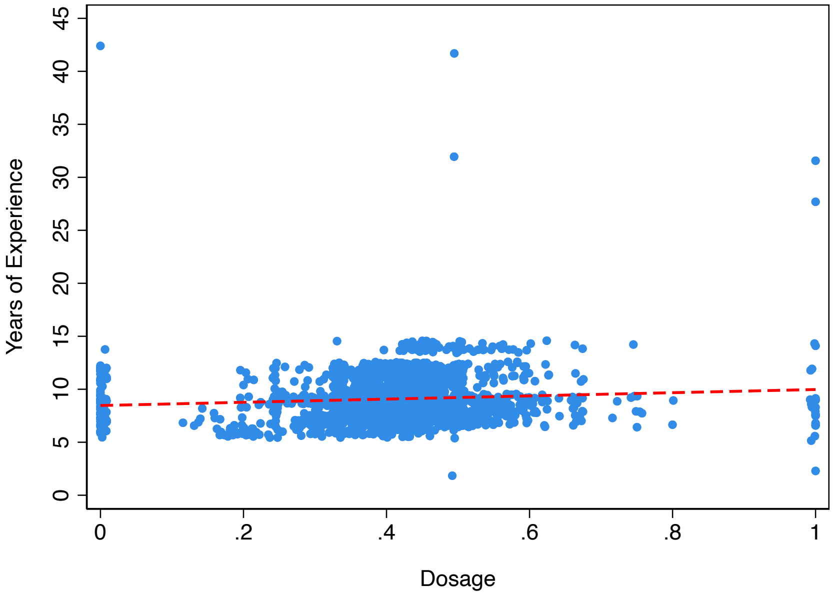
A. Full Sample

B. Distributions by the Number of Games Called

Notes: In 2023, games held on Tuesdays to Thursdays used robot and games on Fridays to Sundays are called without robot. “Dosage” is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$. Panel A plots the dosage for full sample of Triple-A league games in 2023 at the umpire-game level. Panel B plots the dosage for the umpire’s first, fifth and tenth game of the season.

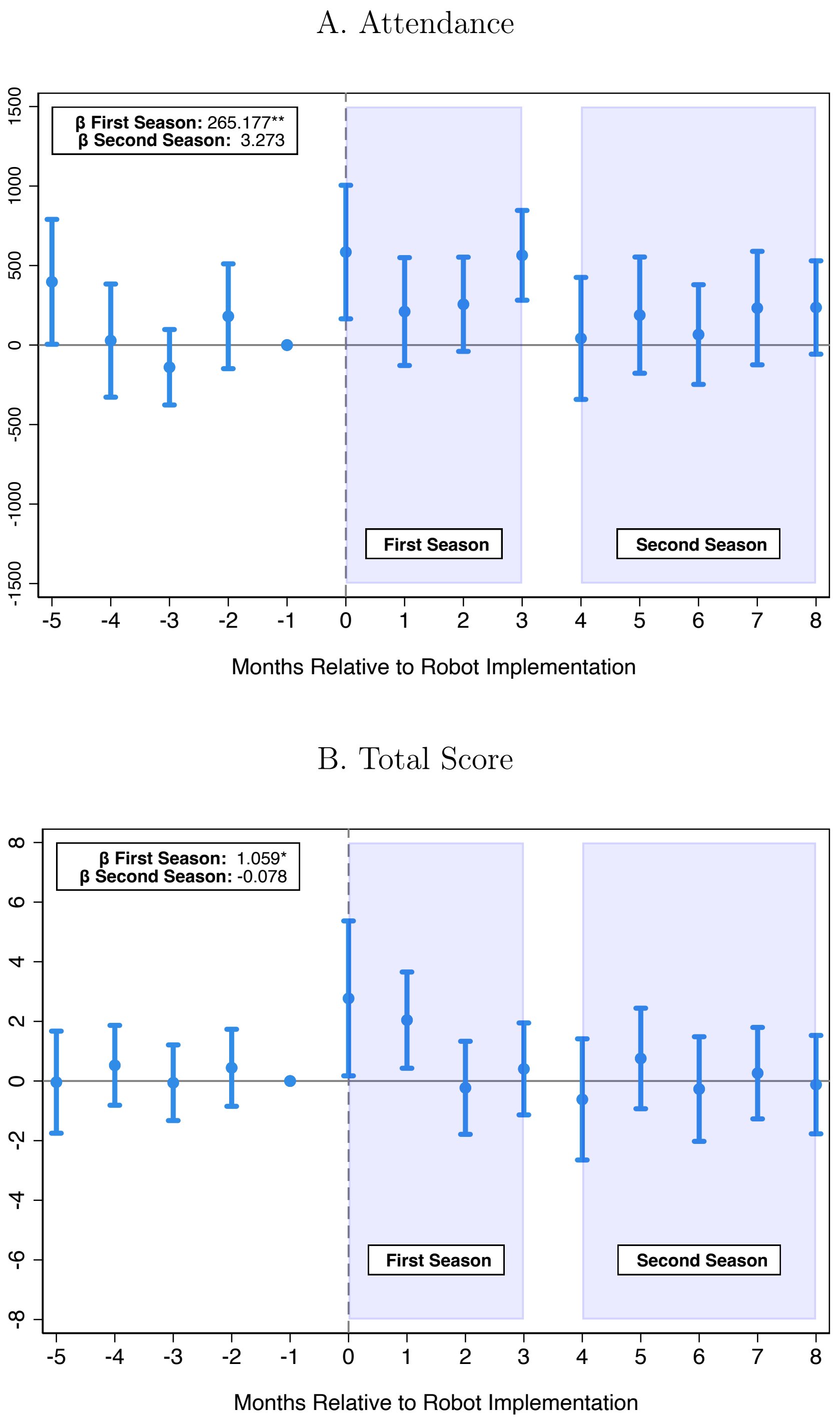
Figure A.17: Are Dosages Quasi-Random?

Years of Experience



Notes: In 2023, games held on Tuesdays to Thursdays used robot and games on Fridays to Sundays are called without robot. "Dosage" is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$. The figure plots the umpire's years of experience on the Y-axis and the dosage on the X-axis.

Figure A.18: Event-Study - Game Outcomes



A. Attendance

B. Total Score

Notes: All regressions include year-by-month and team fixed effects and league specific time trend. Standard errors are clustered at the team-level. "First Season' indicate that it is the first season of robot implementation and "Second Season' indicate that it is the second season of robot implementation. The sample includes games from Single-A Florida State and Carolina leagues. * $p<0.1$,** $p<0.05$, *** $p<0.01$.

Table A.1: Player Outcomes Summary Statistics

	Full Sample		With Robot		Without Robot	
	Mean	SD	Mean	SD	Mean	SD
Pitch Characteristics						
Pitch Distance from Center	0.991	(0.544)	1.072	(0.559)	0.988	(0.543)
Pitch Distance from Border	0.403	(0.282)	0.426	(0.344)	0.402	(0.278)
Pitched Inside	0.511	(0.500)	0.426	(0.494)	0.515	(0.500)
Pitched Edge	0.596	(0.491)	0.618	(0.486)	0.595	(0.491)
Batter Swung	0.489	(0.500)	0.478	(0.500)	0.489	(0.500)
Game Characteristics - Pitchers						
Pitcher Strike Out	4.718	(2.351)	4.584	(2.234)	4.724	(2.356)
Pitcher Hit Allowed	4.946	(2.254)	4.606	(2.228)	4.959	(2.253)
Pitcher Home Run Allowed	0.589	(0.802)	0.615	(0.817)	0.588	(0.801)
Pitcher Walk Allowed	1.952	(1.394)	2.196	(1.471)	1.942	(1.390)
Game Characteristics - Batters						
Batter Struck Out	0.959	(0.899)	1.025	(0.920)	0.956	(0.898)
Batter Hit	0.926	(0.892)	0.928	(0.895)	0.926	(0.891)
Batter Home Run	0.108	(0.330)	0.119	(0.346)	0.107	(0.329)
Batter Walked	0.428	(0.652)	0.543	(0.732)	0.423	(0.648)
Number of Games	66,297		2,731		63,566	
Number of Pitchers	7,821		1,948		7,657	
Number of Batters	7,496		1,537		7,383	

Notes: Distances are measured in feet. “Pitched Inside” indicates that a pitch falls inside the strike zone and “Pitched Edge” indicates that a pitch falls between 0.5 and 1.5 feet from the center of the strike zone. Outcomes in “Game Characteristics” panel are collapsed at the game-level. I restrict the sample to pitchers who faced at least 15 batters in an appearance and batters with at least 3 at-bats. A batter’s at-bat ends in a walk (or base on balls) when the count reaches 4 balls. Robot is used in games in Single-A Florida from 2021, in Triple-A Pacific Coast League in 2022 and in select games in both Triple-A leagues in 2023.

Table A.2: Summary Statistics - Are Robot-Assisted Umpires Different?

	Never-Treated Umpires		Robot-Assisted Umpires		Post-Robot Umpires	
	Mean	SD	Mean	SD	Mean	SD
Pitch Characteristics						
Called Strike	0.338	(0.473)	0.337	(0.473)	0.337	(0.473)
Called Correctly	0.953	(0.212)	0.953	(0.211)	0.953	(0.211)
Residualized Accuracy	0.0000	(0.2016)	0.0000	(0.2013)	0.0001	(0.2006)
Horizontal distance	0.740	(0.434)	0.737	(0.435)	0.739	(0.436)
Vertical distance	0.844	(0.567)	0.840	(0.557)	0.838	(0.557)
Game Characteristics						
Ejection by Umpire	0.070	(0.309)	0.061	(0.286)	0.062	(0.290)
Number of Games	15,067		10,750		9,630	
Number of Pitches	2,089,927		1,493,225		1,337,175	
Number of Umpires	313		70		62	

Notes: "Never-Treated Umpires" are those who have never been assisted by the robots, "Robot-Assisted Umpires" include those who have been assisted by the robots and "Post-Robot Umpires" include those who have been assisted and subsequently moved to another league without the robots. The statistics are from before any robot implementation and from the Minor Leagues. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. "Residualized accuracy" residualizes whether a decision was correct for pitch location and team-by-year fixed effect to account for pitch coordinates that may depend on calibration for each stadium and stringer plotting coordinates. Distances are in feet.

Table A.3: Do Umpires Lose Skill? - Restricting to More Ambiguous Decisions

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.123***	(0.018)	-0.037***	(0.005)	0.864	4,166,996
Correctly Called Strike	0.015***	(0.005)	0.004**	(0.002)	0.267	4,166,996
Correctly Called Ball	0.108***	(0.017)	-0.041***	(0.006)	0.597	4,166,996
Incorrectly Called Strike	-0.111***	(0.018)	0.047***	(0.007)	0.126	4,166,996
Incorrectly Called Ball	-0.012***	(0.002)	-0.011***	(0.001)	0.010	4,166,996
Called Strike	-0.097***	(0.017)	0.052***	(0.007)	0.393	4,166,996

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Sample is restricted to pitches that are within 0.5 feet from the nearest border of the strike zone from the outside and within 0.2 feet from the nearest border of the strike zone from the inside. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.4: Do Umpires Suffer More From Decision Biases?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Omission Bias (3-0)	-0.077***	(0.023)	0.029***	(0.008)	0.088	58,580
Omission Bias (0-2)	-0.002	(0.003)	-0.003**	(0.001)	0.010	62,490
Gambler's Fallacy	-0.012*	(0.006)	-0.003*	(0.002)	0.011	61,839

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. Outcomes are indicators that a call is incorrectly called given a situation. The outcome for "Omission Bias (3-0)" is that a pitch was incorrectly called strike when the count is 3-0. The outcome for "Omission Bias (0-2)" is that a pitch was incorrectly called ball when the count is 3-0. The outcome for "Gambler's Fallacy" is that a pitch was incorrectly called ball when the previous pitch was called strike. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.5: Summary Statistics - Are Umpires who Move Different?

	Umpires who Stayed		Moved Umpire		Promoted Umpire	
	Mean	SD	Mean	SD	Mean	SD
Pitch Characteristics						
Called Strike	0.330	(0.470)	0.328	(0.470)	0.325	(0.469)
Called Correctly	0.950	(0.219)	0.944	(0.230)	0.936	(0.244)
Residualized Accuracy	-0.0000	(0.2070)	0.0000	(0.2151)	-0.0002	(0.2253)
Horizontal distance	0.738	(0.432)	0.723	(0.424)	0.722	(0.424)
Vertical distance	0.837	(0.553)	0.846	(0.550)	0.862	(0.553)
Game Characteristics						
Ejection by Umpire	0.075	(0.353)	0.066	(0.293)	0.063	(0.287)
Number of Games	915		5,444		2,850	
Number of Pitches	132,937		782,866		407,161	
Number of Umpires	86		161		80	

Notes: "Umpires who Stayed" are umpires who have not moved between the leagues in 2022, "Moved Umpire" represents the umpires who primarily called games in different league than in 2022 and "Promoted Umpire" represents the umpires who moved up the league class from 2022. The statistics are from 2021 and from the Minor Leagues as 2020 season was canceled due to the COVID-19 pandemic. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. "Residualized accuracy" residualizes whether a decision was correct for pitch location and team-by-year fixed effect to account for pitch coordinates that may depend on calibration for each stadium and stringer plotting coordinates. Distances are in feet.

Table A.6: Are Results Sensitive to Different Control Groups?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Control Group: Umpires Switched League						
Correctly Called	0.062***	(0.009)	-0.025***	(0.004)	0.951	4,814,684
Correctly Called Strike	0.017***	(0.004)	-0.011***	(0.002)	0.294	4,814,684
Correctly Called Ball	0.045***	(0.007)	-0.014***	(0.003)	0.657	4,814,684
Incorrectly Called Strike	-0.056***	(0.009)	0.031***	(0.004)	0.042	4,814,684
Incorrectly Called Ball	-0.006***	(0.001)	-0.007***	(0.001)	0.007	4,814,684
Called Strike	-0.039***	(0.007)	0.021***	(0.003)	0.336	4,814,684
Control Group: Promoted Umpires						
Correctly Called	0.061***	(0.009)	-0.027***	(0.004)	0.951	3,879,268
Correctly Called Strike	0.019***	(0.004)	-0.011***	(0.002)	0.296	3,879,268
Correctly Called Ball	0.042***	(0.007)	-0.015***	(0.003)	0.656	3,879,268
Incorrectly Called Strike	-0.055***	(0.009)	0.034***	(0.004)	0.042	3,879,268
Incorrectly Called Ball	-0.006***	(0.001)	-0.007***	(0.001)	0.007	3,879,268
Called Strike	-0.036***	(0.007)	0.023***	(0.003)	0.338	3,879,268
Control Group: Single-A Umpires						
Correctly Called	0.061***	(0.009)	-0.031***	(0.004)	0.951	3,191,497
Correctly Called Strike	0.019***	(0.004)	-0.013***	(0.002)	0.297	3,191,497
Correctly Called Ball	0.042***	(0.007)	-0.018***	(0.003)	0.655	3,191,497
Incorrectly Called Strike	-0.054***	(0.009)	0.039***	(0.004)	0.042	3,191,497
Incorrectly Called Ball	-0.006***	(0.002)	-0.008***	(0.001)	0.007	3,191,497
Called Strike	-0.035***	(0.007)	0.026***	(0.003)	0.338	3,191,497
Control Group: Major League Umpires						
Correctly Called	0.065***	(0.008)	-0.027***	(0.003)	0.913	4,723,320
Correctly Called Strike	0.029***	(0.005)	-0.014***	(0.002)	0.266	4,723,320
Correctly Called Ball	0.036***	(0.005)	-0.013***	(0.002)	0.647	4,723,320
Incorrectly Called Strike	-0.057***	(0.008)	0.034***	(0.004)	0.077	4,723,320
Incorrectly Called Ball	-0.008***	(0.001)	-0.007***	(0.001)	0.010	4,723,320
Called Strike	-0.028***	(0.005)	0.020***	(0.002)	0.343	4,723,320

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Each panel uses different control groups: "Umpires Switched League" refers to any umpires who switched league between seasons; "Promoted Umpires" refers to umpires who moved up in the league system; "Single-A Umpires" and "Major League Umpires" are the umpires working in the lowest and highest system, respectively. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.7: Do Players Alter Their Behaviors?

	Treated Players					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Pitch Distance from Center	-0.016***	(0.003)	-0.008***	(0.003)	0.980	17,327,268
Pitch Distance from Border	-0.017***	(0.002)	-0.003**	(0.001)	0.404	17,327,268
Pitched Inside	0.070***	(0.002)	0.005**	(0.002)	0.527	17,327,268
Pitched Edge	0.001	(0.002)	0.002	(0.002)	0.588	17,327,268
Batter Swung	-0.014***	(0.003)	0.000	(0.003)	0.491	17,325,584

Notes: All regressions include a vector of covariates at the pitch-level, year-by-month, umpire and player and team-by-year fixed effects, otherwise specified. For distance outcomes, pitch location controls are omitted. Standard errors are clustered at the player-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. Distances are measured in feet. "Pitched Inside" indicates that a pitch falls inside the strike zone and "Pitched Edge" indicates that a pitch falls between 0.5 and 1.5 feet from the center of the strike zone. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.8: Do Umpires Lose Skill? - Reweighing for Player Responses

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.080***	(0.012)	-0.024***	(0.004)	0.931	8,685,514
Correctly Called Strike	0.013***	(0.004)	-0.005***	(0.001)	0.286	8,685,514
Correctly Called Ball	0.067***	(0.010)	-0.020***	(0.004)	0.645	8,685,514
Incorrectly Called Strike	-0.074***	(0.012)	0.030***	(0.004)	0.060	8,685,514
Incorrectly Called Ball	-0.006***	(0.001)	-0.006***	(0.001)	0.009	8,685,514
Called Strike	-0.061***	(0.010)	0.025***	(0.004)	0.346	8,685,514

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Regressions reweigh the data to have same distribution of pitches as the pitches in the Major League in 2021. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.9: Do Pitch-Level Controls Matter?

	Treated Umpires					
Main Specification	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
without Pitch-level Controls						
Correctly Called	0.051***	(0.008)	-0.015***	(0.002)	0.932	8,864,801
with Pitch-level Controls						
Correctly Called	0.064***	(0.009)	-0.020***	(0.003)	0.932	8,864,667
Dosage Specification	With Robot	S.E.	Dosage	S.E.	Baseline Mean	N
without Pitch-level Controls						
Correctly Called	0.041***	(0.001)	-0.017**	(0.007)	0.885	330,188
with Pitch-level Controls						
Correctly Called	0.052***	(0.001)	-0.020***	(0.007)	0.885	330,185

Notes: All regressions for pitch-level outcomes include year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Treated Games" indicate that robot is assisting umpires calling the game and "Dosage" is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the data before 2021, the year of first robot implementation for the Main Specification panel. Baseline mean is calculated using the first game of the year without robot-assistance for the Dosage Specification panel. * p<0.1,** p<0.05, *** p<0.01.

Table A.10: Does Data Quality Matter?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.033***	(0.006)	-0.015***	(0.002)	0.919	8,815,175
Correctly Called Strike	0.038***	(0.007)	-0.020***	(0.003)	0.267	8,815,175
Correctly Called Ball	-0.005	(0.004)	0.005***	(0.001)	0.652	8,815,175
Incorrectly Called Strike	-0.035***	(0.005)	0.017***	(0.003)	0.075	8,815,175
Incorrectly Called Ball	0.002	(0.002)	-0.002***	(0.000)	0.006	8,815,175
Called Strike	0.003	(0.004)	-0.003**	(0.001)	0.342	8,815,175

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. I use the manually-plotted coordinates for all pitches. Standard errors are clustered at the umpire-level. “Robot Mo.” indicate that robot is assisting umpires calling the game and “Post-Robot Mo.” indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.11: Do Umpires Lose Skill? - Restricting to Triple-A and Major League Umpires

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.089***	(0.003)	-0.036***	(0.003)	0.913	1,920,905
Correctly Called Strike	0.046***	(0.003)	-0.021***	(0.002)	0.261	1,920,905
Correctly Called Ball	0.043***	(0.003)	-0.015***	(0.002)	0.652	1,920,905
Incorrectly Called Strike	-0.081***	(0.003)	0.044***	(0.003)	0.079	1,920,905
Incorrectly Called Ball	-0.008***	(0.001)	-0.008***	(0.001)	0.008	1,920,905
Called Strike	-0.035***	(0.004)	0.024***	(0.002)	0.340	1,920,905

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. I restrict the sample to the Major League umpires and umpires treated in Triple-A implementation in 2022. Baseline mean is calculated using the data before 2022, the year of robot implementation for the Triple-A umpires. * p<0.1,** p<0.05, *** p<0.01.

Table A.12: Do High-Skilled Umpires Lose More Skill?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Bottom Quartile						
Correctly Called	0.105***	(0.009)	-0.009	(0.006)	0.901	459,518
Correctly Called Strike	0.058***	(0.011)	-0.005	(0.004)	0.250	459,518
Correctly Called Ball	0.047***	(0.009)	-0.004	(0.004)	0.651	459,518
Incorrectly Called Strike	-0.096***	(0.005)	0.011	(0.007)	0.090	459,518
Incorrectly Called Ball	-0.010*	(0.005)	-0.003*	(0.001)	0.009	459,518
Called Strike	-0.038***	(0.011)	0.007	(0.005)	0.340	459,518
IQR						
Correctly Called	0.091***	(0.007)	-0.046***	(0.005)	0.948	353,885
Correctly Called Strike	0.036***	(0.004)	-0.022***	(0.003)	0.287	353,885
Correctly Called Ball	0.054***	(0.007)	-0.024***	(0.003)	0.661	353,885
Incorrectly Called Strike	-0.082***	(0.005)	0.056***	(0.006)	0.048	353,885
Incorrectly Called Ball	-0.009**	(0.003)	-0.010***	(0.001)	0.004	353,885
Called Strike	-0.046***	(0.006)	0.034***	(0.004)	0.335	353,885
Top Quartile						
Correctly Called	0.084***	(0.008)	-0.047***	(0.002)	0.961	232,069
Correctly Called Strike	0.043***	(0.008)	-0.032***	(0.003)	0.291	232,069
Correctly Called Ball	0.041***	(0.010)	-0.015***	(0.003)	0.670	232,069
Incorrectly Called Strike	-0.063***	(0.006)	0.061***	(0.003)	0.034	232,069
Incorrectly Called Ball	-0.021***	(0.007)	-0.014***	(0.001)	0.004	232,069
Called Strike	-0.019*	(0.009)	0.029***	(0.003)	0.325	232,069

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. “Robot Mo.” indicate that robot is assisting umpires calling the game and “Post-Robot Mo.” indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. I restrict the sample to never-treated umpires and umpires treated in Triple-A implementation in 2022. Umpires’ skill are measured in 2021 and divided into quartiles. Baseline mean is calculated using the data before 2022, the year of robot implementation for the Triple-A umpires. * p<0.1,** p<0.05, *** p<0.01.

Table A.13: Do Umpires Improve with Experience?

	Experience	S.E.	Outcome Mean	N
All Umpires				
Accuracy	0.0007	(0.0005)	0.9162	893
$\mathbb{1}(Ejection)$	0.0007	(0.0008)	0.0566	893
Number of Ejection	0.0002	(0.0010)	0.0685	893
Umpires Experience $\leq$ 5 yrs.				
Accuracy	0.0020***	(0.0006)	0.9467	317
$\mathbb{1}(Ejection)$	0.0040*	(0.0023)	0.0593	317
Number of Ejection	0.0040	(0.0029)	0.0699	317
Umpires Experience $\in$ (5,10] yrs.				
Accuracy	-0.0052***	(0.0017)	0.9216	208
$\mathbb{1}(Ejection)$	-0.0008	(0.0026)	0.0650	208
Number of Ejection	-0.0017	(0.0038)	0.0808	208

Notes: All regressions for pitch-level outcomes include umpire fixed effects. Standard errors are clustered at the umpire-level. Experience is measured in years. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Outcomes are collapsed at the umpire-by-year level. * p<0.1,** p<0.05, *** p<0.01.

Table A.14: Do More-Experienced Umpires Lose More Skill?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Umpire Experience $\leq$ 6 years						
Correctly Called	0.094***	(0.008)	-0.053***	(0.003)	0.953	157,135
Correctly Called Strike	0.014*	(0.008)	-0.031***	(0.002)	0.288	157,135
Correctly Called Ball	0.080***	(0.011)	-0.022***	(0.002)	0.666	157,135
Incorrectly Called Strike	-0.084***	(0.008)	0.064***	(0.002)	0.042	157,135
Incorrectly Called Ball	-0.010***	(0.003)	-0.011***	(0.001)	0.004	157,135
Called Strike	-0.069***	(0.011)	0.033***	(0.002)	0.330	157,135
Umpire Experience $\in$ (6, 9] years						
Correctly Called	0.079***	(0.004)	-0.042***	(0.003)	0.948	460,100
Correctly Called Strike	0.039***	(0.004)	-0.022***	(0.002)	0.284	460,100
Correctly Called Ball	0.040***	(0.005)	-0.020***	(0.003)	0.664	460,100
Incorrectly Called Strike	-0.074***	(0.004)	0.052***	(0.004)	0.046	460,100
Incorrectly Called Ball	-0.005**	(0.002)	-0.011***	(0.001)	0.005	460,100
Called Strike	-0.035***	(0.005)	0.031***	(0.003)	0.331	460,100
Umpire Experience $>$ 9 years						
Correctly Called	0.092***	(0.008)	-0.018***	(0.005)	0.895	1,100,884
Correctly Called Strike	0.058***	(0.006)	-0.009***	(0.003)	0.249	1,100,884
Correctly Called Ball	0.035***	(0.008)	-0.009***	(0.003)	0.646	1,100,884
Incorrectly Called Strike	-0.084***	(0.008)	0.021***	(0.006)	0.096	1,100,884
Incorrectly Called Ball	-0.008***	(0.002)	-0.003***	(0.001)	0.009	1,100,884
Called Strike	-0.026***	(0.008)	0.012***	(0.004)	0.345	1,100,884

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. I restrict the sample to the Major League umpires and umpires treated in Triple-A implementation in 2022. Umpires' years of experience are measured in 2021. Baseline mean is calculated using the data before 2022, the year of robot implementation for the Triple-A umpires. * $p<0.1$,** $p<0.05$, *** $p<0.01$.

Table A.15: Do Umpires Lose More Skill if They Are Further Away from Promotion?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Umpire without MLB Experience						
Correctly Called	0.082***	(0.005)	-0.049***	(0.002)	0.953	512,674
Correctly Called Strike	0.028***	(0.004)	-0.026***	(0.002)	0.290	512,674
Correctly Called Ball	0.054***	(0.006)	-0.023***	(0.002)	0.664	512,674
Incorrectly Called Strike	-0.073***	(0.004)	0.060***	(0.002)	0.042	512,674
Incorrectly Called Ball	-0.009***	(0.003)	-0.011***	(0.001)	0.004	512,674
Called Strike	-0.045***	(0.006)	0.035***	(0.002)	0.332	512,674
Umpire with MLB Experience						
Correctly Called	0.093***	(0.005)	-0.015***	(0.004)	0.897	1,205,489
Correctly Called Strike	0.053***	(0.006)	-0.009***	(0.003)	0.249	1,205,489
Correctly Called Ball	0.040***	(0.005)	-0.007***	(0.003)	0.647	1,205,489
Incorrectly Called Strike	-0.085***	(0.005)	0.020***	(0.005)	0.094	1,205,489
Incorrectly Called Ball	-0.007***	(0.002)	-0.004***	(0.001)	0.009	1,205,489
Called Strike	-0.032***	(0.006)	0.011***	(0.003)	0.343	1,205,489

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. “Robot Mo.” indicate that robot is assisting umpires calling the game and “Post-Robot Mo.” indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. I restrict the sample to the Major League umpires and umpires treated in Triple-A implementation in 2022. Umpires’ prior experience in the Major League is measured in 2021. Baseline mean is calculated using the data before 2022, the year of robot implementation for the Triple-A umpires. * p<0.1,** p<0.05, *** p<0.01.

Table A.16: 2023 Triple-A Implementation Summary Statistics

	Full Sample		With Robot		Without Robot	
	Mean	SD	Mean	SD	Mean	SD
Pitch Characteristics						
Called Strike	0.310	(0.462)	0.308	(0.462)	0.311	(0.463)
Called Correctly	0.918	(0.275)	0.940	(0.237)	0.901	(0.299)
Residualized Accuracy	-0.0000	(0.2630)	0.0297	(0.2316)	-0.0220	(0.2820)
Horizontal distance	0.763	(0.490)	0.757	(0.488)	0.768	(0.492)
Vertical distance	0.839	(0.575)	0.836	(0.573)	0.842	(0.576)
Game Characteristics						
Ejection by Umpire	0.056	(0.283)	0.043	(0.234)	0.065	(0.314)
Number of Games	2,126		888		1,238	
Number of Pitches	330,188		140,588		189,600	
Number of Umpires	71		64		70	
Dosage	0.420	(0.137)				

Notes: A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. "Residualized accuracy" residualizes whether a decision was correct for pitch location and team-by-year fixed effect to account for pitch coordinates that may depend on calibration for each stadium and stringer plotting coordinates. Distances are in feet. Robot is used in games played between Tuesday and Thursday in both Triple-A leagues in 2023. "Dosage" is $\frac{\text{\# of games with robot}}{\text{\# of total games}}$.

Table A.17: 2020 COVID Pause Summary Statistics

	Never-Returned Umpires		Returned Umpires	
	Mean	SD	Mean	SD
Pitch Characteristics				
Called Strike	0.334	(0.472)	0.333	(0.471)
Called Correctly	0.957	(0.203)	0.952	(0.213)
Residualized Accuracy	-0.0001	(0.1945)	0.0003	(0.2034)
Horizontal distance	0.745	(0.437)	0.738	(0.431)
Vertical distance	0.849	(0.575)	0.857	(0.575)
Game Characteristics				
Ejection by Umpire	0.062	(0.279)	0.068	(0.304)
Number of Games	1,382		3,358	
Number of Pitches	193,254		465,809	
Number of Umpires	83		84	

Notes: A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. Distances are in feet. “Never-Returned Umpires” are those who have not returned to call games since the COVID-19 pandemic and “Returned Umpires” are those who have returned from the COVID-19 canceled season. The statistics are from 2019 before the COVID-19 canceled 2020 season and from the Minor Leagues.

Table A.18: Are Umpires Following the Other Guideline?

	Treated Umpires					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitch-level Outcomes						
Correctly Called	0.038***	(0.006)	-0.009***	(0.002)	0.932	8,864,667
Correctly Called Strike	0.013***	(0.004)	-0.006***	(0.001)	0.281	8,864,667
Correctly Called Ball	0.025***	(0.003)	-0.004**	(0.002)	0.651	8,864,667
Incorrectly Called Strike	-0.024***	(0.004)	0.007***	(0.001)	0.059	8,864,667
Incorrectly Called Ball	-0.014***	(0.002)	0.002***	(0.001)	0.008	8,864,667
Called Strike	-0.011***	(0.003)	0.002	(0.002)	0.340	8,864,667

Notes: All regressions for pitch-level outcomes include a vector of covariates at the pitch-level, year-by-month, umpire and team-by-year fixed effects. Standard errors are clustered at the umpire-level. "Robot Mo." indicate that robot is assisting umpires calling the game and "Post-Robot Mo." indicate that the umpire returned following robot-assistance. A pitch is correctly called if it crosses the strike zone and called strike or missed the strike zone and called ball. A pitch is incorrectly called if it misses the strike zone but called strike or crosses the strike zone but called ball. Accuracy is measured using the alternate strike zone implemented for the robot-assisted games on July 20, 2021. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.19: Do the Leagues Benefit from Robot Implementation?

	Treated Team					
	First Season	S.E.	Second Season	S.E.	Baseline Mean	N
Game-level Outcomes						
Attendance	265.2**	(123.1)	3.3	(121.2)	2135.2	6,826
Duration (in min.)	21.1***	(2.9)	15.5***	(2.9)	166.2	7,219
Num. of Pitches	11.2***	(2.7)	0.5	(3.6)	272.4	7,220
Total Score	1.1*	(0.5)	-0.1	(0.5)	8.5	7,220
Delay	-17.9	(24.3)	-16.8	(20.3)	64.2	845

Notes: All regressions include year-by-month and team fixed effects and league specific time trend. Standard errors are clustered at the team-level. “First Season’ indicate that it is the first season of robot implementation and “Second Season’ indicate that it is the second season of robot implementation. The sample includes games from Single-A Florida and Carolina leagues. Baseline mean is calculated using the data from 2019, the season before robot implementation. * p<0.1,** p<0.05, *** p<0.01.

Table A.20: Do Players Perform Better Following Robot Implementation?

	Treated Players					
	Robot Mo.	S.E.	Post-Robot Mo.	S.E.	Baseline Mean	N
Pitcher Outcomes						
Pitcher Strike Out	-0.015	(0.072)	-0.129	(0.080)	4.650	118,876
Pitcher Walk Allowed	0.125***	(0.046)	-0.030	(0.048)	1.908	118,876
Pitcher Hit Allowed	-0.039	(0.070)	-0.040	(0.059)	5.170	118,876
Pitcher Home Run Allowed	0.039*	(0.023)	0.013	(0.018)	0.559	118,876
Batter Outcomes						
Batter Struck Out	0.004	(0.013)	0.027*	(0.014)	0.914	1,089,716
Batter Walked	0.031***	(0.009)	-0.013	(0.009)	0.393	1,089,716
Batter Hit	-0.002	(0.009)	-0.024***	(0.009)	0.938	1,089,716
Batter Home Run	0.009**	(0.004)	0.000	(0.003)	0.099	1,089,716

Notes: All regressions include year-by-month, umpire and player and team-by-year fixed effects. Standard errors are clustered at the player-level. “Robot Mo.” indicate that robot is assisting umpires calling the game and “Post-Robot Mo.” indicate that the players are playing in games called by umpires following robot-assistance. The outcomes are collapsed at the player-by-game level. I restrict the sample to pitchers who faced at least 15 batters in an appearance and batters with at least 3 at-bats. A batter’s at-bat ends in a walk (or base on balls) when the count reaches 4 balls. Baseline mean is calculated using the data before 2021, the year of first robot implementation. * $p<0.1$,** $p<0.05$, *** $p<0.01$.

A.2 Baseball Game Description

Baseball is often referred to as America's pastime and has captivated fans for over a century. In this section, I briefly describe a typical baseball game.

Setup A baseball game is played between two teams, each consisting of nine players. One team takes on the role of the batting team, while the other becomes the fielding team. The objective of the batting team is to score runs, while the fielding team aims to prevent runs. The field is divided into four bases arranged in a diamond pattern: first base, second base, third base, and home plate.

Game The game begins with a pitcher, standing on the pitcher's mound, throwing a pitch. The pitcher's objective is to throw the baseball to the batter while trying to prevent them from hitting it. The batter stands at home plate, ready to swing at any pitch that comes his way. The batter's goal is to hit the pitched ball and safely reach a base. To score a run, a batter must make a complete circuit around all four bases, returning to home plate.

A game consists of nine innings, with each team having the opportunity to bat and field. Each inning is divided into two halves: the top and bottom. The team with the most runs at the end of the nine innings wins.

Inning An inning refers to the period of play. A standard baseball game comprises nine innings, although extra innings can occur if the game is tied after the regulation nine innings. In the top half of the inning, the team (usually an away team) is designated as the "batting team" and gets its chance to score runs. Each player gets a chance to bat in a specific order ("batting lineup") until three outs are accumulated. At the end of the top half of an inning, there is a brief break as the teams switch roles. At the end of the bottom half of an inning, another short break occurs as the teams switch roles again. This pattern continues until nine innings are completed.

At-Bat An “at-bat” refers to the specific plate appearance of a batter during a game. Each at-bat involves a sequence of pitches thrown by the pitcher, and the count represents the number of balls and strikes during this plate appearance. The count consists of two numbers. The first number represents the number of balls, and the second number represents the number of strikes. The count starts with 0 balls and 0 strikes. A ball is called when the pitcher throws a pitch outside the strike zone, and the batter does not swing. A strike is called when the pitcher throws a pitch inside the strike zone and the batter does not swing, or the batter swings and misses. If a batter accumulates four balls, he is awarded first base, a situation known as a “walk.” If a batter accumulates three strikes, he is called “out.”

A.3 More Details about Robot Implementation

In 2021, the Major League successfully implemented an automated ball-and-strike system at Single-A Florida State League. The system expanded to the Triple-A Pacific Coast League in 2022 and the entire Triple-A in 2023. The technology is used in two different systems: full ABS in which the game is called with the robot and a "Challenge" system.

Challenge System In the Single-A Florida State League, in 2022, the Major League explored an alternative system called a "Challenge" system. In select games with the challenge system, umpires call pitches as before without the robot, but the players (pitcher, catcher and batter) can appeal these calls and check with the robot. Each team were given three challenges in a game and if an appeal is successful, then the team retains the challenge.[1,2] The challenge system expanded to Triple-A leagues in 2023 and was also used in select games.[3]

Data In the data, I observe the number of successful and unsuccessful challenges in these select games, but I do not know the exact pitch that resulted in a challenge. I observe the ultimate call on the pitch.

In an average game, the umpire calls about 150 pitches. The average number of challenges is about 4.4 per game and the average success rate is about 33%. Therefore, about 1.4 pitches are overturned in a game with the challenge system (or about 0.9%). The average accuracy in these games is 90.1%, but with 1.4 overturned pitches, the accuracy will be about 89.2%. The main result of the paper is that the umpires' accuracy declines following the implementation of robots. If anything, the challenge system biases me away from finding the negative results.

Sample Restriction To account for the alternating implementation of the challenge system and full robot-assisted games, I impose the following sample restrictions.

[1] A challenge must be made within 2 seconds of the call.

[2] The average number of challenges is about 4.4 per game and the average success rate is about 33%.

[3] Games held on Fridays to Sundays used the challenge system.

In the main difference-in-differences specification (Equation 1.1), I drop the games with the full robot system in the second year of the implementation. For example, if an umpire worked in the Triple-A Pacific in 2022 and continued to work in the Triple-A Leagues in 2023, I drop the full robot-assisted games in 2023. This restriction is made to estimate the effect of robot implementation on human skills.

In the dosage model (Equation 1.3), I exploit the alternating system and keep all these games.

A.4 Pitch Coordinates

The Major League provides detailed pitch-level data including the X and Y coordinates for every thrown pitch. There are two different coordinate systems: manually-plotted and pitch-tracking technology-plotted coordinates.

A.4.1 Manually-Plotted Pitch Coordinates

The league hires "stringers" who digitally score games to provide data that are available on the MLB Stats API. Most stadiums hire about three stringers and each work about 25-30 games in a season. New stringers go through an 8-10 weeks training program and score practice games with the more experienced stringers before working alone. The main task of these stringers is to enter the results of every pitch and play. They work at the press box in each stadium with the Major League representative who oversees the job. In particular, to record pitch coordinates, the stringer "clicks" on the software and records where he sees the pitch go through the plate. These are, therefore, recorded in the pixel coordinate system.

A.4.2 Technology-Plotted Pitch Coordinates

Starting in 2003, the Major League started installing technology that tracks pitches in Major League stadiums. "Questec" was used to measure umpires' performances, yet it was criticized for reliability.

In 2005, the league started installing cameras in stadiums to track pitches and completed in 2007. "PITCHf/x" used three cameras, strategically located in the stadium, to track pitches as it goes through the strike zone. Sportvision, the company that developed PITCHf/x, claims that the system tracks pitches with an accuracy of better than one inch.

In 2017, the league replaced "PITCHf/x" with "TrackMan" which uses cameras and Doppler radar to track pitches. And in 2020, "TrackMan" was replaced with "Hawk-Eye" which uses the optical-tracking and vision-processing system that enables tracking pitches with an accuracy of better than 0.1 inch.

The pitch coordinates from these systems are recorded as X and Y coordinates in feet.

A.4.3 Transforming Coordinates

To create a dataset with a uniform coordinate system, I convert the manually-plotted pitch coordinates in pixels into feet. First, each stringer might have a unique bias that could depend on the vantage point. To account for this, I transform the coordinates at the team(stadium)-by-year level.[4] At the team-by-year level, I generate a heatmap of umpires' ball-strike decisions and find the corners of the strike zone. In particular, I find the region where the pitches are called strike more than 95% of the time and take the top, bottom, left and right points. Using these four points, I linearly transform the pitch coordinates into feet-unit as I know the four corners of the strike zone in feet.

All of the analyses include team-by-year fixed effects to account for these transformations unless otherwise mentioned.

[4] I do not observe who the stringer is for each game, so I cannot employ stringer-FEs.

A.5 Players' Productivity

Table A.20 displays per-game outcomes for players.[5] In the season with robot assistance, as they adjusted their behaviors, the number of walks increased for both pitchers and batters. This is likely due to the pitcher pitching closer to the edge and the batter swinging less (See Table A.7). Batters, following robot exposure, become less offensively productive in the following season. They strike out 0.03 more and generate 0.02 hits less per game. Collectively, these findings suggest the players' productivity is negatively impacted by the robot adoption despite small change in their strategic behaviors. A potential explanation includes an adjustment in the prior season that hindered the growth of players relative to players who didn't have to adjust. Readjusting to inconsistent calls by readjusting strategies might have affected performances.

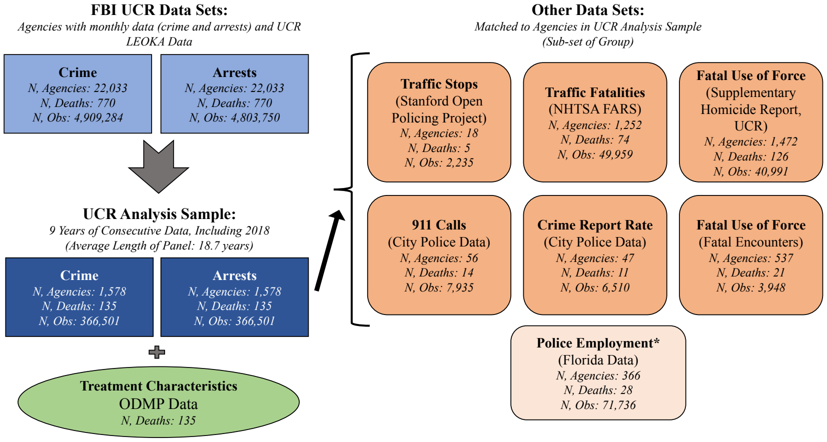
[5] I restrict to pitchers who faced at least 15 batters in an appearance and batters with at least 3 at-bats.

Appendix B

Appendix to “The Impact of Fear on Police Behavior and Public Safety”

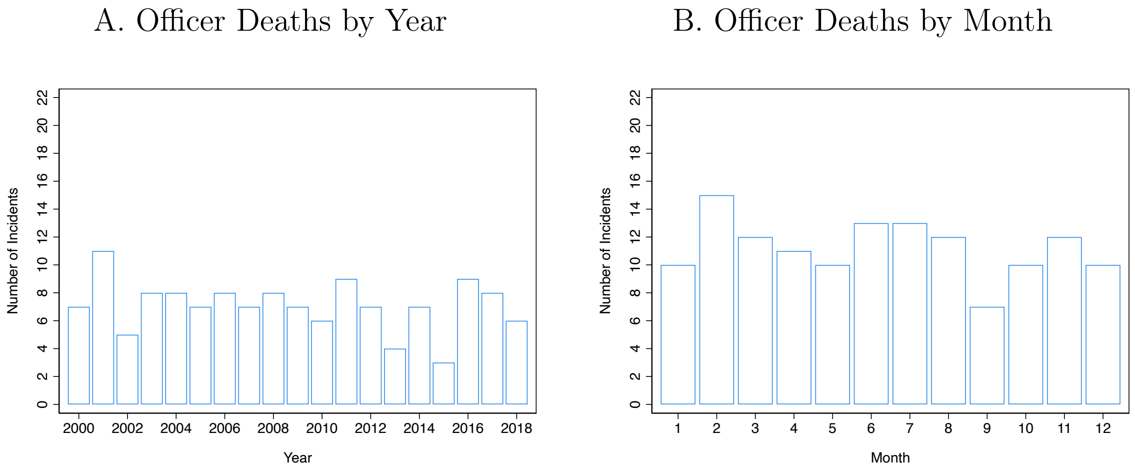
B.1 Figures & Tables

Figure B.1: Sample Construction and Restrictions



Notes: This figure displays the sample construction and restrictions for the various outcomes used in the study. The left columns display how the analysis sample restrictions in the UCR data, namely the restriction for at least 9 years of consecutive monthly data that includes 2018, impacts the set of agencies and treatment events in the data. This base sample is then merged to different sources of outcome data on the right, which creates different sub-sets used for analysis. The Florida employment data uses a slightly different analysis sample in the UCR, which contains Florida agencies that regularly report crimes and arrests to the UCR at the *annual* level, given that no agencies in Florida report monthly crime data to the UCR.

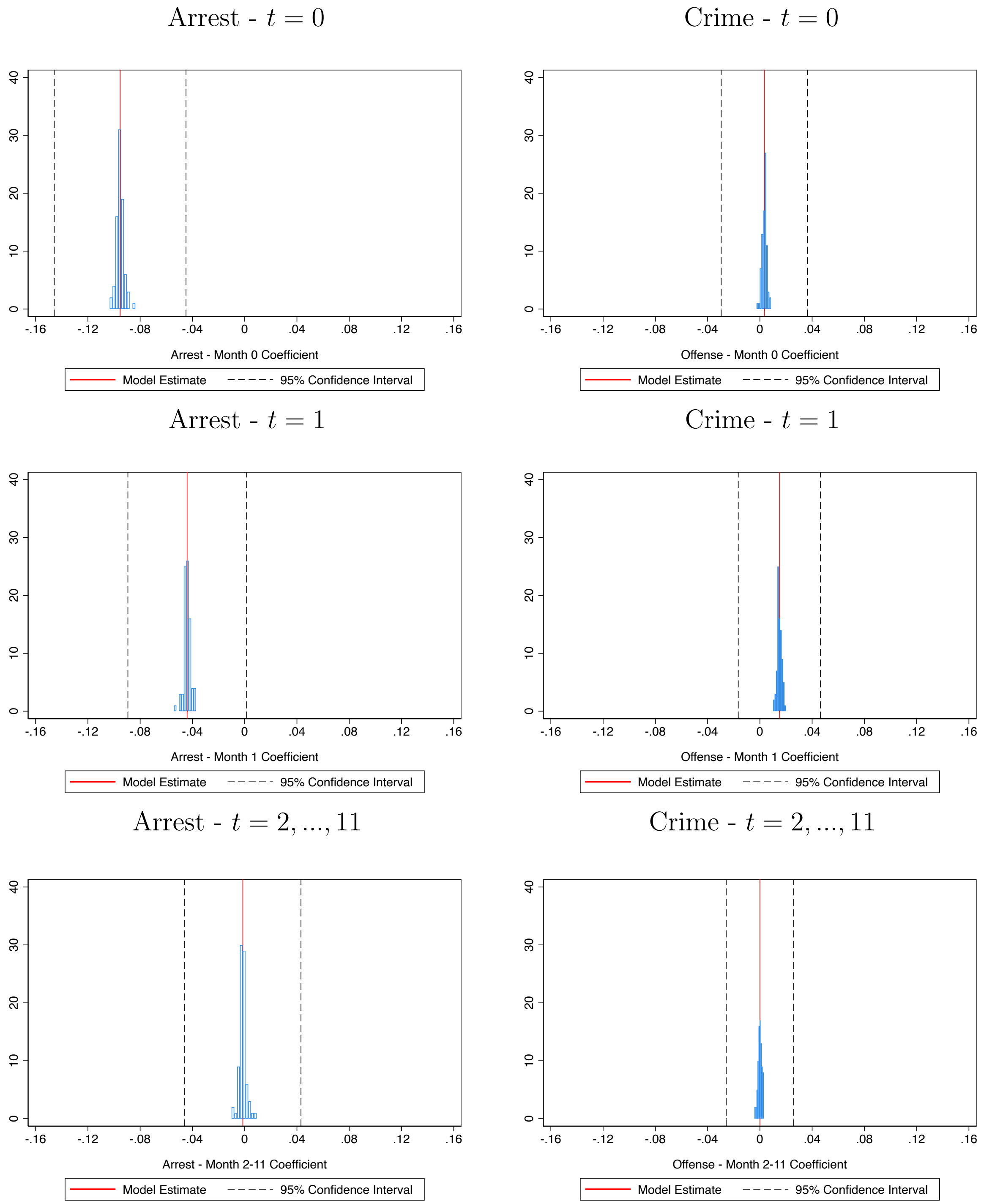
Figure B.2: Variation in Officer Deaths


A. Officer Deaths by Year

B. Officer Deaths by Month

Notes: In 1,578 departments in our sample, there are a total of 135 officer death events in which 151 officers were killed.

Figure B.3: Distribution of Coefficients Dropping Single Treated Agency



Arrest - $t = 0$

Crime - $t = 0$

Arrest - $t = 1$

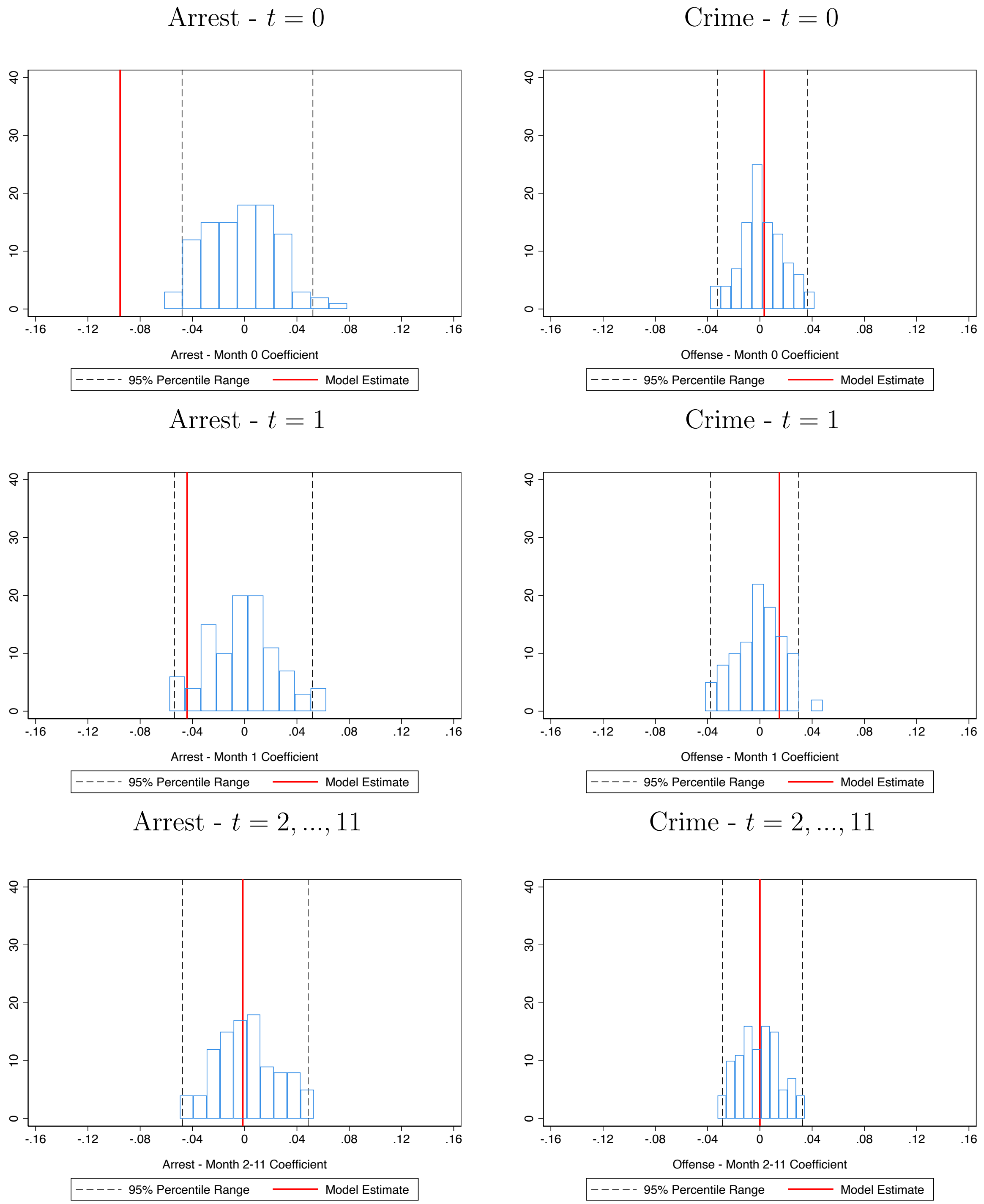
Crime - $t = 1$

Arrest - $t = 2, ..., 11$

Crime - $t = 2, ..., 11$

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Standard errors are clustered at the department level. We re-estimate the model dropping one treatment city at a time. There are 82 treated cities.

Figure B.4: Placebo Treatment Timing


Arrest - $t = 0$

Crime - $t = 0$

Arrest - $t = 1$

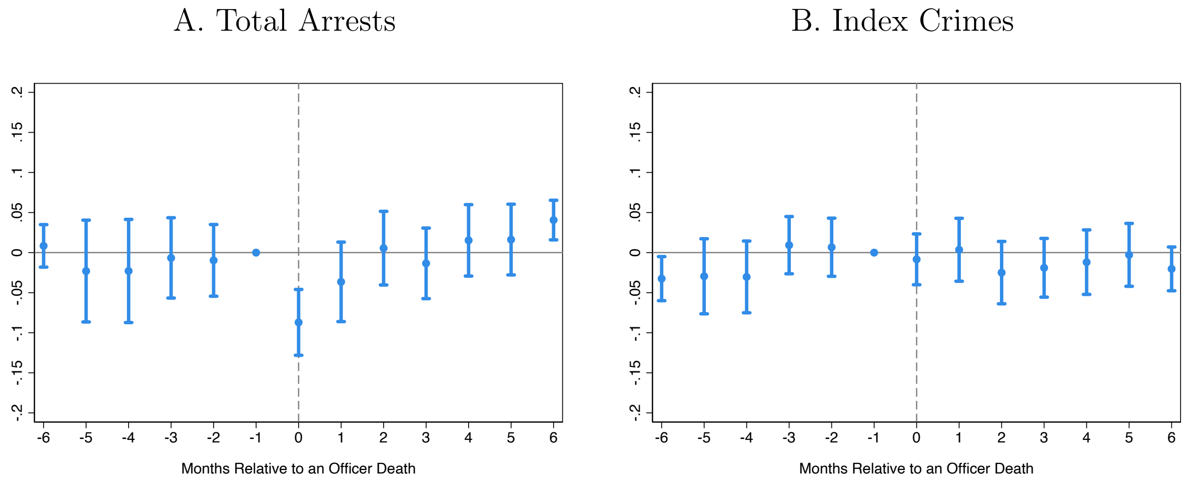
Crime - $t = 1$

Arrest - $t = 2, ..., 11$

Crime - $t = 2, ..., 11$

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Standard errors are clustered at the department level. The timing of officer deaths among treated agencies is randomized holding the number of officer deaths per agency constant. The model is re-estimated 100 times to construct the placebo distribution.

Figure B.5: Event-Study: Sun and Abraham (2021)

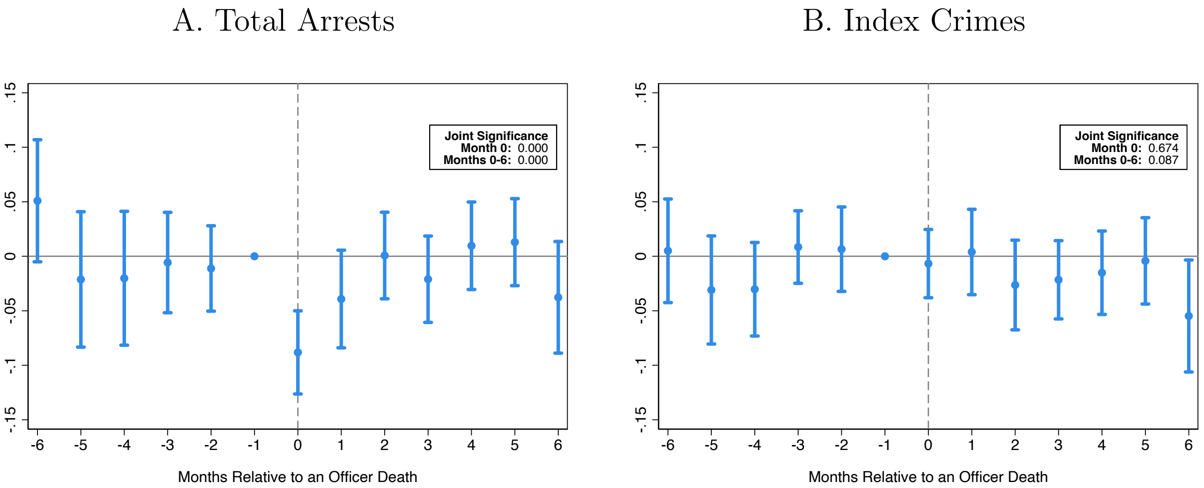


A. Total Arrests

B. Index Crimes

Notes: This figure plots Sun and Abraham (2021)'s proposed "interaction-weighted" coefficient estimator. This estimator combines cohort-specific treatment effects, based on treatment timing, using strictly positive weights. To estimate this model, we include a separate panel for each treatment event, rather than each city. All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Months -6 and 6 include all months before month -6 and all months after month 6, respectively. Standard errors are clustered at the department level.

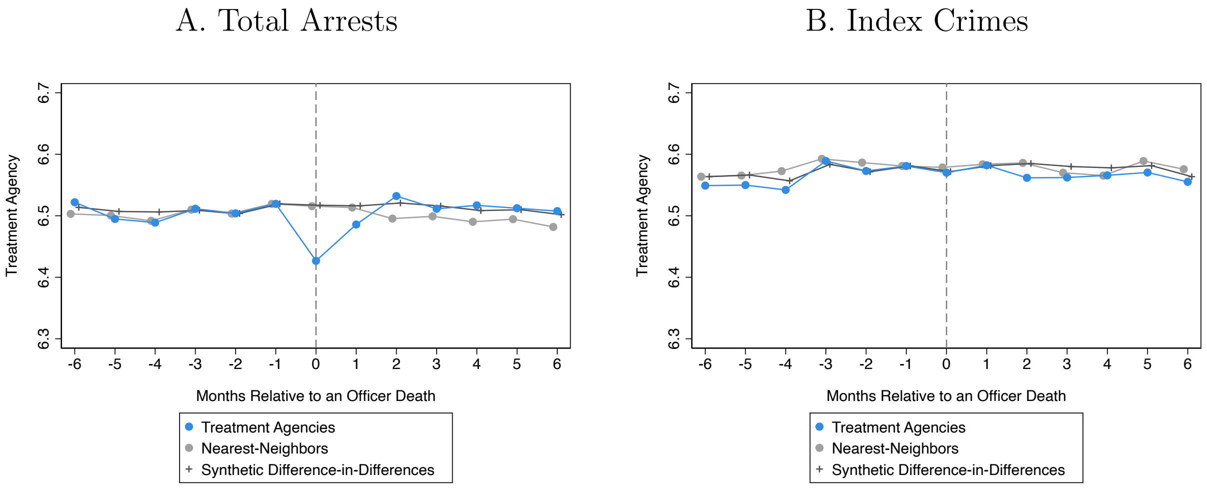
Figure B.6: Event-Study: Omitting Agency-Specific Linear Time Trends


A. Total Arrests

B. Index Crimes

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects. Months -6 and 6 include all months before month -6 and all months after month 6, respectively. Standard errors are clustered at the department level.

Figure B.7: Raw Data: Nearest Neighbor Matching and Synthetic Control

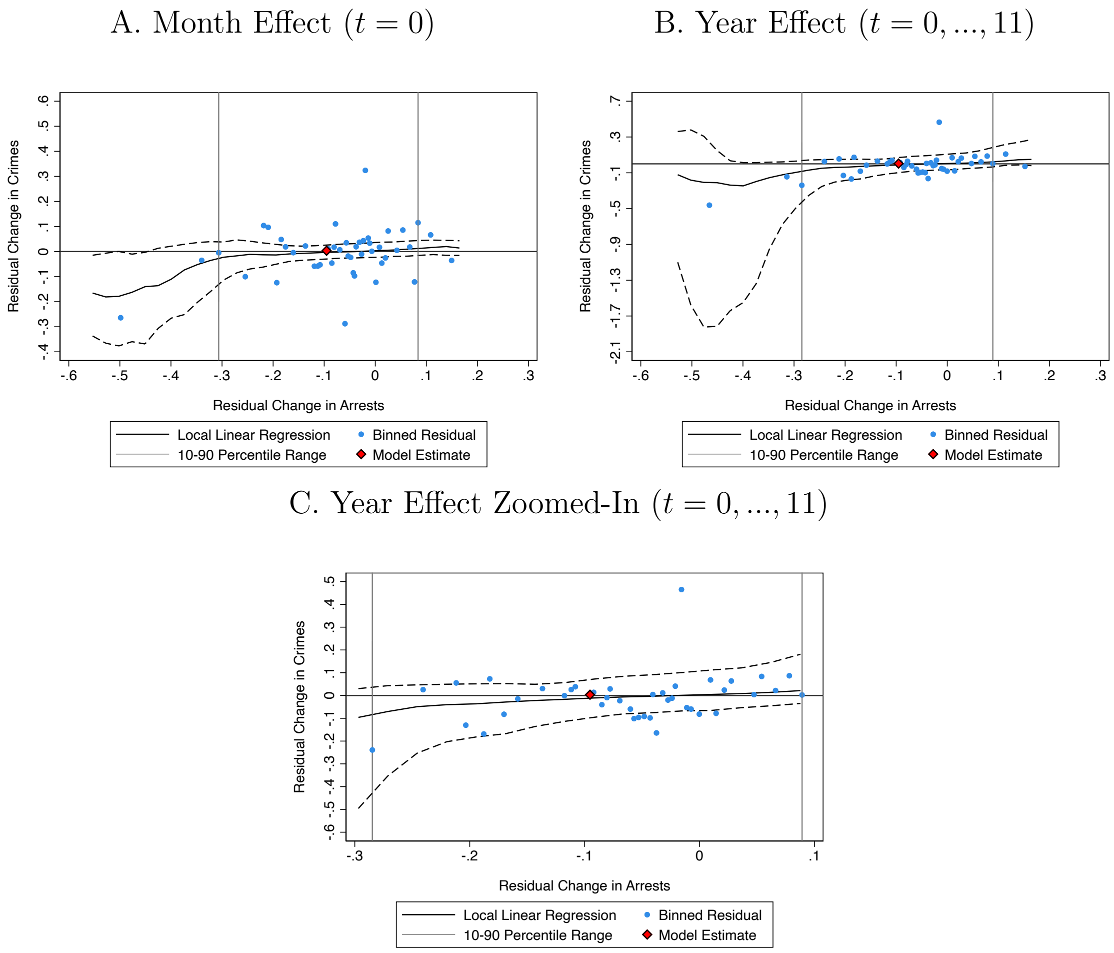


A. Total Arrests

B. Index Crimes

Notes: This figure plots the data around the officer death events. The blue lines plot the raw outcomes of the treated agencies. The light gray lines use the nearest-neighbor matching approach to match treatment event to 5 control agencies using information on demographic characteristics in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment. The darker gray lines use the synthetic difference-in-differences estimation method. A set of 100 nearest-neighbor agencies that do not experience officer death within a year of treatment agency's officer death event is generated by matching on demographic characteristics in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment. Then, from this set, a synthetic control agency is created by matching on demographic characteristics in the treatment year. The synthetic difference-in-differences is estimated and control and treatment series of all periods are obtained. There are 120 matched pairs in both sets. In Panel A, "Nearest-Neighbors" line ranges from 5.25 to 5.29 and "Synthetic Difference-in-Differences" line ranges from 4.57 to 4.59. In Panel B, "Nearest-Neighbors" line ranges from 5.14 to 5.17 and "Synthetic Difference-in-Differences" line ranges from 4.41 to 4.44.

Figure B.8: Arrest to Crime Curve



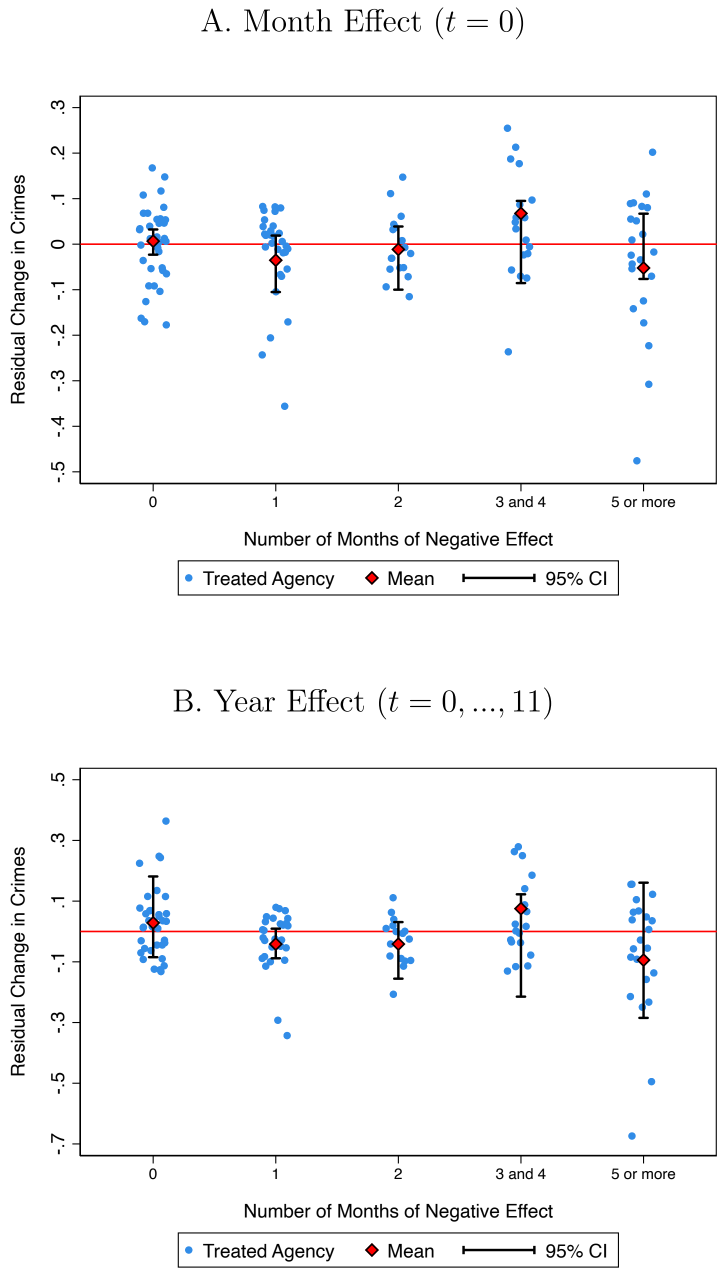
A. Month Effect ($t = 0$)

B. Year Effect ($t = 0, ..., 11$)

C. Year Effect Zoomed-In ($t = 0, ..., 11$)

Notes: The residual changes in arrest and crime are estimated conditional on covariates, a department-specific linear time trend, department-by-calendar month and year-by-month fixed effects and differenced relative to the month prior to a line-of-duty death. The x-axis on all plots shows the residual change in arrests in the month of an officer death. Figure A shows the residual change in crime in the month of an officer death. The Year Effect plots the average monthly residual change in crimes in the year following the officer death event. Each plot has 50 binned values of the residuals. Residuals that are below 5th percentile or above 95th percentile are dropped from the plots. Standard errors (dashed lines) are produced by reproducing the results through block bootstrapping (re-sampling police department panels) 200 times and plotting the 5th and 95th percentile of the local linear regression lines from these iterations. The gray bars represent the 90-10 percentile range.

Figure B.9: Crime Impact by Length of Arrest Decline

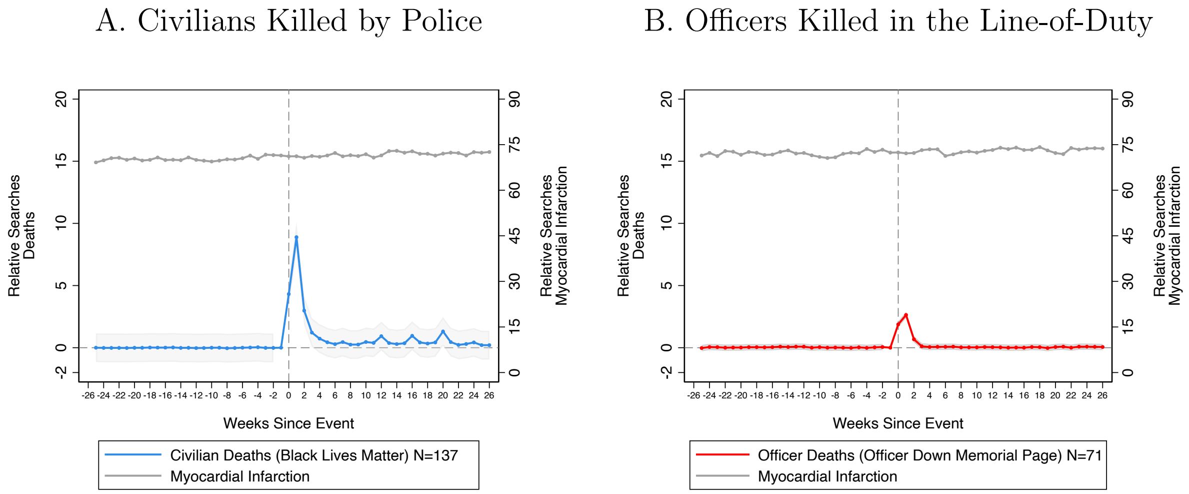


A. Month Effect ($t = 0$)

B. Year Effect ($t = 0, ..., 11$)

Notes: The residual changes in arrest and crime are estimated conditional on covariates, a department-specific linear time trend, department-by-calendar month and year-by-month fixed effects and differenced relative to the month prior to a line-of-duty officer death. The length of arrest effect (x-axis) is determined by the number of consecutive months where the department's estimated arrest residuals are more negative than the residual for the month prior to the line-of-duty officer death. Each plot shows the treated department's values of the residuals, during the month of the officer death, or the average effect for the year following an officer death. The gray bars represent the 95% confidence interval for each duration of arrest decline calculated using a bootstrapping approach with 200 replications. The bootstrap re-samples police departments and recalculates the arrest decline duration as well as the corresponding residual change in crime for each bin in each iteration.

Figure B.10: Google Trends Analysis, Search Volume *Relative* to Benchmarks

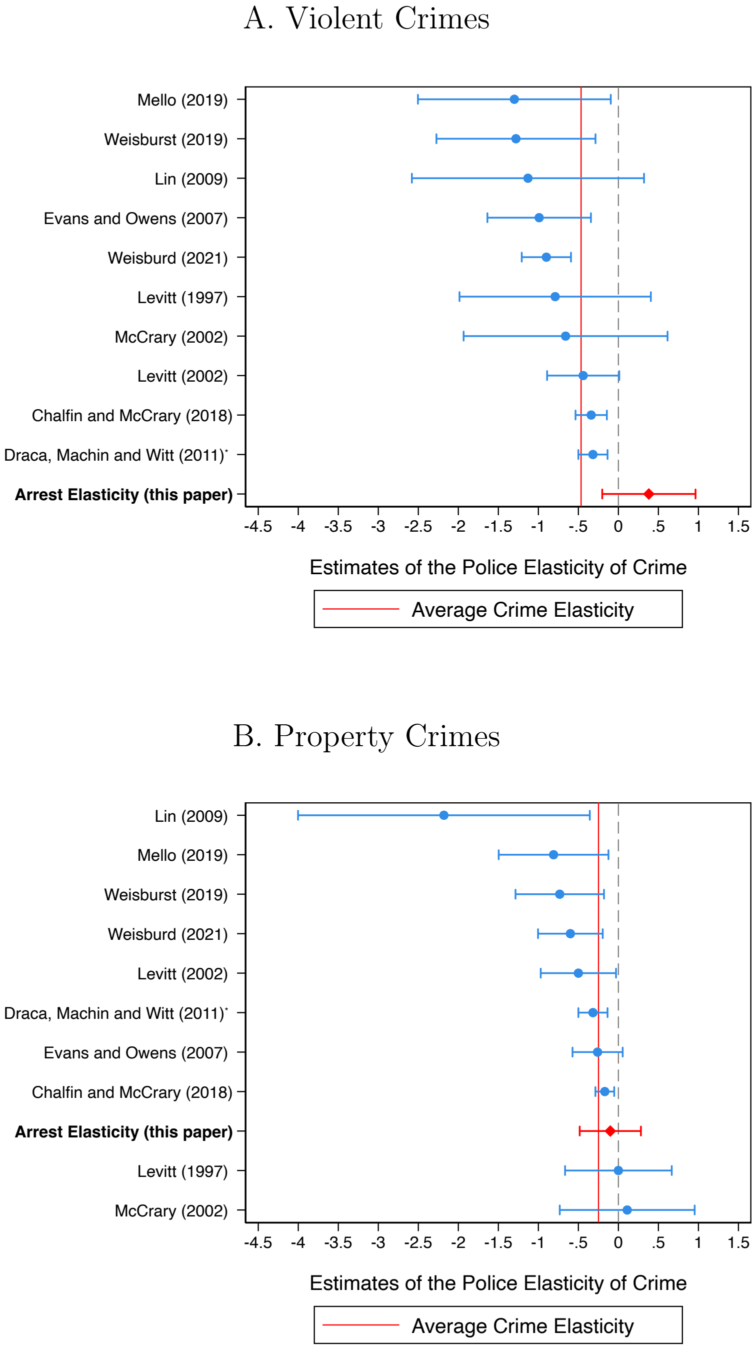


A. Civilians Killed by Police

B. Officers Killed in the Line-of-Duty

Notes: Each search term is an exact first and last name for the individual. We identify high-profile civilian deaths using a list compiled by *Black Lives Matter*, and identify officer deaths by linking the FBI LEOKA data we use in this project to records from the *Officer Down Memorial Page* to obtain officer names. Each search is centered around the time period of -1. Each search is benchmarked by topical searches for the most common cause of death, heart disease, which is relatively stable in popularity across time and locations within the U.S. Google Trends plots relative search intensity with a maximum search popularity in each search of 100. Relative search intensity is calculated in the year around the event in the state of the event. The gray line plots the search popularity for myocardial infraction. The gray shaded area represents the 95% confidence interval from regressing search popularity on weeks with the individual fixed effect.

Figure B.11: Arrest-to-Crime Elasticity (this paper) vs. Police Manpower-to-Crime Elasticities



A. Violent Crimes

B. Property Crimes

Notes: The estimates of the police elasticities of violent and property crimes are from recent articles. Draca et al. (2011) estimates an elasticity of total crime with respect to police employment. For the Levitt (1997) estimates, we take the elasticity estimates from McCrary (2002) correcting for a coding error in the original paper. The estimates from this paper use the crime elasticity with respect to changes in total arrest enforcement. The red bars represent the average elasticities of all articles excluding our estimates, weighted by the inverse of their variance.

Table B.1: Summary Demographic Characteristics

		Full Sample			Treated Agencies	
	Mean	S.D.	N	Mean	S.D	N
Characteristics of Cities						
Number of Police Officers	75.2	(349.7)	29564	582.8	(1397.1)	1544
Number of Officers Killed by Felony	0.005	(0.085)	29564	0.096	(0.332)	1544
Number of Officers Assaulted	10.8	(48.1)	29564	74.9	(176.6)	1544
% Black	7.7	(12.0)	29564	15.0	(17.8)	1544
% Hispanic	16.8	(20.8)	29564	22.6	(21.2)	1544
% White	68.0	(24.7)	29564	54.2	(24.6)	1544
% Male	48.8	(3.4)	29564	48.9	(1.8)	1544
% Female-Headed Household	31.3	(8.2)	29564	33.8	(7.1)	1544
% Age <14	20.2	(4.7)	29564	20.8	(4.4)	1544
% Age 15-24	14.3	(6.8)	29564	16.6	(6.9)	1544
% Age 25-44	27.2	(5.2)	29564	28.4	(3.9)	1544
% Age >45	38.3	(8.6)	29564	34.2	(7.8)	1544
% < High School	15.9	(11.0)	29564	17.7	(9.4)	1544
% High School Graduate	28.3	(9.5)	29564	25.7	(7.1)	1544
% Some College	28.3	(7.3)	29564	29.4	(5.7)	1544
% College Graduate or More	27.6	(16.1)	29564	27.2	(13.3)	1544
Unemployment Rate	4.8	(3.1)	29564	5.6	(2.3)	1544
Poverty Rate	12.7	(8.7)	29564	15.7	(7.5)	1544
Median Household Income	45658.5	(20918.3)	29564	40249.9	(15112.0)	1544
Population	41205.4	(133018.3)	29564	243160.3	(504777.6)	1544
Number of Agencies	1578					
Number of Treated Agencies	82					

Notes: The characteristics information are from the data with crime activity outcomes. Officer related information are from the FBI's Law Enforcement Officer Killed or Assaulted (LEOKA) that covers the period 2000-2018. Demographics data come from the 2000 U.S. Census and the American Community Survey 5-year estimates from 2010 to 2018. For years 2001 to 2009, the demographics information are linearly interpolated.

Table B.2: Robustness Specifications

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
(1) Baseline Specification									
Murder Offenses	0.391***	(0.058)	0.033	(0.039)	0.015	(0.013)	0.22	2.35	354504
adj. for Officer Death	0.052	(0.047)	0.031	(0.039)	0.015	(0.012)	0.22	2.34	354495
Arrests	-0.095***	(0.026)	-0.044*	(0.023)	-0.001	(0.023)	151.9	964.5	354507
Violent Crimes	-0.036	(0.027)	0.039	(0.029)	-0.034*	(0.018)	18.3	165.8	354507
Property Crimes	0.010	(0.018)	0.012	(0.016)	0.002	(0.014)	121.6	857.7	354507
(2) Restrict to Treated Cities									
Murder Offenses	0.393***	(0.058)	0.031	(0.039)	0.013	(0.013)	2.35	2.35	18510
Arrests	-0.097***	(0.026)	-0.044**	(0.022)	-0.005	(0.021)	964.5	964.5	18510
Violent Crimes	-0.037	(0.028)	0.035	(0.030)	-0.036*	(0.018)	165.8	165.8	18510
Property Crimes	0.010	(0.020)	0.013	(0.016)	0.005	(0.014)	857.7	857.7	18510
(3) Separate Panel for Each Event									
Murder Offenses	0.379***	(0.057)	0.034	(0.038)	0.014	(0.011)	0.64	6.51	366498
Arrests	-0.100***	(0.024)	-0.050**	(0.020)	-0.008	(0.018)	255.4	1888.9	366501
Violent Crimes	-0.024	(0.025)	0.048*	(0.028)	-0.022	(0.016)	43.9	415.4	366501
Property Crimes	0.012	(0.016)	0.015	(0.013)	0.005	(0.010)	235.5	1935.9	366501
(4) Counting Multiple Officer Deaths Additively									
Murder Offenses	0.359***	(0.056)	0.035	(0.032)	0.019*	(0.011)	0.22	2.35	354504
Arrests	-0.085***	(0.023)	-0.043**	(0.021)	-0.004	(0.021)	151.9	964.5	354507
Violent Crimes	-0.025	(0.022)	0.038	(0.025)	-0.026	(0.016)	18.3	165.8	354507
Property Crimes	0.009	(0.017)	0.011	(0.014)	0.001	(0.012)	121.6	857.7	354507

Table B.2: Robustness Specifications (Continued)

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
(5) Drop Agency × Month									
Murder Offenses	0.393***	(0.058)	0.033	(0.037)	0.016	(0.013)	0.22	2.35	354504
Arrests	-0.092***	(0.026)	-0.040*	(0.024)	-0.002	(0.023)	151.9	964.5	354507
Violent Crimes	-0.036	(0.025)	0.037	(0.028)	-0.033*	(0.018)	18.3	165.8	354507
Property Crimes	0.011	(0.019)	0.013	(0.018)	0.002	(0.014)	121.6	857.7	354507
(6) Add State-by-Year FE									
Murder Offenses	0.389***	(0.058)	0.032	(0.039)	0.013	(0.013)	0.22	2.35	354504
Arrests	-0.102***	(0.026)	-0.049**	(0.023)	-0.005	(0.022)	151.9	964.5	354507
Violent Crimes	-0.036	(0.027)	0.039	(0.030)	-0.028	(0.018)	18.3	165.8	354507
Property Crimes	0.004	(0.018)	0.007	(0.015)	-0.003	(0.013)	121.6	857.7	354507
(7) Remove DUI Arrests									
Murder Offenses	0.391***	(0.058)	0.033	(0.039)	0.015	(0.013)	0.22	2.35	354504
Arrests	-0.090***	(0.026)	-0.037	(0.024)	0.002	(0.023)	139.2	895.4	354507
Violent Crimes	-0.036	(0.027)	0.039	(0.029)	-0.034*	(0.018)	18.3	165.8	354507
Property Crimes	0.010	(0.018)	0.012	(0.016)	0.002	(0.014)	121.6	857.7	354507
(8) Levels Model									
Murder Offenses	1.337***	(0.502)	0.053	(0.271)	-0.153	(0.130)	0.22	2.35	354504
Arrests	-69.192*	(36.695)	-21.615	(51.944)	-3.457	(47.503)	151.9	964.5	354507
Violent Crimes	-4.655	(8.450)	2.090	(9.000)	-5.475	(9.548)	18.3	165.8	354507
Property Crimes	-8.650	(21.749)	12.234	(20.065)	-24.597	(26.627)	121.6	857.7	354507

Table B.2: Robustness Specifications (Continued)

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
(9) Per Capita Model (Per 100K Residents)									
Murder Offenses	1.944***	(0.407)	0.133	(0.113)	0.013	(0.042)	0.29	0.65	354504
Arrests	-41.918***	(10.609)	-22.632**	(9.960)	-6.843	(9.320)	456.1	457.1	354507
Violent Crimes	-1.752	(1.446)	0.863	(1.484)	-1.676	(1.090)	32.2	51.9	354507
Property Crimes	-1.383	(6.385)	3.669	(5.623)	-0.121	(5.216)	293.2	344.9	354507
(10) Inverse Hyperbolic Sine Model									
Murder Offenses	0.498***	(0.074)	0.039	(0.049)	0.020	(0.016)	0.11	0.72	354504
Arrests	-0.097***	(0.026)	-0.045*	(0.024)	-0.002	(0.023)	4.8	6.4	354507
Violent Crimes	-0.042	(0.031)	0.045	(0.033)	-0.041**	(0.019)	2.0	4.1	354507
Property Crimes	0.010	(0.019)	0.011	(0.017)	0.002	(0.014)	4.4	6.2	354507
(11) Sun & Abraham (2020) IW Estimator									
Murder Offenses	0.380***	(0.044)	0.032	(0.034)	0.011	(0.007)	0.64	6.51	366498
Arrests	-0.090***	(0.024)	-0.040*	(0.021)	0.003	(0.009)	255.4	1888.9	366501
Violent Crimes	-0.029	(0.024)	0.043	(0.027)	-0.028***	(0.007)	43.9	415.4	366501
Property Crimes	0.012	(0.017)	0.014	(0.015)	0.005	(0.006)	235.5	1935.9	366501
(12) Drop Time Trend									
Murder Offenses	0.376***	(0.059)	0.017	(0.039)	-0.002	(0.011)	0.22	2.35	354504
Arrests	-0.138***	(0.028)	-0.089***	(0.024)	-0.049**	(0.023)	151.9	964.5	354507
Violent Crimes	-0.044	(0.029)	0.030	(0.031)	-0.041**	(0.020)	18.3	165.8	354507
Property Crimes	-0.007	(0.022)	-0.007	(0.019)	-0.017	(0.017)	121.6	857.7	354507
(13) Nearest Neighbor Matching									
Murder Offenses	0.380***	(0.058)	0.015	(0.041)	0.002	(0.017)	1.7	6.5	31097
Arrests	-0.126***	(0.026)	-0.064***	(0.022)	-0.025	(0.018)	699.7	1987.7	31098
Violent Crimes	-0.046	(0.033)	0.052	(0.034)	-0.039**	(0.018)	122.8	423.4	31098
Property Crimes	-0.014	(0.020)	-0.010	(0.018)	-0.023	(0.014)	645.4	1947.9	31098

Notes: The baseline specification is a replicate of output in Table 2.2 and each subsequent model is a variant of this baseline. Model (2) restricts the sample to treated cities. Model (3) uses a separate panel for each officer death treatment rather than each department. Model (4) counts multiple death events additively rather than as a single event. Model (5) drops the agency-by-month fixed effect. Model (6) adds state by year fixed effects. Model (7) removes the DUI arrests counts from the total arrests. Models (8), (9) and (10) consider alternate functional forms, using a levels, a per capita and an inverse hyperbolic sine, respectively. Model (11) uses Sun and Abraham (2021)'s proposed estimator. Model (12) drops the department-specific linear time trends and Model (13) uses a nearest neighbor matching approach. Standard errors are clustered at the department level. * p<0.1,** p<0.05, *** p<0.01.

Table B.3: Predicting Treatment-Specific Synthetic Control Arrest Effects

Predicting Agency-Level Arrest Effect	Agency	S.E.	Incident	S.E.	All	S.E.
Log Population	0.024**	(0.012)			0.020	(0.012)
% White Population	0.002	(0.001)			0.002	(0.001)
% Less than HS	0.006**	(0.003)			0.005*	(0.003)
% Poverty	-0.003	(0.003)			-0.000	(0.003)
Crime Rate	0.006	(0.008)			0.006	(0.008)
Officer Non-White			-0.022	(0.038)	-0.039	(0.038)
Officer Female			-0.079	(0.076)	-0.062	(0.076)
During Traffic Stop			-0.009	(0.049)	0.004	(0.048)
Not Cleared within 48 hrs			-0.099**	(0.043)	-0.084*	(0.045)
Weighted Mean	-0.081		-0.081		-0.081	
Variance	0.029		0.029		0.029	
F-Statistic	2.098		1.793		1.777	
p-value	0.071		0.135		0.080	
R-squared	0.084		0.059		0.127	
Observations	120		120		120	

Notes: A set of 100 nearest-neighbor agencies that do not experience officer death within a year of treatment agency's officer death event is generated by matching on demographic characteristics in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment. Then, from this set, a synthetic control agency is created by matching on demographic characteristics in the treatment year. There are 120 matched pairs. The synthetic difference-in-differences is estimated and post-period treatment effects are obtained. The table shows the results of regressing agency-level treatment effect for each respective post-period on covariates. The covariates are the first reported measure for each department. "Weighted Mean" shows the average treatment effect weighted by inverse of the standard error squared and "Variance" is the variance of the treatment effects. * p<0.1,** p<0.05, *** p<0.01.

Table B.4: Agency-Level Characteristics by Predicted Arrest Effect Size

Agency-Level Characteristics by Predicted Arrest Effect Size (Months 0 & 1)	Full Sample	Top Quartile $E(\tau \mid X) <$ -0.084	IQR $E(\tau \mid X) \in$ $(-0.084, -0.024)$	Bottom Quartile $E(\tau \mid X) >$ -0.024
Log Population	12.041	10.932	12.215	12.803
% White Population	51.423	60.546	48.826	47.494
% Less than HS	22.613	14.976	22.655	30.167
% Poverty	14.596	11.402	15.198	16.587
Crime Rate	4.753	3.509	5.042	5.419
Officer Non-White	0.217	0.133	0.233	0.267
Officer Female	0.050	0.133	0.033	0.000
During Traffic Stop	0.133	0.200	0.133	0.067
Not Cleared within 48 hrs	0.158	0.367	0.133	0.000

Notes: A set of 100 nearest-neighbor agencies that do not experience officer death within a year of treatment agency's officer death event is generated by matching on demographic characteristics in the treatment year and lagged monthly crime and arrest levels in the year prior to treatment. Then, from this set, a synthetic control agency is created by matching on demographic characteristics in the treatment year. There are 120 matched pairs. The synthetic difference-in-differences is estimated and post-period treatment effects are obtained. This table shows the department characteristics splitting the sample by the predicted treatment effect size from Table B.3. * p<0.1,** p<0.05, *** p<0.01.

Table B.5: Additional Outcomes

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
A. 911 Call Outcomes									
Officer-Initiated Interactions	-0.047*	(0.027)	-0.020	(0.022)	0.019	(0.029)	4996.5	8292.8	5873
Officer Presence	0.002	(0.002)	0.003	(0.003)	0.006	(0.004)	0.2	0.3	5813
B. Employment Outcomes, Florida									
Full-Time Equivalent Officers	0.005	(0.014)	0.003	(0.014)	0.004	(0.014)	108.0	512.7	71736
Number of Hired Officers	0.229	(0.204)	-0.231	(0.157)	-0.080	(0.065)	0.8	2.9	71736
Number of Fired Officers	0.022	(0.075)	0.109*	(0.058)	0.003	(0.022)	0.1	0.4	71736
Number of Officer Deaths	0.630***	(0.049)	0.024	(0.031)	0.002	(0.006)	0.0	0.1	71736
Number of Officer Quits	-0.047	(0.062)	0.039	(0.072)	-0.042**	(0.021)	0.6	2.4	71736
C. Traffic Accidents									
Fatal Traffic Accidents	-0.023	(0.045)	-0.016	(0.031)	-0.025*	(0.013)	0.26	1.60	283906
Accidents involving Alcohol	0.012	(0.043)	-0.004	(0.032)	-0.018	(0.022)	0.09	0.57	256978
D. Fatal Use-of-Force									
Supplementary Homicide Report	0.024	(0.025)	-0.024	(0.018)	0.003	(0.006)	0.02	0.16	359733
Fatal Encounters	0.044	(0.037)	-0.025	(0.039)	0.030**	(0.014)	0.03	0.26	172760
E. Accidental Officer Death									
Murder Offenses	0.006	(0.040)	0.061	(0.044)	0.005	(0.015)	0.23	2.45	329669
Arrests	-0.019	(0.026)	0.008	(0.031)	0.011	(0.030)	155.2	967.9	329672
Violent Crimes	0.031	(0.045)	0.004	(0.044)	0.019	(0.023)	19.0	183.3	329672
Property Crimes	0.009	(0.027)	-0.049	(0.037)	-0.005	(0.024)	125.1	986.9	329672

Notes: All regressions include department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions in Panels A, C, D and E additionally include a vector of covariates at the department-by-year level. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$ and outcome means are given in levels. Standard errors are clustered at the department level. "Officer-Initiated Interactions" refers to the number of officer-initiated incidents (e.g. traffic stops, on-sight investigations). "Officer Presence" represents the proportion of Census block groups visited by an officer in response to a 911 call. "Employment Outcomes, Florida" panel uses the officer-level data from the Florida Department of Law Enforcement and covers all law enforcement agency officer employment spells from 2000 to 2016. Reasons for termination include violations of policies or standards, failure to qualify, misconduct, etc. Officer quits include all voluntary separations. "Accidents involving alcohol" is the number of fatal traffic accidents with at least one driver with the blood alcohol concentration 0.01 g/dL or higher involved in a crash. Fatal Use-of-Force panel includes two measures of civilians killed by police. First measure is a count of deaths at the hands of officers from the Supplementary Homicide Report of the FBI UCR series. Second, *Fatal Encounters* is a count of civilians killed by police from a crowd-sourced data series, which we restrict to the sample period of 2010-2018 for data quality reasons. Both measures exclude records of deaths of suspects involved in the line-of-duty officer death event during month 0, as well as records of civilian deaths that occur before the officer death in month 0. "Accidental Officer Death" panel shows the four main outcomes using the accidental officer death as a treatment instead of felonious death. There are 73 officer accidental death events. * p<0.1,** p<0.05, *** p<0.01.

B.2 Heterogeneity by Offense Type and Arrestee Demographics

In this appendix, we include additional heterogeneity analysis by both crime type and demographics of arrestees.

B.2.1 Crime and Arrest Sub-Types

Next, we estimate the baseline model separately for each crime and arrest sub-type in the analysis to explore which categories are driving changes in the aggregate outcome sums. Table B.6 displays the sub-type results for index crime arrests and index crimes. For index crime arrests, we find significant decreases in robbery, aggravated assault, and motor vehicle theft arrests. There is a long-term decline in aggravated assault arrests; here, we are cautious to interpret this as a treatment effect given the lack of long-term effects for any other subcategory of serious arrests. For index crime, we observe no significant changes in any category in the first month of treatment or the month after.

The results for "quality of life" arrests and "non-index" arrests provide a more detailed picture of what types of arrests are reduced as a result of treatment. Table B.7 shows that there are large and significant declines in arrests for weapons offenses, prostitution, driving under the influence of alcohol (DUI) (which is classified as a mid-level "non-index" offense), drug sale, drug possession, and arrests that are uncategorized in the UCR.[1] Several of these declines correspond to reductions that are greater than 10%. The results imply that over the two month period following an officer death, officers make 1.5 fewer arrests for weapons offenses, 3 fewer arrests for prostitution, 19 fewer DUI arrests, 9 fewer arrests for drug sales, 22 fewer arrests for drug possession, and 27 fewer uncategorized arrests in each treated city.[2]

[1] The results also show marginally significant second month effects for other assault and vandalism.

[2] We assume that uncategorized arrests are likely to be for offenses that are not listed as options for reporting in UCR. Given the broad number of offense categories available for reporting in UCR, we argue that these arrests are for other low-level offenses.

Given that we observe a large reduction in DUI arrests, we explicitly measure the subset of fatal traffic accidents that involve a drunk driver (Table B.5). These alcohol-related accidents do not respond to the reduction in DUI arrests associated with an officer death. Likewise, as discussed above, the decline in total arrests persists after excluding DUI arrests (see Table B.2, specification (7)).

B.2.2 Demographics of Arrestees

Another treatment dimension of interest is who is affected by the reduction in arrests that we observe. We investigate whether the declines are concentrated among particular demographic groups by regressing demographic-specific measures of log arrests on our treatment, using our preferred specification. Table B.8 shows that we observe arrest declines across all race, gender, and age groups following an officer death in the line-of-duty. While the point estimates vary somewhat across groups, we cannot reject that any of the demographic sub-group declines differ in magnitude from the total arrest effect of a 9.5% decline. The share of Black arrestees, 36%, and male arrestees, 76%, exceeds their respective population shares of 15% and 49% in the treatment sample. As a result, the equivalent percent declines across groups leads to a reduction in the disparity, in levels, of arrests across races and genders.

Table B.6: Index Crimes and Arrests by Type

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
A. Murder Outcomes									
Murder Offenses	0.391***	(0.058)	0.033	(0.039)	0.015	(0.013)	0.22	2.35	354504
Murder Arrests	0.111**	(0.044)	0.071	(0.043)	-0.000	(0.023)	0.17	1.57	354507
B. Index Arrests									
Rape	-0.014	(0.029)	-0.042	(0.033)	-0.001	(0.018)	0.28	2.08	354507
Robbery	-0.094***	(0.035)	-0.059	(0.047)	0.003	(0.023)	1.7	15.6	354507
Aggravated Assault	-0.088**	(0.035)	-0.036	(0.028)	-0.056**	(0.025)	6.4	44.3	354506
Burglary	0.004	(0.040)	0.022	(0.045)	0.014	(0.028)	3.7	20.7	354507
Theft	-0.072*	(0.042)	-0.034	(0.042)	-0.022	(0.034)	14.9	82.6	354507
Motor Vehicle Theft	-0.098*	(0.055)	-0.118*	(0.062)	-0.044	(0.062)	1.4	11.8	354507
C. Index Crime									
Rape	-0.040	(0.035)	0.042	(0.038)	-0.006	(0.021)	1.3	10.1	353656
Robbery	-0.004	(0.030)	0.009	(0.032)	-0.017	(0.017)	5.9	61.0	354382
Aggravated Assault	-0.044	(0.034)	0.036	(0.030)	-0.034	(0.021)	11.1	94.8	354355
Burglary	0.041	(0.029)	0.023	(0.031)	0.010	(0.020)	24.0	175.3	354478
Theft	-0.026	(0.029)	-0.013	(0.026)	-0.022	(0.022)	81.9	541.9	354506
Motor Vehicle Theft	0.026	(0.033)	-0.009	(0.031)	0.011	(0.023)	15.7	140.5	354389

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$ and outcome means are given in levels. Standard errors are clustered at the department level. * p<0.1,** p<0.05, *** p<0.01.

Table B.7: Non-Index Arrest Outcomes by Type

	1st Month		2nd Month		Long-Term		Outcome Mean		
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N
A. Non-Index Arrests									
Manslaughter	0.013	(0.024)	0.014	(0.024)	-0.005	(0.010)	0.01	0.10	354507
Arson	0.023	(0.041)	-0.058	(0.041)	-0.012	(0.022)	0.15	0.85	354507
Other Assault	-0.028	(0.034)	-0.058*	(0.035)	-0.002	(0.030)	13.6	89.2	354507
Weapons	-0.083**	(0.042)	-0.007	(0.038)	-0.018	(0.023)	2.3	17.1	354507
Prostitution	-0.079*	(0.042)	-0.104*	(0.057)	-0.038	(0.041)	1.2	15.5	354507
Other Sex Offense	-0.052	(0.034)	-0.042	(0.040)	-0.010	(0.028)	0.92	6.68	354507
Family Offense	-0.022	(0.050)	0.057	(0.043)	0.032	(0.040)	0.58	4.14	354506
DUI	-0.164***	(0.048)	-0.108***	(0.042)	-0.031	(0.034)	12.7	69.1	354507
Drug Sale	-0.154*	(0.088)	-0.101	(0.091)	-0.108	(0.110)	3.8	35.4	354506
Forgery	-0.006	(0.039)	-0.037	(0.043)	-0.002	(0.028)	1.04	5.38	354507
Fraud	-0.011	(0.046)	-0.007	(0.046)	0.053	(0.033)	1.71	8.29	354507
Embezzlement	-0.028	(0.046)	-0.017	(0.033)	0.019	(0.025)	0.23	1.07	354507
Stolen Property	0.008	(0.048)	0.056	(0.047)	0.056	(0.042)	1.49	7.49	354505
Runaway	0.034	(0.041)	0.015	(0.043)	0.011	(0.045)	1.16	7.87	354507
B. Quality of Life Arrests									
Disorderly Conduct	-0.013	(0.049)	-0.023	(0.050)	0.011	(0.043)	5.3	29.4	354506
Curfew/Loitering	-0.069	(0.067)	0.018	(0.059)	-0.019	(0.065)	2.3	30.7	354507
Vandalism	-0.069	(0.042)	-0.073*	(0.043)	-0.040	(0.035)	2.9	17.1	354507
Gambling	-0.049	(0.031)	-0.004	(0.032)	-0.016	(0.021)	0.06	0.65	354506
Vagrancy	0.007	(0.077)	-0.006	(0.075)	0.042	(0.075)	0.55	6.02	354507
Drunkenness	-0.056	(0.068)	0.015	(0.064)	-0.010	(0.060)	8.9	44.3	354507
Liquor	-0.058	(0.071)	-0.053	(0.068)	-0.001	(0.059)	5.0	27.8	354507
Drug Possession	-0.107**	(0.054)	-0.109*	(0.060)	-0.044	(0.063)	17.5	102.8	354507
Uncategorized Arrests	-0.100*	(0.059)	-0.003	(0.043)	0.056	(0.044)	40.1	260.5	354507

Notes: All regressions include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$ and outcome means are given in levels. Standard errors are clustered at the department level. * p<0.1,** p<0.05, *** p<0.01.

Table B.8: Heterogeneity, Arrestee Demographics

	1st Month		2nd Month		Long-Term		Outcome Mean			p-value
	(t=0)	S.E.	(t=1)	S.E.	(t=2,...,11)	S.E.	Full	Treated	N	Diff. total
Policing Activity										
Total Arrests	-0.095***	(0.026)	-0.044*	(0.023)	-0.001	(0.023)	151.9	964.5	354507	
Black	-0.069**	(0.029)	-0.006	(0.030)	0.015	(0.022)	40.0	353.1	354507	0.499
White	-0.107***	(0.029)	-0.062**	(0.025)	-0.005	(0.024)	108.2	590.7	354507	0.760
Male	-0.093***	(0.026)	-0.042*	(0.023)	-0.003	(0.022)	114.1	736.6	354507	0.951
Female	-0.097***	(0.029)	-0.049*	(0.028)	0.004	(0.025)	37.8	227.9	354507	0.959
Adult	-0.096***	(0.028)	-0.043*	(0.025)	0.000	(0.024)	130.5	832.7	354507	0.980
Juvenile	-0.097**	(0.042)	-0.077*	(0.045)	-0.019	(0.036)	21.3	131.9	354507	0.980

Notes: Regressions in include a vector of covariates at the department-by-year level, department-by-calendar month and year-by-month fixed effects and department-specific linear time trends. Regressions also include a dummy variable for 12 or more months after the occurrence of an officer death. Outcomes are defined as $Y_{it} = log(y_{it} + 1)$ and outcome means are given in levels. Standard errors are clustered at the department level. The last column reports the p-value from testing whether the first month effects of the sub-group are equal to the total arrests effect. Juvenile is defined to be people arrested under 18 years of age. * p<0.1,** p<0.05, *** p<0.01.

B.3 Nearest Neighbor and Synthetic Control Methods

This appendix details the estimation methods for the Nearest Neighbor and Synthetic Control estimates used in the paper. These estimates are used in Table B.3, Figure 2.5, specification (13) of Table B.2, Figure B.7, and Table B.4.

The purpose of these exercises is twofold. First, we aim to estimate effects where treatment events are matched to highly similar control units based on pre-treatment characteristics. This matching to control units allows us to compare post-treatment outcomes across treatment and control units in a parsimonious way, omitting x-variables and time trends in these matched comparisons. Second, the synthetic control methods allow us to recover an individual treatment effect for each event in our data, which we use to examine heterogeneity of treatment effects.

Below are the steps used in these analyses:

1. Apply the Nearest Neighbor matching algorithm to treatment events with at least one year of pre-treatment data. For each treatment, we restrict the pool of possible controls to agencies that do not have an officer death event in the year prior or after the treatment death event. We use a matching algorithm that minimizes the distance between matched covariates. The covariates we use are counts of violent and property crimes and arrests for periods -1, -2, and -3, and the slope of these outcomes between periods -3 to -12, as well as the treatment year city-level poverty rate, share white, share with a high school degree or less education, and log population.

2. Obtain the set of 100 nearest neighbors from step (1). This is the donor pool for the synthetic control analysis.

3. Using the 10 closest nearest neighbors for each treatment from step (1), estimate a parsimonious difference-in-difference model in this sample that contains no x-variables or unit-specific time trends. This estimate is the "Nearest Neighbor" specification and

is reported in specification (13) of Table B.2. The average outcomes of treatment and control using this method are also presented in the top panel of Figure B.7.

4. Conduct a synthetic control match using the synthetic difference-in-differences command. This match allows each treatment's matched control unit to be a weighted average of multiple units in the donor pool in (2), which has a size of 100 for each treatment. The match uses the following x-covariates: treatment year city-level poverty rate, share white, share with a high school degree or less education, and log population.

5. Run the synthetic difference-in-differences command to obtain a treatment event specific estimate for each treatment, $\hat{\tau}_i$. This command requires the treatment and synthetic control units to have matched pre-treatment trends, but permits pre-treatment levels to differ, similar to a traditional difference-in-differences model. The average outcomes for all treated units versus all synthetic control units is shown in the bottom panel of Figure B.7.

6. Estimate a standard error for each $\hat{\tau}_i$ using placebo methods. For each treatment, randomly draw a control unit from the donor pool in (2) and assign this unit as the treated agency. Estimate the synthetic control estimate for this placebo agency, $\hat{\tau}_b$. Repeat this exercise 100 times to obtain the distribution of $\hat{\tau}_b$; use the standard deviation of this distribution as the estimate of the standard error, $se(\hat{\tau}_i)$.

7. Use $\hat{\tau}_i$ as the outcome for the heterogeneity tests in Table B.3. This exercise asks how treatment agency and incident characteristics relate to the size of the arrest treatment effect for the first treatment month. Weight these regressions by $1/se(\hat{\tau}_i)^2$ using the estimates from (5) and (6).

8. From the tests in (7), estimate predictions of $\hat{\tau}_i$ (for the one month arrest effect) based on pre-treatment characteristics, $E(\hat{\tau}_i|X)$. This prediction is constructed using a "leave-out" version of Column 3 of Table B.3, where each estimate of $E(\hat{\tau}_i|X)$ is determined from

all treatments other than i.

9. Bin treatment events by values of $E(\hat{\tau}_i|X)$ for arrests into: top quartile, bottom quartile and inter-quartile range. Within these bins, plot the distribution of $\hat{\tau}_i$ for arrest and crime outcomes over time. This analysis shows how the crime and arrest effects vary for agencies with "predicted" large versus small arrest declines, where predictions only leverage variation in pre-treatment covariates. This analysis is shown in Figure 2.5. Summary statistics of treatment events/agencies in each bin are shown in Table B.4.

We next note additional features of the analysis. First, to the extent possible, we use a common set of matching variables in the nearest neighbor, synthetic control, and heterogeneity prediction table described above, for consistency and transparency.

Second, we define our donor pool for the synthetic control exercise using the 100 nearest neighbors determined from the Nearest Neighbor matching algorithm. We do this for computational reasons. Our synthetic control analysis takes approximately one week to run on our machines, and this run-time increases when more units are added to the donor pool.

B.4 Google Search Trends Description

Each search term is an exact first and last name for the individual in the U.S. state where the death occurred. We identify high-profile civilian deaths using a list compiled by *Black Lives Matter*, and identify officer deaths by linking the FBI LEOKA data we use in this project to records from the *Officer Down Memorial Page* to obtain officer names. Each search is centered around the time period of -1. Further, each search is benchmarked by topical searches for the most common cause of death, heart disease, which is relatively stable in popularity across time and locations within the U.S. Google Trends plots relative search intensity with a maximum search popularity in each search of 100. A benchmark would not be necessary if Google Trends data contained absolute search volume, but unfortunately this data series only includes relative measures of search volume that are a function of the topics and terms used to pull the data. The use of a benchmark is therefore critical to this analysis, as it helps to rescale other outcomes in terms of their importance over time and across geographic areas.

B.5 Data Appendix

B.5.1 Data Sources

Law Enforcement Officers Killed or Assaulted (UCR LEOKA) The FBI's Law Enforcement Officers Killed or Assaulted (LEOKA) data set contains detailed information on total officer employment and officers that are killed or assaulted in the field for each month. We use officers feloniously killed in the line-of-duty as a measure of officer deaths and all assaults on sworn officers whether or not the officers suffered injuries. We verify each officer fatality event in the sample using the web resource *Officer Down Memorial Page* (ODMP) and exclude death events from LEOKA that are not able to be verified in ODMP. This website is also used to gather characteristics of the fatality event and officer who was killed, which is used in the heterogeneity analysis. We utilize the version cleaned and formatted by Jacob Kaplan available from ICPSR (Kaplan, 2020b). This dataset covers the period 2000-2018.

Crime Offense Data (UCR Crime) and Arrest Data (UCR Arrest) The Uniform Crime Report Offenses Known and Clearances By Arrest (UCR Crime) data set contains offenses reported to law enforcement agencies. The crimes reported are homicide, forcible rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft for each month. The Uniform Crime Report Arrests by Age, Sex, and Race (UCR Arrest) data set contains the number of arrests for each crime type by age, sex and race at the month level. We use the total arrests and arrest sub-types in our analysis. We utilize the version cleaned and formatted by Jacob Kaplan available from ICPSR (Kaplan, 2020a). This dataset covers the period 2000-2018. We include all departments that consistently and continuously report monthly data on *both* crime and arrests for at least 9 years in this period, up until and including the last year of the data, 2018.

Use-of-Force Data (UCR Supplementary Homicide Reports) The Uniform Crime Report Supplementary Homicide Reports (UCR Supplementary Homicide Reports) data set contains the number of homicides. We utilize the version cleaned and formatted by Jacob Kaplan available from ICPSR (Kaplan, 2020c) covering the period 2000-2018. We use the "felons killed by police" circumstance in our analysis after restricting the sample to the agencies with other UCR outcomes. We exclude treatment events in which a suspect was killed during the officer fatality event in order to measure the police behavioral response to an officer fatality, rather than features of the event itself.

Use-of-Force Data (Fatal Encounters) Fatal Encounters is a national crowd-sourced database of all deaths through police interaction. We remove suicidal deaths from our analysis and restrict the sample to the agencies with other UCR outcomes. As in the UCR

Supplementary Homicide Report, we exclude treatment events in which a suspect was killed during the officer fatality event. Fatal Encounters was established in 2013 and backfills earlier record years which causes quality to decrease in earlier record years. To address this issue, we restrict attention to the period 2010-2018.

Employment Data: Florida Department of Law Enforcement (FDLE) Florida Department of Law Enforcement (FDLE) has information on all officer employment spells employed including the employing agency, start and end dates of the spell and the reason for separation. We restrict attention to all law enforcement agency officer employment spells that cover the period 2000 to 2016.

Traffic Stop Data We use the standardized traffic stop data from the Stanford Open Policing Project. Each row of the data represents a traffic stop that include information on date, location, subject and officer characteristics and stop characteristics. We collapse the data at city-month level and drop the first and last month for each city to account for incomplete months. We then use the intersection between this data set and our analysis sample.

Traffic Accident Data: Fatality Analysis Reporting System (FARS) We use the Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA) to create measure of traffic fatalities and those involving alcohol. The data include information on fatal injuries in a vehicle crashes. We collapse the accident-level data at city-month level to generate counts. For the accidents involving alcohol, we use the number of drunk drivers involved in a crash. This data element is most reliable from 2008 to 2014 when drivers with the blood alcohol concentration (BAC) 0.01 g/dL or greater are counted. Prior to 2008, all individuals involved in accidents are counted. After 2014, the BAC level measure is changed to 0.001 g/dL or greater for counting. The data covers 2000 to 2018 for any accidents and 2008 to 2014 for accidents involving alcohol.

911 Call Dispatch Data We have hand-collected administrative 911 dispatch call records through submitting open-records requests to cities across the U.S. The data sets for each city vary in the way that they record calls and must be cleaned in order to harmonize the data across cities. Each data set collected is first cleaned to categorize calls into records of interactions that were initiated by officers and civilian complainant calls. Officer-initiated interactions are sometimes included in dispatch data when an officer reports his location in such an interaction to a dispatcher and these may include records of officers assisting other officers in distress, assisting the fire department, or responding to traffic violations. We also calculate the share of calls that result in an officer writing a crime incident report or "Crime Report Rate (911 Calls)" through examining the outcome or disposition of each call which is coded as a field in our data. Lastly, we also construct a measure of officer presence. In 51 of the 56 cities in this sample, we geocode calls to Census block groups and we calculate the share of Census block groups with a 911 call or officer-initiated interaction.

Demographic Data (U.S. Census and American Community Survey) We use the 2000 United States Census and the American Community Survey (ACS) 5-year estimates from 2010 to 2018 to provide information on city characteristics. Specifically, we report each city's population, share Black, Hispanic and white, share male, the share of female-headed household, the share in each age category, the share in each education category, the unemployment rate, the poverty rate and median household income. We linearly interpolate these covariates for the years 2001 to 2009.

B.5.2 Sample Restrictions

The UCR data suffer from reporting and measurement issues. To alleviate concerns about data quality, we take following procedures to extensively clean the outcomes of interest. First, we restrict our analysis to municipal police departments serving cities with population larger 2,000 residents and to the period 2000-2018. Then, we keep departments that consistently report these outcomes after replacing any negative arrest or crime values as missing. Specifically, we only retain agencies that report both crimes and arrests monthly each year in the period 2000-2018 (for example, this procedure drops agencies that report annually or biannually). To increase sample size, we include any agency that reports at least 9 years of consecutive data through 2018, or agencies that begin reporting between 2000-2010.

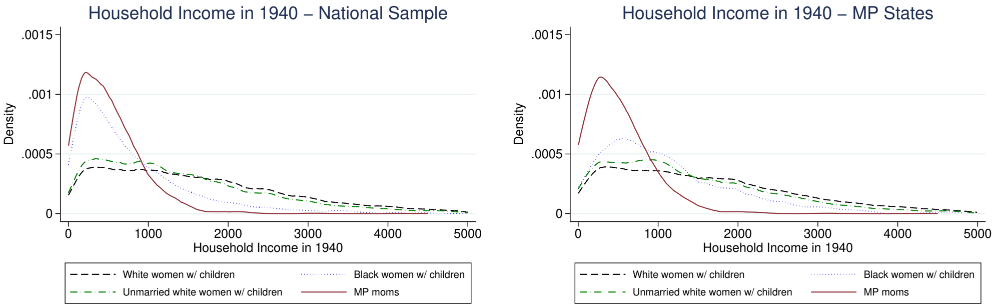
We merge the UCR data together using the originating agency identifiers, the Traffic Stop, FARS and 911 Calls data using the city name and Census data using the Federal Information Processing Standards (FIPS) Place code.

Appendix C

Appendix to "The Impact of Cash Transfers to Poor Mothers on Family Structure and Maternal Well-Being"

C.1 Figures & Tables

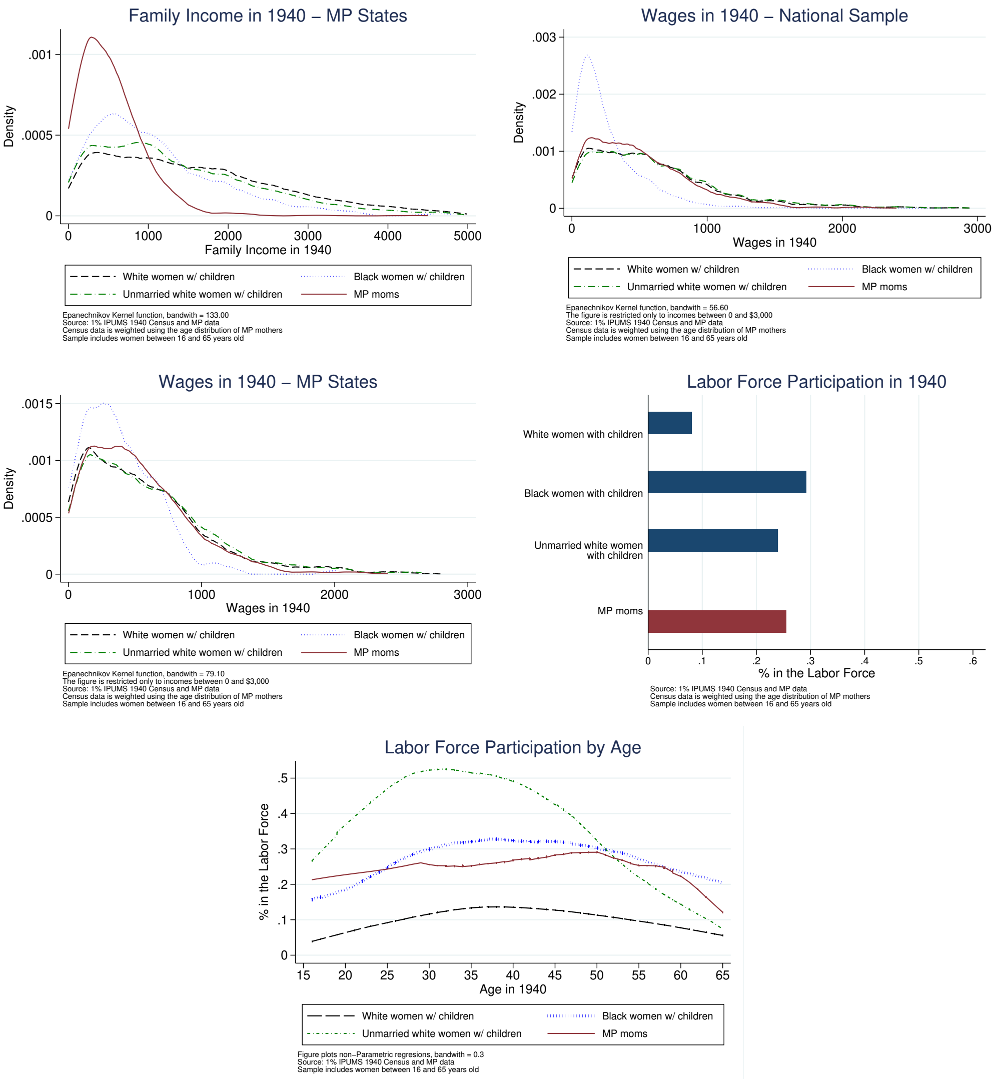
Figure C.1: Census Figures



Epanechnikov Kernel function, bandwith = 95.00
Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

Epanechnikov Kernel function, bandwith = 133.60
Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

Figure C.1: Census Figures (Continued)



Family Income in 1940 – MP States

Epanechnikov Kernel function, bandwith = 133.00
Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

Wages in 1940 – National Sample

Epanechnikov Kernel function, bandwith = 56.60
The figure is restricted only to incomes between 0 and $3,000
Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

Wages in 1940 – MP States

Epanechnikov Kernel function, bandwith = 79.10
The figure is restricted only to incomes between 0 and $3,000
Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

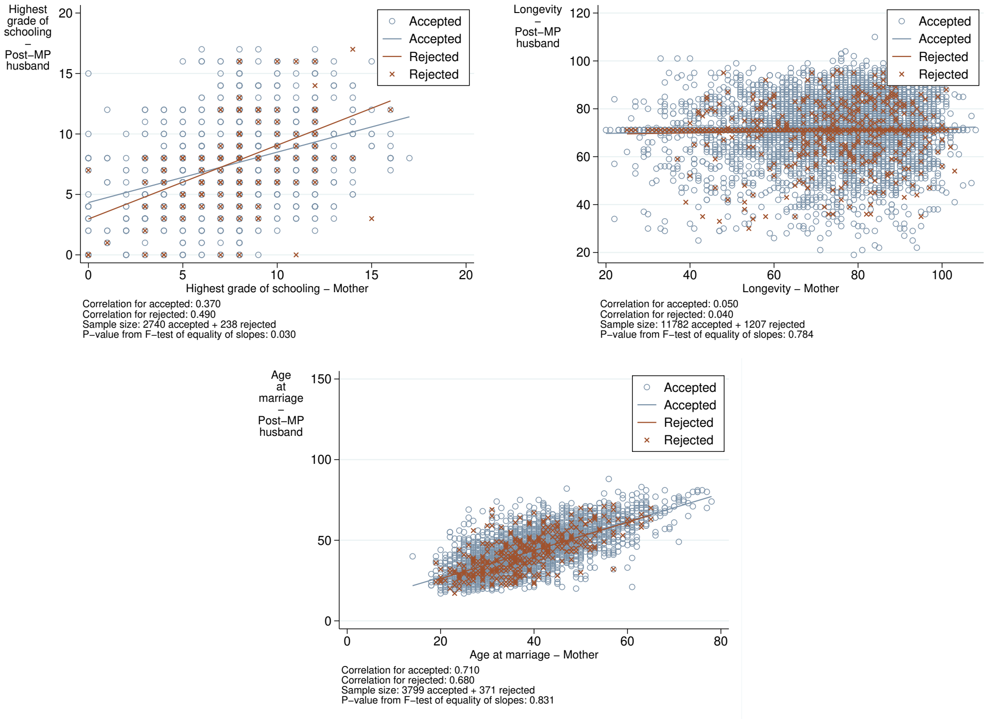
Labor Force Participation in 1940

Source: 1% IPUMS 1940 Census and MP data
Census data is weighted using the age distribution of MP mothers
Sample includes women between 16 and 65 years old

Labor Force Participation by Age

Figure plots non–Parametric regresions, bandwith = 0.3
Source: 1% IPUMS 1940 Census and MP data
Sample includes women between 16 and 65 years old

Figure C.2: Cash transfers do not change the degree of assortative mating in education, longevity and age at marriage



Correlation for accepted: 0.370
Correlation for rejected: 0.490
Sample size: 2740 accepted + 238 rejected
P–value from F–test of equality of slopes: 0.030

Correlation for accepted: 0.050
Correlation for rejected: 0.040
Sample size: 11782 accepted + 1207 rejected
P–value from F–test of equality of slopes: 0.784

Correlation for accepted: 0.710
Correlation for rejected: 0.680
Sample size: 3799 accepted + 371 rejected
P–value from F–test of equality of slopes: 0.831

Figure C.3: Share of MP applicants remarrying by age

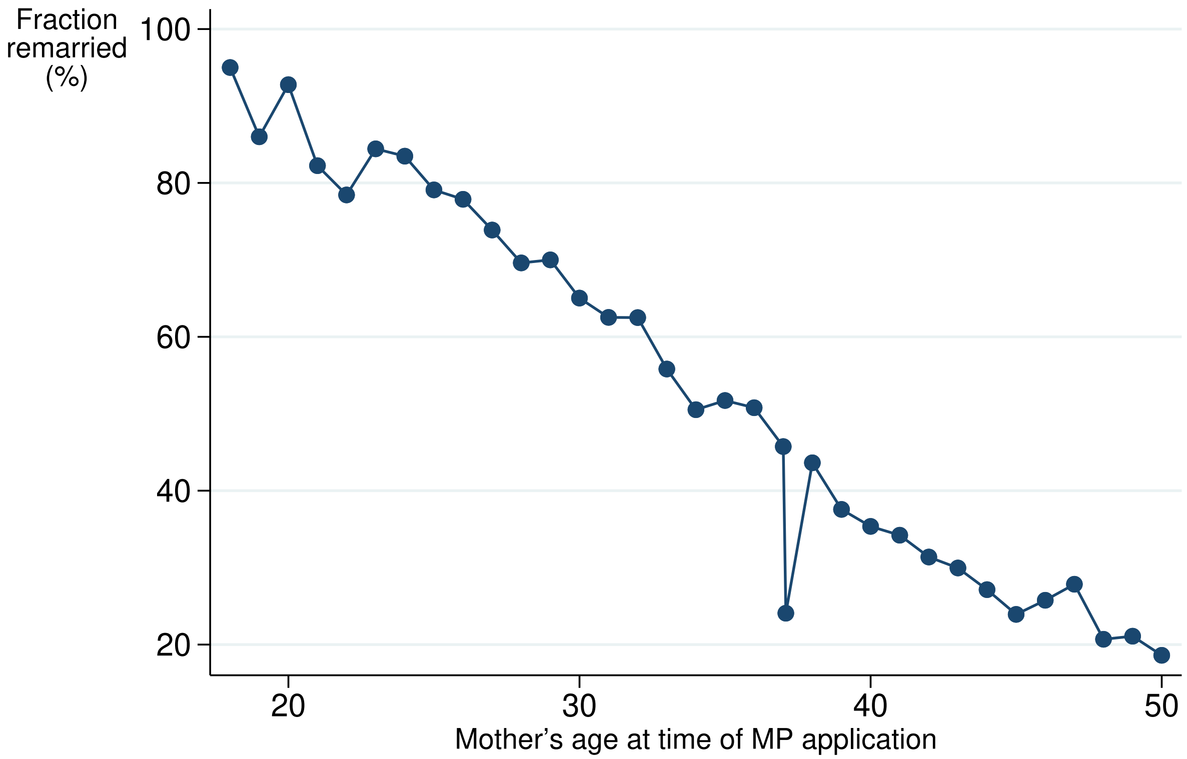
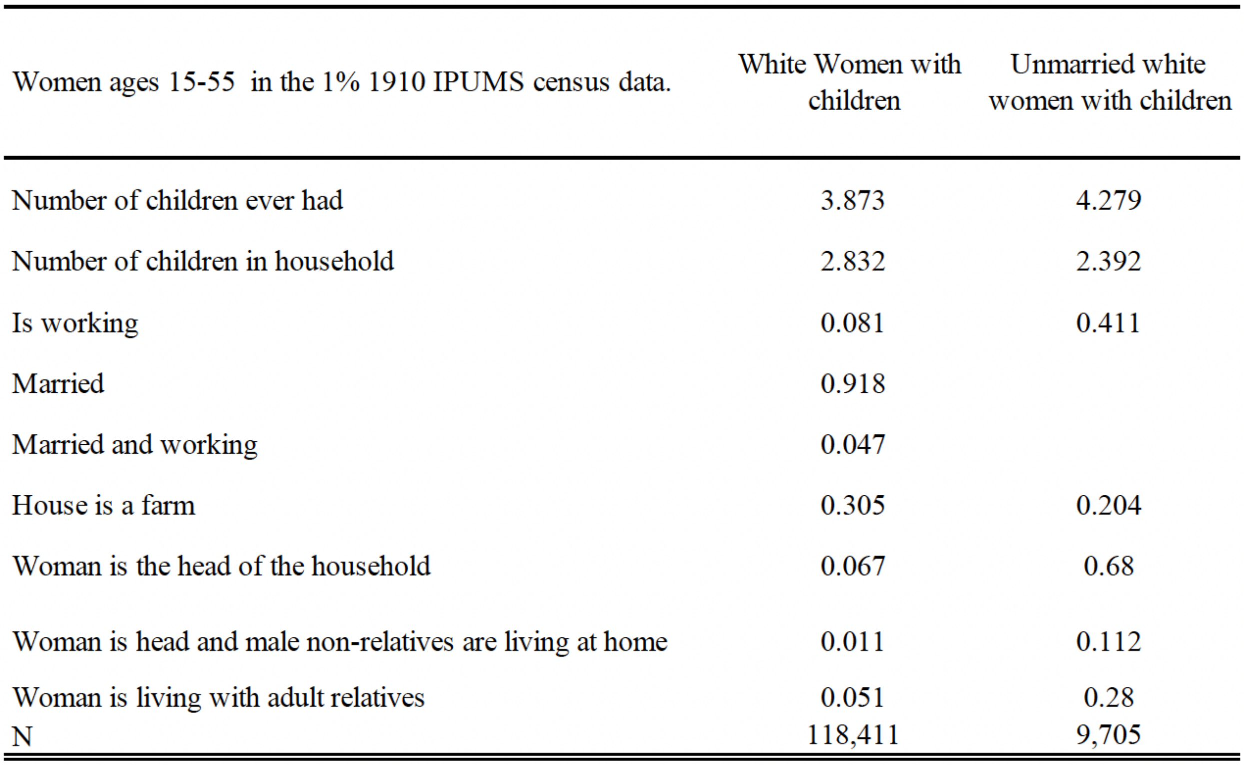


Figure is restricted to mothers between 18 and 50 years old at the time of application that had information on remarriage. 78.10000000000001% of the sample.

Notes: The figure plots the fraction remarrying by age. The 5th percentile of the age at remarriage is 24 and the 95th is 52.

Table C.1: The Status of Poor Women with Children in 1910



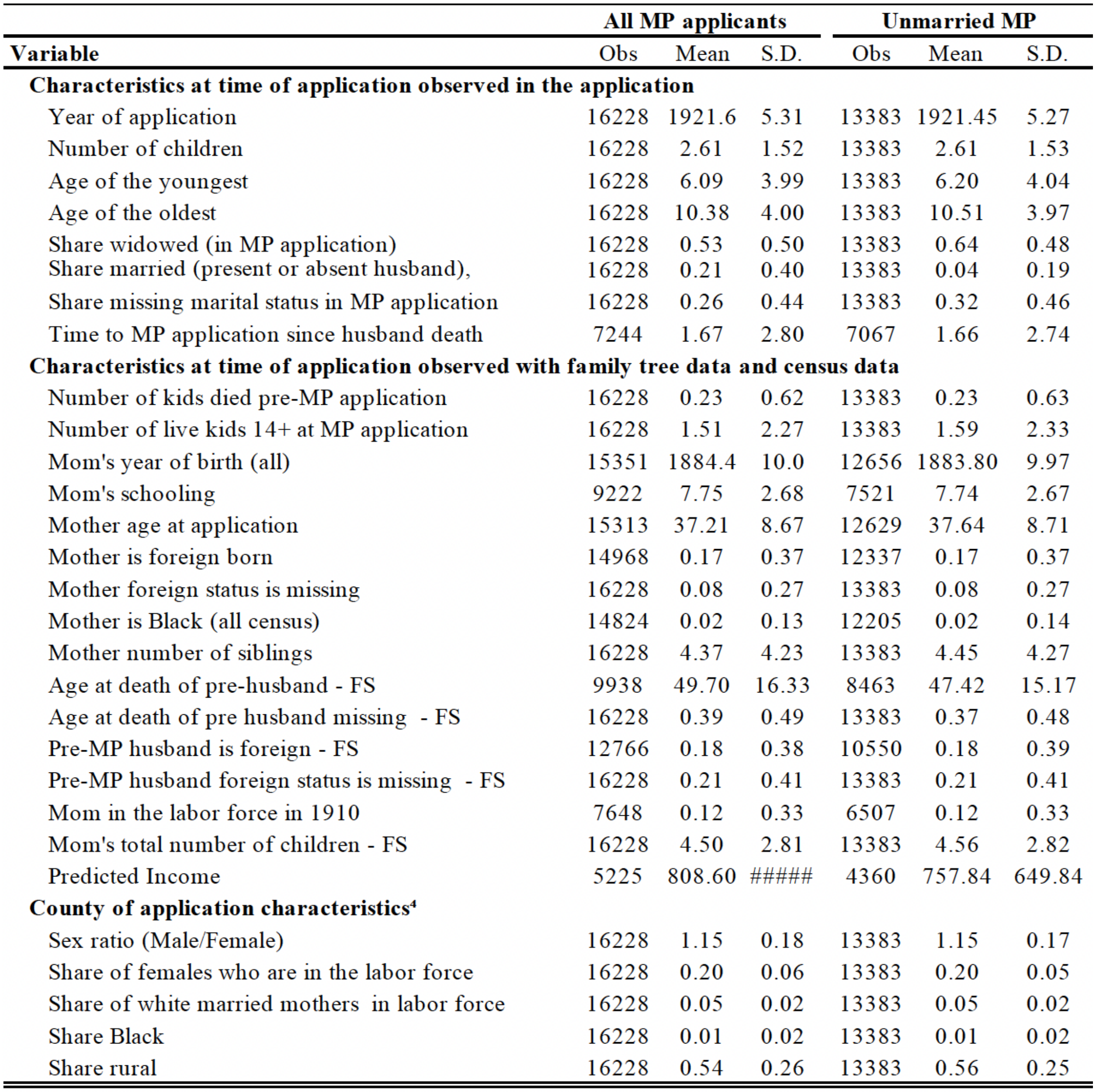
Women ages 15-55 in the 1% 1910 IPUMS census data.	White Women with children	Unmarried white women with children
Number of children ever had	3.873	4.279
Number of children in household	2.832	2.392
Is working	0.081	0.411
Married	0.918	
Married and working	0.047	
House is a farm	0.305	0.204
Woman is the head of the household	0.067	0.68
Woman is head and male non-relatives are living at home	0.011	0.112
Woman is living with adult relatives	0.051	0.28
N	118,411	9,705

Notes: Author's computation using data from the 1910 census.

Table C.2: Summary Statistics for MP Applicants

	All MP applicants			**Unmarried MP**		
Variable	Obs	Mean	S.D.	Obs	Mean	S.D.
Found remarriage information	16228	0.84	0.37	13383	0.84	0.36
Share accepted	16228	0.90	0.30	13383	0.90	0.30
Dependent variables						
Remarrriage rates						
Mom ever remarried	13638	0.47	0.50	11286	0.48	0.50
% remarried within 1 years[2]	11509	0.02	0.15	9423	0.03	0.16
% remarried within 2 years	11509	0.08	0.28	9423	0.09	0.29
% remarried within 3 years	11509	0.14	0.34	9423	0.15	0.35
% remarried within 5 years	11509	0.21	0.41	9423	0.22	0.41
Among moms that remarried						
Duration to remarriage in years	4255	6.71	7.73	3572	6.36	7.55
Mom age at remarriage	4240	38.89	9.98	3558	38.77	9.80
Post-MP husband						
age at remarriage - FS	4179	43.31	12.63	3507	43.27	12.53
longevity - FS	6384	71.30	12.02	5435	71.28	12.04
died before 1940 - FS	4850	0.18	0.38	4123	0.19	0.39
wage income - 1940	3301	693.60	770.05	2815	674.77	759.27
highest schooling grade - 1940	3460	7.59	2.75	2955	7.56	2.72
occ earnings score - latest census[3]	3932	40.49	29.56	3328	39.68	29.62
occ income score - latest census[3]	4206	20.24	10.79	3556	20.09	10.83
was a farmer - latest census[3]	5264	0.11	0.31	4457	0.12	0.32
lives in owned housing unit - 1920	2843	0.56	0.50	2418	0.57	0.49
foreign born - FS	5522	0.16	0.37	4673	0.16	0.36
foreign status is missing in FS	6384	0.14	0.34	5435	0.14	0.35
No. of children at time of marriage - FS	4255	0.56	1.11	3572	0.57	1.10
Quality of match						
Age gap - FS	5771	4.22	8.68	4874	4.32	8.71
Education gap - 1940	2978	-0.23	2.88	2545	-0.23	2.83
Other Maternal outcomes						
Mom's longevity	12989	74.29	15.04	10749	74.32	14.84
Mom died before 1940	13064	0.17	0.38	10810	0.18	0.38
Mom's income in 1940	8226	130.3	306.9	6697	125.40	305.68
Mom's occupation score 1940	9358	4.66	8.81	7635	4.48	8.67
Mom in the labor force in 1940	9351	0.26	0.44	7630	0.25	0.43
Mom worked in 1940	9358	0.24	0.42	7635	0.23	0.42
Mom was married in 1940	9330	0.45	0.50	7615	0.42	0.49
Mom's household income in 1940	9070	956.0	1050.3	7398	955.59	1053.2
Mom's no. of own kids living together in 1940	9358	1.74	1.59	7635	1.71	1.57
Number of kids born after MP application	16228	0.27	0.83	13383	0.26	0.82

Table C.2: Summary Statistics for MP Applicants (Continued)



Variable	All MP applicants			Unmarried MP		
	Obs	Mean	S.D.	Obs	Mean	S.D.
Characteristics at time of application observed in the application						
Year of application	16228	1921.6	5.31	13383	1921.45	5.27
Number of children	16228	2.61	1.52	13383	2.61	1.53
Age of the youngest	16228	6.09	3.99	13383	6.20	4.04
Age of the oldest	16228	10.38	4.00	13383	10.51	3.97
Share widowed (in MP application)	16228	0.53	0.50	13383	0.64	0.48
Share married (present or absent husband),	16228	0.21	0.40	13383	0.04	0.19
Share missing marital status in MP application	16228	0.26	0.44	13383	0.32	0.46
Time to MP application since husband death	7244	1.67	2.80	7067	1.66	2.74
Characteristics at time of application observed with family tree data and census data						
Number of kids died pre-MP application	16228	0.23	0.62	13383	0.23	0.63
Number of live kids 14+ at MP application	16228	1.51	2.27	13383	1.59	2.33
Mom's year of birth (all)	15351	1884.4	10.0	12656	1883.80	9.97
Mom's schooling	9222	7.75	2.68	7521	7.74	2.67
Mother age at application	15313	37.21	8.67	12629	37.64	8.71
Mother is foreign born	14968	0.17	0.37	12337	0.17	0.37
Mother foreign status is missing	16228	0.08	0.27	13383	0.08	0.27
Mother is Black (all census)	14824	0.02	0.13	12205	0.02	0.14
Mother number of siblings	16228	4.37	4.23	13383	4.45	4.27
Age at death of pre-husband - FS	9938	49.70	16.33	8463	47.42	15.17
Age at death of pre husband missing - FS	16228	0.39	0.49	13383	0.37	0.48
Pre-MP husband is foreign - FS	12766	0.18	0.38	10550	0.18	0.39
Pre-MP husband foreign status is missing - FS	16228	0.21	0.41	13383	0.21	0.41
Mom in the labor force in 1910	7648	0.12	0.33	6507	0.12	0.33
Mom's total number of children - FS	16228	4.50	2.81	13383	4.56	2.82
Predicted Income	5225	808.60	#####	4360	757.84	649.84
County of application characteristics[4]						
Sex ratio (Male/Female)	16228	1.15	0.18	13383	1.15	0.17
Share of females who are in the labor force	16228	0.20	0.06	13383	0.20	0.05
Share of white married mothers in labor force	16228	0.05	0.02	13383	0.05	0.02
Share Black	16228	0.01	0.02	13383	0.01	0.02
Share rural	16228	0.54	0.26	13383	0.56	0.25

Notes: [1]Unmarried MP applicants include widowed, divorced and never married women. [2]People who remarried and have missing dates are dropped. The duration measure starts at 0.5 (the variable is duration + 0.5, so we assume that marriages occur uniformly within a year). We also assume that if women married the same year they applied for the pension (and the exact data of marriage is missing) that the marriage took place after the MP application. [3]Defined from pre marriage data: uses 1940 if available, then 1930, then 1920, then 1910. Never uses a measure that is observed post-MP marriage. [4]Measured in year of application. Yearly measures are constructed through linear interpolation using census data from 1910, 1920 and 1930. All measures use the universe of people who are between 18 and 55 years old. Sample restriction: we drop mothers that applied after 1930 or records for mothers that applied multiple times so mothers only appear once in the data and individuals who we discovered in the family tree were not the mother (a handful of grandmothers, sisters and step-mothers).

Table C.3: Accepted Moms are slightly worse off at time of application

	MP admin data	Newly collected data					
Outcome:	Number of kids on application[1]	Number of kids died before MP application	Number of live kids 14+ at MP application	Mom age at application	Mom number of siblings	Mom foreign born	Mom is Black
Panel A: All Moms (County and Year of Application FE)							
Mean of outcome for rejected	2.2000	0.198	1.631	37.824	4.14	0.155	0.017
Accepted	0.421	0.023	-0.193	-0.712	0.105	0.006	0.004
OLS (unadjusted se)	(0.042)***	(0.017)	(0.063)***	(0.253)***	(0.115)	(0.010)	(0.004)
Robust standard errors	[0.038]***	[0.016]	[0.067]***	[0.267]***	[0.117]	[0.010]	[0.004]
Clustered at county	{0.058}***	{0.016}	{0.072}***	{0.272}***	{0.130}	{0.009}	{0.004}
Clustered at county*year	(0.045)***	(0.016)	(0.069)***	(0.274)***	(0.114)	(0.010)	(0.005)
Observations	16228	16228	16228	15313	16228	14968	14824
R-squared	0.083	0.052	0.041	0.033	0.094	0.125	0.058
Panel B: Unmarried Moms (County and Year of Application FE)							
Mean of outcome for rejected	2.182	0.196	1.727	38.372	4.22	0.159	0.018
Accepted	0.441	0.034	-0.224	-0.779	0.049	0.007	0.003
OLS (unadjusted se)	(0.046)***	(0.019)*	(0.071)***	(0.278)***	(0.127)	(0.012)	(0.004)
Robust standard errors	[0.042]***	[0.018]*	[0.076]***	[0.296]***	[0.131]	[0.012]	[0.005]
Clustered at county	{0.057}***	{0.019}*	{0.097}**	{0.318}**	{0.139}	{0.009}	{0.005}
Clustered at county*year	(0.047)***	(0.017)*	(0.080)***	(0.311)**	(0.133)	(0.011)	(0.005)
Observations	13383	13383	13383	12629	13383	12337	12205
R-squared	0.092	0.058	0.047	0.044	0.101	0.130	0.063

Table C.3: Accepted Moms are slightly worse off at time of application (Continued)

	Newly collected data						
Outcome:	In labor force 1910	Work 1910	Occupational score 1910[2]	Mom education 1940	Years from Pre-MP husband death[3]	Longevity of Pre-MP husband	Predicted Income (based on Iowa census data)
Panel A: All Moms (County and Year of Application FE)							
Mean of outcome for rejected	0.140	0.151	2.407	7.654	2.214	51.418	824.642
Accepted	-0.007	-0.013	-0.286	0.018	-0.38	-1.759	-50.700
OLS (unadjusted se)	(0.013)	(0.012)	(0.228)	(0.107)	(0.127)***	(0.608)***	(34.287)
Robust standard errors	[0.014]	[0.013]	[0.252]	[0.106]	[0.140]***	[0.637]***	[33.650]
Clustered at county	{0.011}	{0.010}	{0.312}	{0.102}	{0.119}***	{0.591}***	{29.663}*
Clustered at county*year	(0.013)	(0.013)	(0.251)	(0.110)	(0.132)***	(0.672)***	(32.908)
Observations	7648	8953	8953	9222	7244	9938	5332
R-squared	0.033	0.039	0.032	0.064	0.067	0.076	0.152
Panel B: Unmarried Moms (County and Year of Application FE)							
Mean of outcome for rejected	0.141	0.153	2.515	7.712	2.222	49.083	768.819
Accepted	-0.006	-0.01	-0.309	-0.051	-0.395	-1.672	-52.257
OLS (unadjusted se)	(0.014)	(0.013)	(0.257)	(0.119)	(0.125)***	(0.614)***	(38.093)
Robust standard errors	[0.015]	[0.014]	[0.283]	[0.115]	[0.141]***	[0.656]**	[37.794]
Clustered at county	{0.011}	{0.010}	{0.282}	{0.112}	{0.121}***	{0.717}**	{37.969}
Clustered at county*year	(0.015)	(0.013)	(0.265)	(0.119)	(0.134)***	(0.699)**	(38.244)
Observations	6507	7515	7515	7521	7067	8463	4453
R-squared	0.039	0.044	0.037	0.063	0.071	0.076	0.207

Notes: Controls include county and year of application fixed effects. The sample drops mothers that applied after 1930, and applications made by a person who is not the mother, keeps only the observations of the first successful attempt (It keeps the application with more children listed if multiple successful applications in the same year. Keep the smallest FS ID if applied successfully more than once the same year, with the same number of children.) The predicted income is obtained using the 1915 Iowa census to estimate the coefficients to predict income for all recipients. The regression includes only the covariates observed in both our data and the Iowa census. It includes widow status, mother's age, number of kids, number of kids at each age, age of youngest and oldest kid at application, number of kids over 14 years old at application, an indicator if the mother is foreign-born, and indicator of being Black, schooling and occupation score.[1]Only includes kids with eligible age. [2]Occupational score inputs zeros for mothers out of the labor force. [3]Death to MP application if >0.

Table C.4: Does accepted status predict missing data for marriage outcomes?

		Data for women known to have remarried		
		Missing Family Search variables		
Outcome:	Remarriage information missing	Duration until remarriage	Age gap	Post-MP Husband Longevity
Panel A: All Moms (County and year FE)				
Mean of outcome for rejected	0.205	0.355	0.121	0.298
Accepted	-0.039	-0.023	-0.024	-0.050
OLS (unadjusted se)	(0.010)***	(0.021)	(0.014)*	(0.020)**
Robust standard errors	[0.011]***	[0.022]	[0.015]	[0.021]**
Clustered at county	{0.014}***	{0.019}	{0.013}*	{0.018}***
Clustered at county*year	(0.011)***	(0.021)	(0.016)	(0.021)**
Observations	16228	6384	6384	6384
R-squared	0.045	0.140	0.035	0.044
Panel B: All Moms (All Controls)				
Mean of outcome for rejected	0.205	0.355	0.121	0.298
Accepted	-0.009	-0.022	-0.020	-0.040
OLS (unadjusted se)	(0.009)	(0.020)	(0.013)	(0.020)**
Robust standard errors	[0.010]	[0.021]	[0.015]	[0.021]*
Clustered at county	{0.011}	{0.022}	{0.014}	{0.019}**
Clustered at county*year	(0.010)	(0.022)	(0.016)	(0.021)*
Observations	16228	6384	6384	6384
R-squared	0.294	0.205	0.114	0.085
Panel C: Unmarried Moms (County and year FE)				
Mean of outcome for rejected	0.203	0.370	0.127	0.300
Accepted	-0.038	-0.033	-0.019	-0.048
OLS (unadjusted se)	(0.011)***	(0.023)	(0.015)	(0.022)**
Robust standard errors	[0.012]***	[0.024]	[0.017]	[0.023]**
Clustered at county	{0.016}**	{0.021}	{0.014}	{0.020}**
Clustered at county*year	(0.014)***	(0.023)	(0.017)	(0.023)**
Observations	13383	5435	5435	5435
R-squared	0.052	0.145	0.041	0.049
Panel D: Unmarried Moms (All Controls)				
Mean of outcome for rejected	0.203	0.370	0.127	0.300
Accepted	-0.009	-0.039	-0.019	-0.045
OLS (unadjusted se)	(0.010)	(0.022)*	(0.015)	(0.021)**
Robust standard errors	[0.010]	[0.023]*	[0.016]	[0.023]**
Clustered at county	{0.013}	{0.021}*	{0.015}	{0.021}**
Clustered at county*year	(0.012)	(0.023)*	(0.016)	(0.022)**
Observations	13383	5435	5435	5435
R-squared	0.307	0.216	0.125	0.096

Table C.4: Does accepted status predict missing data for marriage outcomes? (Continued)

	Data for women known to have remarried				
	Missing Post-MP census variables				
Outcome:	Education (1940)	Education gap (1940)	Occupational score (earliest)	Farmer (earliest)	Income (1940)
Panel A: All Moms (county and year FE)					
Mean of outcome for rejected	0.513	0.599	0.325	0.199	0.539
Accepted	-0.036	-0.042	-0.010	-0.021	-0.043
OLS (unadjusted se)	(0.023)	(0.023)*	(0.021)	(0.018)	(0.023)*
Robust standard errors	[0.023]	[0.022]*	[0.021]	[0.019]	[0.023]*
Clustered at county	{0.020}*	{0.015}***	{0.020}	{0.015}	{0.023}*
Clustered at county*year	(0.022)	(0.020)**	(0.025)	(0.019)	(0.022)*
Observations	6384	6384	6384	6384	6384
R-squared	0.048	0.049	0.055	0.032	0.045
Panel B: All Moms (All Controls)					
Mean of outcome for rejected	0.513	0.599	0.325	0.199	0.539
Accepted	-0.024	-0.030	-0.001	-0.014	-0.030
OLS (unadjusted se)	(0.023)	(0.023)	(0.021)	(0.017)	(0.023)
Robust standard errors	[0.022]	[0.022]	[0.021]	[0.018]	[0.023]
Clustered at county	{0.020}	{0.015}*	{0.021}	{0.016}	{0.023}
Clustered at county*year	(0.021)	(0.020)	(0.024)	(0.018)	(0.022)
Observations	6384	6384	6384	6384	6384
R-squared	0.102	0.102	0.079	0.079	0.098
Panel C: Unmarried Moms (County and year FE)					
Mean of outcome for rejected	0.511	0.606	0.330	0.203	0.535
Accepted	-0.033	-0.046	-0.011	-0.019	-0.039
OLS (unadjusted se)	(0.025)	(0.025)*	(0.023)	(0.020)	(0.025)
Robust standard errors	[0.025]	[0.024]*	[0.024]	[0.020]	[0.025]
Clustered at county	{0.019}*	{0.017}***	{0.020}	{0.018}	{0.022}*
Clustered at county*year	(0.023)	(0.022)**	(0.024)	(0.020)	(0.023)*
Observations	5435	5435	5435	5435	5435
R-squared	0.055	0.057	0.059	0.034	0.049
Panel D: Unmarried Moms (All Controls)					
Mean of outcome for rejected	0.511	0.606	0.330	0.203	0.535
Accepted	-0.027	-0.038	-0.005	-0.015	-0.029
OLS (unadjusted se)	(0.025)	(0.025)	(0.023)	(0.019)	(0.025)
Robust standard errors	[0.024]	[0.024]	[0.023]	[0.020]	[0.025]
Clustered at county	{0.019}	{0.017}**	{0.022}	{0.019}	{0.022}
Clustered at county*year	(0.022)	(0.022)*	(0.023)	(0.019)	(0.022)
Observations	5435	5435	5435	5435	5435
R-squared	0.117	0.115	0.087	0.088	0.108

Table C.4: Does accepted status predict missing data for marriage outcomes? (Continued)

	All women					
	Missing Post-MP census variables					
Outcome:	Longevity	Household Income (1940)	LFP (1930)	LFP (1940)	Occupation Score (1930)	Location (1940)
Panel A: All Moms (county and year FE)						
Mean of outcome for rejected	0.244	0.520	0.396	0.521	0.886	0.520
Accepted	-0.038	-0.069	-0.066	-0.069	-0.036	-0.069
OLS (unadjusted se)	(0.011)***	(0.014)***	(0.013)***	(0.014)***	(0.010)***	(0.014)***
Robust standard errors	[0.012]***	[0.014]***	[0.013]***	[0.014]***	[0.009]***	[0.014]***
Clustered at county	{0.016}**	{0.014}***	{0.016}***	{0.014}***	{0.012}***	{0.014}***
Clustered at county*year	(0.012)***	(0.014)***	(0.014)***	(0.014)***	(0.010)***	(0.014)***
Observations	16228	16228	16228	16228	16228	16228
R-squared	0.054	0.052	0.088	0.052	0.042	0.052
Panel B: All Moms (All Controls)						
Mean of outcome for rejected	0.244	0.520	0.396	0.521	0.886	0.520
Accepted	-0.006	-0.032	-0.034	-0.032	-0.022	-0.032
OLS (unadjusted se)	(0.009)	(0.013)**	(0.012)***	(0.013)**	(0.010)**	(0.013)**
Robust standard errors	[0.010]	[0.013]**	[0.013]***	[0.013]**	[0.009]**	[0.013]**
Clustered at county	{0.013}	{0.013}**	{0.013}**	{0.013}**	{0.009}**	{0.013}**
Clustered at county*year	(0.011)	(0.013)**	(0.013)***	(0.013)**	(0.009)**	(0.013)**
Observations	16228	16228	16228	16228	16228	16228
R-squared	0.352	0.157	0.215	0.157	0.075	0.157
Panel C: Unmarried Moms (County and year FE)						
Mean of outcome for rejected	0.242	0.529	0.401	0.529	0.888	0.529
Accepted	-0.035	-0.076	-0.066	-0.076	-0.034	-0.076
OLS (unadjusted se)	(0.012)***	(0.015)***	(0.014)***	(0.015)***	(0.011)***	(0.015)***
Robust standard errors	[0.013]***	[0.015]***	[0.015]***	[0.015]***	[0.010]***	[0.015]***
Clustered at county	{0.018}**	{0.016}***	{0.018}***	{0.016}***	{0.010}***	{0.016}***
Clustered at county*year	(0.014)**	(0.015)***	(0.016)***	(0.015)***	(0.011)***	(0.015)***
Observations	13383	13383	13383	13383	13383	13383
R-squared	0.060	0.054	0.091	0.054	0.043	0.054
Panel D: Unmarried Moms (All Controls)						
Mean of outcome for rejected	0.242	0.529	0.401	0.529	0.888	0.529
Accepted	-0.006	-0.041	-0.035	-0.041	-0.021	-0.041
OLS (unadjusted se)	(0.010)	(0.014)***	(0.013)***	(0.014)***	(0.011)*	(0.014)***
Robust standard errors	[0.011]	[0.014]***	[0.014]**	[0.014]***	[0.010]**	[0.014]***
Clustered at county	{0.014}	{0.014}***	{0.014}**	{0.014}***	{0.008}**	{0.014}***
Clustered at county*year	(0.012)	(0.014)***	(0.015)**	(0.014)***	(0.010)**	(0.014)***
Observations	13383	13383	13383	13383	13383	13383
R-squared	0.361	0.159	0.218	0.159	0.077	0.159

Notes: Please refer to Table 3.1 for a full description of the controls, restrictions and checks.

Table C.5: Welfare recipients are not less likely to remarry

Data source:	FamilySearch	Census		
Dependent variable	Ever remarried = 1	Married in 1920	Married in 1930, all	Married in 1940, all
Mean of Y for rejected	0.47	0.39	0.41	0.43
Panel A: County and year FE only				
Accepted	-0.002	-0.084	-0.013	-0.002
Robust standard errors	(0.017)	(0.027)***	(0.020)	(0.022)
R-squared	0.036	0.088	0.073	0.035
Panel B: Main results (Full controls)				
Accepted	-0.014	-0.099	-0.012	-0.006
Robust standard errors	(0.016)	(0.026)***	(0.018)	(0.020)
Clustered at county	[0.020]	[0.022]***	[0.019]	[0.020]
Clustered at county*year	{0.016}	{0.027}***	{0.018}	{0.020}
R-squared	0.228	0.189	0.199	0.219
Observations	11286	3522	9155	7615
Panel C: Checks				
1- Correction for OVB (Oster 2017)	[-0.02;-0.01]	[-0.11;-0.09]	[-0.02;-0.01]	[-0.01;-0.01]
2- Semi-parametric sample selection correction (Newey, 2009)			0.41	
Accepted	-0.014	-0.100	-0.013	-0.005
95% Confidence interval	[-0.05;0.02]	[-0.14;-0.06]	[-0.05;0.03]	[-0.05;0.03]
F-Stat (first stage)	72.37	13.05	24.20	62.77
3- Drop if quality of match low				
Accepted	-0.027	-0.097***	-0.021	0.000
Clustered at county	(0.028)	(0.025)	(0.022)	(0.025)
Observations	5463	1538	4495	3752
4 - IPW	0.009	-0.069***	-0.022	-0.009
	(0.025)	(0.021)	(0.021)	(0.026)
5 - Causal Forest ATE	-0.020	-0.087***	-0.027*	-0.010
	(0.014)	(0.024)	(0.016)	(0.018)
6 - Causal Forest ATT	-0.027	-0.084***	-0.034	-0.012
	(0.020)	(0.030)	(0.023)	(0.025)
Observations	11286	3522	9155	7615

Notes: Sample includes only mothers that were not married at MP application (or whose marital status is missing). See Table 3.1 for other sample restrictions. Panel B controls for county and year-of-application fixed effects and individual, county and state controls. Individual controls: Kids: MP age of the youngest and oldest, MP dummies for number, FS number older than 14, FS number that died before MP, FS number with dates missing. Mother: last name length, dummies for divorced, widowed and missing marital status, age at application, missing age, number of siblings, foreign, missing nativity, first husband's longevity, first husband's longevity is missing. County controls: for ages 18-55: sex ratio (M/F), shares of white married mothers in the labor force, black and rural. County controls match linear interpolated information from the 1910, 1920 and 1930 census with the year of MP application. State controls: manufacturing wages, education/labor laws (age must enter school, work permit age, and continuation school law in place), state expenditures in logs (education, charity, and social programs), state laws concerning MP transfers (work required, reapplication required, maximum amount for the first child and for each additional child). Omitted variable bounds: We use Oster (2017) to construct omitted variable bias (OVB) bounds. We assume that the R-max is 1.3 times greater than the R-squared from panel B. We assume delta = (-1, 1) for lower and upper bounds. Sample Selection Correction: We follow the two-step estimation suggested by Newey (2009) to correct for sample selection. First, we regress the dummy indicating whether the outcome is missing on RA fixed effects (73 dummies) and all other controls. We report the F-statistic of the test of relevance of these dummies. Second, we estimate a linear regression of the outcome on controls and on a fourth degree polynomial of predicted values from the first stage. We jointly bootstrap the two stages and report the 95% bias corrected confidence interval clustered at the county level, from 200 repetitions. Quality of match: Regressions that drop low quality matches (quality measure below its median) include all controls and cluster the standard errors at the county level. The quality of match between census, family search and administrative data is constructed as the weighted sum of variables that access the similarity between first name, last name, full name, age and place of birth in each dataset. IPW: We estimate the average treatment effect using the estimated probability weights to address for potential missing outcomes. The standard errors are clustered at the county level and a logit model is used to predict the accepted status. Causal Forest: We implement the generalized random forest algorithm proposed by Athey, Tibshirani, and Wager (2019). We estimate the average treatment effects using a doubly robust augmented-inverse-propensity weighting estimation method and report the ATE and ATT. See Appendix for more details.

Table C.6: Does welfare increase quality of Post-MP husband?

Data source:	Family Search		Censuses					Summary index using…		
Outcome:	Post-MP Husband Longevity	Age gap (shifted by 2.5 years)[1]	Occ Score[2]	Mean Occ Wage, 1990	Median Occ Earning, 1900	Post-MP Husband Education	Education gap[3]	Equal weights [4]	Equal weights (no age, education gap)	Satisfaction weights [5]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: County and year FE										
Mean of outcome for rejected	70.13	6.66	21.22	14558.51	595.66	7.80	1.82	-0.05	-0.05	0.36
Accepted	1.800	0.230	-0.719	-597.405	-16.204	-0.374	-0.085	0.096	0.080	-0.010
Robust standard errors	(0.851)**	(0.312)	(0.727)	(567.974)	(11.701)	(0.211)*	(0.162)	(0.057)*	(0.055)	(0.018)
Clustered at county	[0.917]*	[0.272]	[0.555]	[589.140]	[14.783]	[0.219]*	[0.183]	[0.045]**	[0.044]*	[0.021]
Clustered at county*year	{0.871}**	{0.293}	{0.714}	{508.397}	{11.837}	{0.206}*	{0.171}	{0.053}*	{0.053}	{0.018}
Observations	4104	4874	3556	4178	4366	2955	2545	4894	4606	2540
Panel B: control for pre-determined variables and other inputs										
Mean of outcome for rejected	73.99	6.345	20.18	13770.60	602.61	7.946	1.818			
Accepted	1.368	0.247	-0.425	455.806	-16.026	-0.334	0.031			
Robust standard errors	(1.136)	(0.502)	(0.999)	(841.439)	(18.781)	(0.270)	(0.192)			
Clustered at county	(1.309)	(0.599)	(0.749)	[957.851]	[18.394]	(0.279)	(0.239)			
Clustered at county*year	{1.097}	{0.482}	{0.962}	{809.914}	{18.955}	{0.277}	{0.210}			
Observations	1,887	1,887	1,887	1866	1887	1,887	1,887			

Table C.6: Does welfare increase quality of Post-MP husband? (Continued)

Data source:	Family Search		Censuses					Summary index using…		
Outcome:	Post-MP Husband Longevity	Age gap (shifted by 2.5 years)[1]	Occ Score[2]	Mean Occ Wage, 1990	Median Occ Earning, 1900	Post-MP Husband Education	Education gap[3]	Equal weights [4]	Equal weights (no age, education	Satisfaction weights [5]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel C: Checks (for panel B)										
1- Correction for O\	[1.32;1.43]	[0.15;0.36]	[-0.45;-0.39]	[138.73;831.53]	[-18.37;-13.24]	[-0.34;-0.33]	[-0.00;0.06]			
2- Semi-parametric sample selection correction (Newey, 2009)										
Accepted	1.368	0.247	-0.425	442.166	-16.026	-0.334	0.031			
95% Confidence int	[-1.22;3.96]	[-0.94;1.43]	[-1.91;1.06]	[-1452.85;2337.18]	[-52.42;20.37]	[-0.89;0.22]	[-0.44;0.50]			
F-Stat	.	.	.			.	.			
Observations	1,887	1,887	1,887	1887	1887	1,887	1,887			
3- Drop if quality of match low										
Accepted	1.397	0.433	0.099	980.055	1.624	-0.364	-0.010			
Clustered at county	(1.464)	(0.713)	(0.973)	(1294.019)	(25.515)	(0.308)	(0.239)			
Observations	1305	1305	1305	1291	1305	1305	1305			

Notes: Standard errors are clustered at the county level. Please refer to Table 3.1 for a full description of the controls, restrictions and checks. Panel C includes the other inputs (Post-MP Husband longevity, age gap, Post-MP Husband latest occupational score, Post-MP Husband 1940 education and education gap) as controls (except if the input is the regression dependent variable). [1]Age gap is defined as the absolute value of the husband's age minus the mother's age minus 2.5. [2]Defined from pre marriage data: uses 1940 if available, then 1930, then 1920, then 1910. Never uses a measure that is observed post-MP marriage. Columns 4 and 5 use the alternative measures of occupation score from Olivetti and Paserman (2015). [3]Education gap is defined as the absolute value of the difference in highest grade between the mother and the husband. [4]Equal Weights regressions give the same weight to each of the quality measures. Values are standardized to zero mean and variance equals one. [5]Satisfaction weights include husband's occupational score, education and longevity. We use the utility function and the parameters defined and calibrated in Grow and Van Bavel (2015) to construct the dependent variable. The equation below presents the utility function. The first term of the equation is the similarity of education, the second term is the earnings prospect and, the last term is the age gap. We follow the same categorization of variables as in the original paper, except for education, where we divide it in 4 quintile categories instead of the four categories in the paper (no schooling, primary, secondary and tertiary). $\alpha_i = a_i + 25$ To take into account, that female agents prefer partners who are about 2.5 years older. The parameters are:$S_{max} = 4$; $Y_{max} = 5$; $A_{max} = 800$; $w_s = 0.385$; $w_y = 1.201$; $w_a = 10.833$. $v_{ij} = \left(\frac{S_{max}-|si-sj|}{S_{max}}\right)^{w_s} \left(\frac{y_i}{Y_{max}}\right)^{w_y} \left(\frac{A_{max}-|\alpha_i-\alpha_j|}{A_{max}}\right)^{w_a}$. All indices use the occupation score defined in Column 3.

Table C.7: Does welfare increase quality of Post-MP husband? Results for additional quality measures?

Outcome:	Post-MP husband is foreign	Post-MP Husband's kids at marriage	Post-MP Husband is a farmer*	Post-MP Husband 1940 income	1939 earnings occupation score	Husband's age at marriage	Mom's Education	Mom's age at marriage	Mom and Husband live together	
									1930	1940
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: County and year FE										
Mean of outcome for rejected	0.683	0.607	0.106	751.3	43.17	44.67	7.801	38.24	0.827	0.920
Accepted	0.000	-0.028	0.015	-89.762	-3.008	1.224	-0.110	0.992	-0.042**	-0.009
	(0.027)	(0.066)	(0.018)	(77.141)	(2.044)	(0.912)	(0.148)	(0.622)	(0.019)	(0.023)
Observations	4,266	3,572	4,457	2,815	3,328	2,637	3,759	3,558	2,987	2,612
Panel B: control for predetermined variables										
Accepted	0.000	-0.029	0.014	-65.456	-3.002	1.140*	-0.021	0.995***	-0.041**	-0.012
	(0.027)	(0.069)	(0.017)	(80.028)	(2.102)	(0.627)	(0.141)	(0.346)	(0.019)	(0.024)
Observations	4,266	3,572	4,457	2,815	3,328	2,637	3,759	3,558	2,987	2,612
Panel C: control for pre-determined variables and other inputs										
Mean of outcome for rejected	0.446	0.698	0.142	804.6	41.84	40.92	8.169	36.69	0.788	0.919
Accepted	-0.008	0.032	-0.010	-87.459	0.566	2.452**	-0.108	1.761***	-0.031	-0.028
	(0.049)	(0.098)	(0.027)	(109.735)	(0.804)	(0.944)	(0.172)	(0.548)	(0.031)	(0.031)
Observations	1,887	1,363	1,887	1,755	1,755	998	1,887	1,363	1,431	1,887
Panel D: control for pre-determined variables and mom's age at marriage										
Mean of outcome for rejected	0.671	0.609	0.107	758.6	41.54	44.69	7.902	38.24	0.817	0.912
Accepted	0.003	-0.029	0.003	-100.713	-3.725	0.071	-0.105	0.000	0.017	0.004
	(0.029)	(0.070)	(0.020)	(107.115)	(2.801)	(0.483)	(0.179)	(0.000)	(0.022)	(0.029)
Observations	3,173	3,558	3,269	2,107	2,210	2,635	2,507	3,558	2,248	1,947

Notes: Standard errors are clustered at the county level. Please refer to Table 3.1 for a full description of the controls, restrictions and checks. Panel C includes the other inputs (Post-MP Husband longevity, age gap, Post-MP Husband latest occupational score, Post-MP Husband 1940 education and education gap) as controls. *Defined from pre marriage data: uses 1910 if available, then 1920, then 1930, then 1940. Never uses a measure that is observed post-MP marriage.

Table C.8: Determinants of remarriage and time to remarriage

Dependent variable:	**Remarried=1**			**Duration to remarriage**		
Sample:	All	Accepted	Rejected	All	Accepted	Rejected
Mean of dependent variable	0.482	0.482	0.474	6.357	6.442	5.471
Accepted	-0.014			1.296***		
	(0.02)			(0.44)		
MP age of youngest sibling	-0.004*	-0.004*	-0.007	-0.042	-0.077	0.079
	(0.00)	(0.00)	(0.01)	(0.06)	(0.06)	(0.17)
MP age of oldest sibling	0.002	0.003	-0.006	-0.017	0.009	-0.065
	(0.00)	(0.00)	(0.01)	(0.06)	(0.06)	(0.15)
# of kids in the app	-0.020***	-0.021***	-0.015	0.291**	0.242*	0.526
	(0.01)	(0.01)	(0.02)	(0.14)	(0.15)	(0.47)
Length of mother's last name	0	0.001	-0.011	-0.065	-0.068	-0.086
	(0.00)	(0.00)	(0.01)	(0.07)	(0.07)	(0.23)
Divorced mother (MP)	0.373***	0.381***	0.002	-0.413	-0.257	-2.668
	(0.02)	(0.03)	(0.12)	(0.78)	(0.88)	(2.84)
Widow mother (MP)	0.418***	0.431***	0.032	-0.937	-1	-2.16
	(0.02)	(0.02)	(0.13)	(0.74)	(0.80)	(2.50)
MP Marital status is missing	0.317***	0.328***				
	(0.03)	(0.04)				
Mother's age at application	-0.028***	-0.029***	-0.022***	0.050**	0.061**	-0.053
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.08)
Missing mother's age at application	-0.157*	-0.138*	-0.313**	62.181***	66.564***	-7.072*
	(0.08)	(0.08)	(0.13)	(6.63)	(7.51)	(3.83)
Number of siblings of the mother	0.009***	0.009***	0.006	-0.012	-0.017	0.084
	(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.11)
Mother is foreign born (FS)	-0.018*	-0.021*	-0.01	0.112	0.155	0.098
	(0.01)	(0.01)	(0.05)	(0.38)	(0.41)	(0.86)
Mother's foreign status is missing	-0.055**	-0.051*	-0.097	-1.694	-1.162	2.503
	(0.02)	(0.03)	(0.06)	(1.46)	(2.12)	(3.37)
Pre-MP husband's longevity	0.003***	0.003***	0	0.014	0.013	-0.011
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.03)
Pre-MP husband's longevity missing	0.048***	0.049***	0.046	-0.14	-0.24	-0.534
	(0.01)	(0.01)	(0.03)	(0.18)	(0.19)	(1.12)
Number of kids older than 14 (FS)	0.015***	0.015***	0.009	-0.165**	-0.186**	0.079
	(0.00)	(0.00)	(0.01)	(0.07)	(0.08)	(0.32)
No. of kids that died before app (FS)	0.003	0.003	0.02	-0.052	0.018	-1.090**
	(0.01)	(0.01)	(0.02)	(0.14)	(0.14)	(0.50)
No. with missing dates of birth/death (FS)	-0.005	0.006	-0.117***	0.544*	0.387	2.445**
	(0.01)	(0.01)	(0.04)	(0.29)	(0.31)	(1.09)
Observations	11286	10237	1049	3572	3259	313

Notes: OLS regressions. S.E. clustered at the county level. Specifications also include year of app FE. State & county covariates not shown.

Table C.9: Do the cash transfers affect Fertility?

Sample:	All mothers				Mothers that were not married at time of			
Data source	Family Search		Census		Family Search		Census	
Outcome	Post MP kids born	Children ever born	Number of own children in household		Post MP kids born	Children ever born	Number of own children in household	
			1930	1940			1930	1940
Mean of Y for rejected	0.25	4.13	2.38	1.54	0.22	4.16	2.39	1.57
Panel A: County and year FE								
Accepted	0.014	0.326	0.218	0.112	0.024	0.337	0.195	0.099
Robust standard errors	(0.022)	(0.078)***	(0.061)***	(0.059)*	(0.023)	(0.086)***	(0.067)***	(0.065)
Clustered at county	[0.022]	[0.064]***	[0.068]***	[0.056]**	[0.025]	[0.081]***	[0.071]***	[0.059]*
Clustered at county*year	{0.022}	{0.071}***	{0.063}***	{0.066}*	{0.023}	{0.081}***	{0.068}***	{0.074}
R-squared	0.037	0.055	0.131	0.126	0.043	0.059	0.136	0.125
Panel B: Main results (Full controls)								
Accepted	-0.023	0.037	-0.069	-0.036	-0.009	0.061	-0.067	-0.056
Robust standard errors	(0.021)	(0.035)	(0.051)	(0.055)	(0.022)	(0.037)*	(0.056)	(0.060)
Clustered at county	[0.018]	[0.032]	[0.051]	[0.049]	[0.021]	[0.034]*	[0.059]	[0.051]
Clustered at county*year	{0.020}	{0.033}	{0.052}	{0.060}	{0.021}	{0.036}*	{0.056}	{0.067}
R-squared	0.160	0.799	0.412	0.279	0.162	0.805	0.407	0.274
Observations	16228	16228	11178	9358	13383	13383	9174	7635

Notes: Please refer to Table 3.1 for a full description of the controls, restrictions and checks.

Table C.10: Heterogeneity in results - controls

Sample	All		Among remarried women only							
Outcome (Y):	Ever remarried?	# kids post MP	Years to remarriage	Equal weights	Utility weighted index	Post-MP Husband Longevity	Post-MP Husband Occ Score	Post-MP Husband Education	Age gap (shifted by 2.5 years)	Education gap
A.All moms										
Accepted	-0.013	-0.023	1.296	0.945	-0.005	1.752	-0.327	-0.115	0.390	-0.039
Clustered at county level	(0.018)	(0.018)	(0.398)***	(0.429)**	(0.018)	(0.899)*	(0.461)	(0.219)	(0.235)*	(0.174)
R-squared	0.212	0.160	0.315	0.058	0.076	0.052	0.089	0.119	0.044	0.068
Mean of outcome for rejected	0.468	0.246	5.719	23.123	0.360	70.242	21.095	7.709	6.557	1.950
Observations	13638	16228	4255	5792	2973	4830	4206	3460	5771	2978
B. all unmarried moms										
Accepted	-0.014	-0.009	1.275	0.921	-0.006	1.821	-0.828	-0.226	0.275	-0.064
Clustered at county level	(0.020)	(0.021)	(0.444)***	(0.448)**	(0.021)	(0.903)**	(0.574)	(0.228)	(0.289)	(0.185)
R-squared	0.228	0.162	0.338	0.068	0.085	0.056	0.095	0.122	0.049	0.081
Mean of outcome for rejected	0.474	0.224	5.471	23.153	0.361	70.129	21.220	7.798	6.661	1.821
Observations	11286	13383	3572	4894	2540	4104	3556	2955	4874	2545
C. unmarried moms and drop if marital status missing at application										
Accepted	-0.048	-0.032	1.112	1.366	-0.018	2.577	-0.542	-0.393	0.323	-0.258
Clustered at county level	(0.020)**	(0.030)	(0.573)*	(0.467)***	(0.023)	(1.011)**	(0.677)	(0.297)	(0.313)	(0.199)
R-squared	0.241	0.174	0.331	0.076	0.094	0.066	0.095	0.117	0.053	0.108
Mean of outcome for rejected	0.529	0.252	5.232	22.960	0.368	69.428	21.162	7.915	6.559	1.885
Observations	7925	9171	2620	3524	1794	2965	2549	2094	3511	1797
D. states that only admit widows										
Accepted	-0.006	0.000	1.467	1.571	0.003	1.587	0.272	-0.457	0.529	-0.212
Clustered at county level	(0.026)	(0.030)	(0.696)**	(0.494)***	(0.023)	(1.908)	(0.837)	(0.330)	(0.235)**	(0.196)
R-squared	0.206	0.156	0.319	0.060	0.094	0.060	0.101	0.106	0.050	0.101
Mean of outcome for rejected	0.491	0.222	5.471	23.028	0.371	69.582	20.867	8.193	6.335	2.229
Observations	4128	4906	1395	1795	920	1507	1199	1053	1790	921
E. states that admit more than just widows										
Accepted	-0.013	-0.025	1.096	0.644	-0.010	1.547	-0.612	0.056	0.333	0.037
Clustered at county level	(0.020)	(0.022)	(0.430)**	(0.529)	(0.025)	(0.892)*	(0.606)	(0.210)	(0.314)	(0.236)
R-squared	0.220	0.165	0.333	0.070	0.089	0.063	0.097	0.147	0.054	0.076
Mean of outcome for rejected	0.458	0.257	5.858	23.170	0.355	70.555	21.180	7.515	6.665	1.833
Observations	9510	11322	2860	3997	2053	3323	3007	2407	3981	2057
P-value of test that D=E	0.869	0.804	0.767	0.191	0.642	0.741	0.295	0.253	0.758	0.616
F. states that regulated/required work										
Accepted	-0.006	-0.012	1.244	0.852	-0.002	1.641	-0.347	0.082	0.488	-0.095
Clustered at county level	(0.023)	(0.024)	(0.468)**	(0.582)	(0.028)	(1.021)	(0.715)	(0.253)	(0.358)	(0.272)
R-squared	0.212	0.149	0.316	0.068	0.092	0.060	0.086	0.144	0.049	0.086
Mean of outcome for rejected	0.444	0.229	6.130	23.011	0.364	70.188	21.763	7.530	6.421	1.918
Observations	6657	8015	2046	2804	1432	2351	2144	1691	2795	1435
G. states that required women to stay										
Accepted	-0.017	-0.030	1.265	1.127	-0.010	1.559	-0.083	-0.245	0.333	-0.034
Clustered at county level	(0.025)	(0.029)	(0.604)**	(0.556)**	(0.019)	(1.492)	(0.599)	(0.327)	(0.296)	(0.214)
R-squared	0.219	0.180	0.343	0.068	0.098	0.070	0.112	0.124	0.058	0.095
Mean of outcome for rejected	0.501	0.270	5.270	23.252	0.355	70.304	20.231	7.951	6.714	1.990
Observations	6981	8213	2209	2988	1541	2479	2062	1769	2976	1543
P-value of test that F=G	0.651	0.511	0.997	0.526	0.919	0.916	0.722	0.445	0.883	0.700

Table C.10: Heterogeneity in results - controls (Continued)

	Ever	# kids post MP	Years to	Equal weights ^{4}	Utility	Post-MP	Post-MP	Post-MP	Age gap	Education gap^3
Outcome (Y):	remarried?		remarriage		weighted	Husband	Husband	Husband	(shifted by 2.5	
					quality index	Longevity	Occ Score^2	Education	years)^1	
H. counties with high share males (sex ratio above median)										
Accepted	0.001	-0.002	1.642	1.183	0.014	2.154	-0.347	-0.292	0.259	0.012
Clustered at county level	(0.020)	(0.022)	(0.475)***	(0.477)**	(0.026)	(1.224)*	(0.642)	(0.320)	(0.276)	(0.226)
R-squared	0.225	0.181	0.282	0.075	0.100	0.063	0.109	0.147	0.059	0.089
Mean of outcome for rejected	0.472	0.228	5.751	23.154	0.351	69.876	21.106	7.965	6.549	2.083
Observations	6778	8095	2228	2995	1544	2511	2089	1787	2983	1547
I. counties with low share males (sex ratio below median)										
Accepted	-0.022	-0.050	1.042	0.505	-0.021	1.051	-0.242	0.034	0.543	-0.253
Clustered at county level	(0.027)	(0.029)*	(0.544)*	(0.598)	(0.027)	(1.016)	(0.747)	(0.252)	(0.382)	(0.238)
R-squared	0.208	0.144	0.377	0.068	0.115	0.073	0.100	0.126	0.065	0.084
Mean of outcome for rejected	0.464	0.268	5.679	23.086	0.369	70.675	21.085	7.459	6.568	1.814
Observations	6860	8133	2027	2797	1429	2319	2117	1673	2788	1431
P-value of test that H=I	0.746	0.104	0.636	0.415	0.296	0.691	0.961	0.411	0.786	0.443
J. counties with high female labor force participation (LFP above median)										
Accepted	-0.020	-0.021	0.994	0.678	0.019	2.338	0.201	-0.401	0.346	-0.127
Clustered at county level	(0.028)	(0.020)	(0.873)	(0.694)	(0.022)	(1.384)*	(0.721)	(0.265)	(0.321)	(0.228)
R-squared	0.203	0.130	0.359	0.060	0.099	0.064	0.105	0.142	0.055	0.115
Mean of outcome for rejected	0.442	0.200	6.076	23.224	0.357	69.557	21.795	8.105	6.583	2.000
Observations	6766	8108	1980	2741	1426	2245	1940	1633	2733	1427
K. counties with low female labor force participation (LFP below median)										
Accepted	-0.005	-0.026	1.468	1.305	-0.021	1.373	-0.466	0.141	0.429	0.101
Clustered at county level	(0.015)	(0.030)	(0.435)***	(0.618)**	(0.026)	(1.210)	(0.715)	(0.282)	(0.333)	(0.236)
R-squared	0.232	0.188	0.288	0.094	0.113	0.084	0.114	0.154	0.065	0.122
Mean of outcome for rejected	0.493	0.290	5.413	23.037	0.362	70.839	20.544	7.372	6.535	1.908
Observations	6872	8120	2275	3051	1547	2585	2266	1827	3038	1551
P-value of test that J=K	0.601	0.741	0.573	0.585	0.231	0.634	0.593	0.186	0.453	0.763
L. moms above median age										
Accepted	-0.010	-0.007	2.470	0.876	-0.010	3.022	-2.963	-0.237	1.166	-0.292
Clustered at county level	(0.021)	(0.010)	(0.659)***	(0.919)	(0.043)	(2.199)	(1.567)*	(0.395)	(0.656)*	(0.318)
R-squared	0.096	0.060	0.181	0.143	0.274	0.130	0.189	0.243	0.137	0.298
Mean of outcome for rejected	0.282	0.032	5.083	24.781	0.387	70.583	23.587	7.354	6.146	2.200
Observations	6407	7214	1091	1594	606	1267	1113	753	1590	607
M. moms below median age										
Accepted	-0.016	-0.032	0.924	1.160	0.001	1.346	0.803	-0.132	0.101	0.075
Clustered at county level	(0.021)	(0.036)	(0.531)*	(0.567)**	(0.022)	(0.916)	(0.742)	(0.243)	(0.381)	(0.195)
R-squared	0.150	0.157	0.370	0.059	0.096	0.064	0.107	0.124	0.057	0.078
Mean of outcome for rejected	0.644	0.420	5.932	22.491	0.353	70.113	20.122	7.812	6.714	1.883
Observations	7231	9014	3164	4198	2367	3563	3093	2707	4181	2371
P-value of test that L=M	0.699	0.559	0.187	0.521	0.164	0.413	0.041	0.773	0.163	0.208
N. moms above median age of youngest										
Accepted	0.002	-0.002	0.684	1.637	0.001	2.590	-0.832	-0.213	0.795	0.200
Clustered at county level	(0.020)	(0.013)	(0.568)	(0.675)**	(0.027)	(1.222)**	(0.932)	(0.367)	(0.470)*	(0.309)
R-squared	0.177	0.096	0.347	0.116	0.222	0.117	0.178	0.212	0.126	0.232
Mean of outcome for rejected	0.348	0.078	5.857	23.737	0.381	70.597	22.240	7.630	6.425	1.800
Observations	5672	6886	1296	1804	802	1470	1283	965	1797	803
O. moms below median age if youngest										
Accepted	-0.029	-0.037	1.377	0.659	-0.003	1.557	0.033	-0.139	0.151	-0.152
Clustered at county level	(0.023)	(0.033)	(0.512)***	(0.629)	(0.023)	(1.076)	(0.754)	(0.285)	(0.328)	(0.219)
R-squared	0.213	0.169	0.344	0.073	0.104	0.072	0.104	0.135	0.055	0.078
Mean of outcome for rejected	0.594	0.422	5.638	22.781	0.351	70.055	20.551	7.746	6.631	2.012
Observations	7966	9342	2959	3988	2171	3360	2923	2495	3974	2175
P-value of test that N=O	0.279	0.744	0.438	0.688	0.548	0.748	0.998	0.998	0.307	0.402

Table C.10: Heterogeneity in results - controls (Continued)

	Ever	# kids post MP	Years to	Equal weights [4]	Utility	Post-MP	Post-MP	Post-MP	Age gap	Education gap[3]
Outcome (Y):	remarried?		remarriage		weighted	Husband	Husband	Husband	(shifted by 2.5	
					quality index	Longevity	Occ Score[2]	Education	years)[1]	
P. county pop receiving aid above median										
Accepted	-0.016	-0.011	1.735	0.831	0.020	1.373	-0.155	0.031	0.158	0.077
Clustered at county level	(0.024)	(0.022)	(0.475)***	(0.585)	(0.021)	(0.903)	(0.728)	(0.252)	(0.339)	(0.260)
R-squared	0.220	0.172	0.195	0.064	0.077	0.050	0.097	0.122	0.045	0.059
Mean of outcome for rejected	0.456	0.235	5.797	23.020	0.358	69.629	21.596	7.556	6.484	1.786
Observations	6785	8114	2292	3009	1534	2512	2202	1823	3001	1537
Q. county pop receiving aid below median										
Accepted	-0.005	-0.031	1.034	1.420	-0.028	1.665	-0.789	-0.288	0.884	-0.139
Clustered at county level	(0.027)	(0.028)	(0.802)	(0.535)***	(0.023)	(1.622)	(0.564)	(0.280)	(0.275)***	(0.239)
R-squared	0.209	0.154	0.432	0.080	0.114	0.081	0.101	0.153	0.066	0.100
Mean of outcome for rejected	0.481	0.258	5.649	23.225	0.362	70.819	20.546	7.857	6.629	2.107
Observations	6853	8114	1963	2783	1439	2318	2004	1637	2770	1441
P-value of test that P=Q	0.979	0.283	0.479	0.452	0.121	0.806	0.438	0.496	0.213	0.663
R. age of widowhood above median										
Accepted	-0.008	-0.022	1.434	1.348	0.024	1.601	0.117	-0.130	0.419	-0.042
Clustered at county level	(0.020)	(0.016)	(0.460)***	(0.641)**	(0.021)	(1.369)	(0.646)	(0.227)	(0.252)*	(0.230)
R-squared	0.232	0.178	0.435	0.096	0.120	0.078	0.114	0.156	0.067	0.108
Mean of outcome for rejected	0.420	0.203	5.967	22.544	0.350	69.761	21.118	7.606	6.298	2.075
Observations	9142	11360	2403	3383	1609	2716	2417	1890	3364	1613
S. age of widowhood below median										
Accepted	-0.038	-0.032	0.991	-0.069	-0.040	2.133	-0.734	-0.119	0.219	0.062
Clustered at county level	(0.028)	(0.065)	(0.507)*	(0.896)	(0.027)	(1.525)	(1.226)	(0.389)	(0.489)	(0.227)
R-squared	0.174	0.165	0.122	0.080	0.141	0.106	0.142	0.175	0.091	0.154
Mean of outcome for rejected	0.598	0.385	5.284	24.104	0.373	70.941	21.060	7.847	6.992	1.788
Observations	4496	4868	1852	2409	1364	2114	1789	1570	2407	1365
P-value of test that R=S	0.320	0.848	0.571	0.244	0.038	0.819	0.573	0.962	0.648	0.191

Notes: Please refer to Table 3.1 for a full description of the controls, restrictions and checks. [1] Age gap is defined as the absolute value of the husband's age minus the mother's age minus 2.5. [2] Defined from pre marriage data: uses 1940 if available, then 1930, then 1920, then 1910. Never uses a measure that is observed post-MP marriage. Columns 4 and 5 use the alternative measures of occupation score from Olivetti and Paserman (2015). [3] Education gap is defined as the absolute value of the difference in highest grade between the mother and the husband. [4] Equal Weights regressions give the same weight to each of the quality measures. Values are standardized to zero mean and variance equals one.

Table C.11: Do the cash transfers affect Labor supply and wages?

Outcome:	Labor force participation			Work		Occupation Score | occupation not missing			Earned Income | income > 0
Sample:	Applied in 1918-1920	Applied in 1928-1930	All	Applied in 1928-1930	All	Applied in 1918-1920	Applied in 1928-1930	All	All
Census Year	1920	1930	1940	1930	1940	1920	1930	1940	1940
Mean of Y for rejected	0.41	0.38	0.21	0.36	0.19	4.47	15.80	15.79	479.08
Panel A: No controls									
Accepted	-0.026	0.050	0.050	0.022	0.054	-0.243	-1.261	0.154	38.030
Robust standard errors	(0.051)	(0.037)	(0.015)***	(0.037)	(0.015)***	(0.759)	(1.047)	(0.729)	(32.968)
Clustered at county	[0.045]	[0.032]	[0.020]**	[0.029]	[0.018]***	[0.539]	[1.114]	[0.597]	[37.463]
Clustered at county*year	{0.043}	{0.033}	{0.016}***	{0.036}	{0.016}***	{0.647}	{1.141}	{0.745}	{30.602}
Bounds for missing data (Lee 2009)	[-0.31;0.15]	[-0.07;0.14]	[-0.12;0.11]	[-0.11;0.10]	[-0.12;0.11]	[-4.47;1.70]	[-4.92;1.64]	[-5.27;5.26]	[-203.18;247.58]
R-squared	0.000	0.001	0.001	0.000	0.001	0.000	0.002	0.000	0.000
Panel B: Full controls									
Accepted	-0.038	0.001	0.027	-0.032	0.032	-0.149	-1.915	-0.349	5.434
Robust standard errors	(0.055)	(0.043)	(0.017)	(0.043)	(0.016)**	(0.861)	(1.297)	(0.795)	(37.659)
Clustered at county	[0.048]	[0.042]	[0.017]	[0.041]	[0.016]**	[0.584]	[0.909]**	[0.614]	[28.688]
Clustered at county*year	{0.046}	{0.037}	{0.017}	{0.041}	{0.016}*	{0.716}	{1.360}	{0.822}	{35.857}
R-squared	0.154	0.132	0.067	0.113	0.061	0.128	0.163	0.108	0.160
Observations	1451	2225	9351	2227	9358	1452	799	2737	2083
Panel C: Checks									
1- Correction for OVB (Oster 2017)	[-0.04;-0.03]	[-0.02;0.02]	[0.02;0.04]	[-0.06;-0.01]	[0.02;0.04]	[-0.18;-0.11]	[-2.30;-1.65]	[-0.55;-0.18]	[-8.04;16.94]
2- Semi-parametric sample selection correction (Newey, 2009)									
Accepted	-0.039	0.001	0.025	-0.031	0.030	-0.168	-1.628	-0.358	7.543
95% Confidence interval	[-0.14;0.06]	[-0.08;0.09]	[-0.01;0.06]	[-0.11;0.05]	[-0.00;0.06]	[-1.38;1.05]	[-3.37;0.11]	[-1.58;0.86]	[-49.45;64.53]
F-Stat	11.31	15.97	116.23	16.13	116.82	11.29	84.57	52.17	74.65
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3- Drop if quality of match low									
Accepted	-0.034	0.038	0.011	0.002	0.020	-0.005	-2.424	0.395	28.263
Clustered at county	(0.083)	(0.058)	(0.021)	(0.054)	(0.022)	(1.150)	(1.575)	(1.256)	(38.861)
Observations	743	1459	4679	1459	4679	743	561	1547	1237
4 - IPW	-0.026	0.050	0.050**	0.022	0.054***	-0.243	-1.261	0.154	38.030
	(0.045)	(0.032)	(0.020)	(0.029)	(0.018)	(0.536)	(1.107)	(0.594)	(37.301)
5 - Causal Forest ATE	-0.033	-0.013	0.034**	-0.039	0.037**	-0.548	-2.246*	-0.609	7.040
	(0.051)	(0.039)	(0.015)	(0.039)	(0.015)	(0.825)	(1.322)	(0.918)	(35.433)
6- Causal Forest ATT	-0.034	-0.026	0.038*	-0.053	0.038*	-0.608	-2.261*	-0.703	3.044
	(0.060)	(0.053)	(0.021)	(0.055)	(0.020)	(0.965)	(1.294)	(1.131)	(42.037)
Observations	1451	2225	9351	2227	9358	1452	799	2737	2083

Notes: Please refer to Table 3.1 for a full description of the controls, restrictions and checks. Refer to Table 3.2 for a description of the quality measure. Note from IPUMS: Census practice on collecting occupational data (in OCC) for persons not currently in the labor force changed over time. In the earliest samples, no time referent was specified for when the person was gainfully employed. In 1900, past occupation was specifically requested for persons unable to secure any work during the preceding year, but not for persons who had permanently retired. Similarly, for the 1910-1930 surveys, occupation was to be reported for persons temporarily unemployed, but not for those permanently retired. This changed markedly in 1940 and 1950. In those years, OCC was reserved for those in the labor force (working, with a job, or looking for work) in the week prior to the census. For 1940 and 1950, past occupation was separately collected via different questions and variables (UOCC and ROCC) for formerly-employed persons not currently in the labor force.

C.2 Model and Proofs

A woman is either single or married. While single, she receives a flow benefit of b and, with probability λ, she finds a potential partner with quality q (the flow utility she would get from marriage) and decides whether to marry him or stay single. For simplicity, we say that marriage lasts forever. The quality of a partner q is distributed $F(q)$ with support $\left[\underline{q}, \bar{q}\right]$ and $\bar{q} > b$. She discounts the future at rate β.

The value of being single is

$$V_s = b + \beta\left(\lambda \int_{q=\underline{q}}^{\bar{q}} \max\left\{V_m(q), V_s\right\}\ \mathrm{d}F(q) + (1-\lambda)V_s\right).$$

The value of being married to a partner with quality q is

$$V_m(q) = q + \beta V_m(q) = \frac{q}{1-\beta}.$$

Since the value of marriage is strictly increasing in q, the agent will follow a cut-off rule. There is a cutoff quality, q^*, such that she will accept all prospects with $q > q^*$. The cutoff rule is implicitly defined as

$$V_m(q^*) = V_s.$$

Considering that, and rearranging the definition of V_s, we can write

$$\begin{aligned}
V_s &= b + \beta V_s + \beta\lambda \int_{q=\underline{q}}^{\bar{q}} \big(\max\left\{V_m(q) - V_s, 0\right\}\big)\ \mathrm{d}F(q),\\
V_s &= b + \beta V_s + \beta\lambda \int_{q=q^*}^{\bar{q}} \big(V_m(q) - V_s\big)\ \mathrm{d}F(q),\\
V_s &= b + \beta V_s + \frac{\beta\lambda}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\ \mathrm{d}q,\\
(1-\beta)V_s &= b + \frac{\beta\lambda}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\ \mathrm{d}q,
\end{aligned}$$

where the third line followed from integration by parts. From the definition of q^*, we have obtained an implicit equation for q^* (which contains no other endogenous variables)

$$q^* = b + \frac{\beta\lambda}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big) \, \mathrm{d}q. \tag{C.1}$$

$$0 = -q^* + b + \frac{\beta\lambda}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big) \, \mathrm{d}q.$$

We can see that this function is continuous and positive at $q^* = b$ and negative at $q^* = \bar{q}$, so there exists a solution. Also, the function is strictly decreasing so its solution is unique.

Intuitively, this equation says that the value of the minimum acceptable marriage, q^* should be equal to the benefit, b, plus the option value of holding out for a good match. Given a reservation quality, q^*, the probability of marriage is $\lambda\,(1 - F(q^*))$ and the average match quality is $\mathbb{E}[q|q > q^*]$. The duration until remarriage is given by $D = 1/\lambda\,(1 - F(q^*))$.

Before proving Proposition 1, we establish the following useful result.

Lemma 1. *The reservation quality, q^*, is increasing in benefits, b. Moreover, the reservation quality is also increasing in the probability of finding prospects, λ, and the distribution of quality $F(q)$ (in the senses of first-order stochastic dominance).*

Proof. This result can be seen on equation 1. An increase in b, λ, or the distribution F increases the right-hand side of the equation which corresponds to the value of waiting. In order to preserve the equality, the cutoff must be higher. Waiting is more attractive when the benefits are higher, the offers appear more often, or the offers are stochastically better. Then, the woman will only find it worthwhile to settle for a higher cutoff quality. □

Now, we are ready to prove Proposition 1.

Proposition 3. $\partial D/\partial b > 0 and\ \partial\mathbb{E}[q|q > q^*]/\partial b > 0$*: An increase in benefits, b, increases the number of periods the woman stays single and the average quality of the marriage.*

Proof. From our previous lemma, an increase in benefits will increase the cutoff quality. Since the probability of marriage is decreasing in the cutoff quality, the increase in benefits decreases the probability of marriage and increases the expected number of periods the woman stays single. The average quality of the marriage increases because the woman now rejects relatively lower quality proposals. □

In order to test the second prediction of Proposition 1, we would need to observe the quality of the marriage. what we observe are several traits that matter for the quality. We assume that there exists a quality function, $q : \mathcal{X} \to [\underline{q}, \bar{q}]$, that maps a vector of characteristics into a single quality index. For exposition, and without loss of generality, we assume that the function q is increasing in each trait.

Proposition 4. *Without further assumptions about the joint distribution of X and the production function $q(X)$, the sign of $\partial \mathbb{E}[x_i|q > q^*]/\partial b$ is ambiguous for all i. However the sign of $\partial \mathbb{E}[x_i|q > q^*, x_{-i}]/\partial b$ is positive for all x_i so long as all relevant X are observed.*

It might seem natural to expect that higher benefits would result in higher (better) traits in the accepted marriages. This is not necessarily true and it could be that every trait becomes worse.

Example 1. Consider a quality function $q(x_1, x_2) = x_1 x_2$. The joint distribution of the traits is uniformly distributed over three mass points $(1, 10); (10, 1); (4, 4)$. Suppose that, initially, the cutoff is $q^* = 10$. The average of each trait conditional on a match is equal to 5. Consider a small increase in the cutoff $(10 < q^* \leq 16)$. The new average of each trait is 4.

As the example shows, each trait could be, on average, lower with a higher cutoff quality. Still, we can predict an increase in a particular trait when conditioning for all the other relevant traits. In order to see this, notice that for a given value of the other traits, a higher cutoff will only eliminate matches where the trait we are interested in was low.

Extensions

Age

We show that the predictions of the model still hold when we incorporate aging considerations. In order to maintain the simple recursive structure of the model, we model aging as a random independent process that moves the agent from a young state to an old state. In the young state, a woman receives a proposal with probability λ_Y. In the old state, she receives a proposal with probability $\lambda_O < \lambda_Y$. There is a probability π of transitioning from young to old and, naturally, no probability of the reverse transition. The transition, or lack of, is realized at the end of each period after the offer has been accepted or rejected.

The old single woman's problem is the same as the original problem. Let us define $V_{s,O}$ and q_O^* as the value of being single and the cutoff quality when old.

The young woman's problem is slightly different. The opportunity cost of accepting a proposal is given by $V := (1-\pi)V_{s,Y} + \pi V_{s,O}$, where $V_{s,Y}$ is the value of being single when young.

$$V_{s,Y} = b + \beta \left(\lambda_Y \int_{q=\underline{q}}^{\bar{q}} \max\left\{ V_m(q), V \right\} \, \mathrm{d}F(q) + (1-\lambda_Y)V \right).$$

The cutoff rule is defined by $V_m(q_Y^*) = q_Y^*/(1-\beta) = V$. Then, $\pi(V_{s,Y} - V_{s,O}) = \frac{\pi}{1-\pi}(V - V_{s,O}) = \frac{\pi}{1-\pi}\frac{q_Y^* - q_O^*}{1-\beta}$.

$$\begin{aligned}
V_{s,Y} &= b + \beta V + \beta\lambda_Y \int_{q=\underline{q}}^{\bar{q}} \big(\max\left\{ V_m(q) - V, 0 \right\} \big) \, \mathrm{d}F(q), \\
V_{s,Y} &= b + \beta V + \beta\lambda_Y \int_{q=q_Y^*}^{\bar{q}} \big(V_m(q) - V \big) \, \mathrm{d}F(q), \\
V_{s,Y} &= b + \beta V + \frac{\beta\lambda_Y}{1-\beta} \int_{q=q_Y^*}^{\bar{q}} \big(1 - F(q) \big) \, \mathrm{d}q, \\
(1-\beta)V &= b + \pi(V_{s,O} - V_{s,Y}) + \frac{\beta\lambda_Y}{1-\beta} \int_{q=q_Y^*}^{\bar{q}} \big(1 - F(q) \big) \, \mathrm{d}q, \\
q_Y^* &= b - \frac{\pi}{1-\pi}\frac{q_Y^* - q_O^*}{1-\beta} + \frac{\beta\lambda_Y}{1-\beta} \int_{q=q_Y^*}^{\bar{q}} \big(1 - F(q) \big) \, \mathrm{d}q,
\end{aligned}$$

This equation takes into account the probability of transitioning into old age. It is easy to see that the cutoff quality will not be the same if $\lambda_Y > \lambda_O$.

Proposition 5. *If the arrival rate λ falls with age then $\partial \mathbb{E}[q|q > q^*]/\partial b > 0$ and $\partial D/\partial b > 0$.*

Proof. First, for the old woman, the analysis of the basic model applies and the result follows immediately. Second, for the young woman, we can apply the same kind of analysis. Higher benefits increase the value of waiting both directly and indirectly. The direct effect comes from enjoying the benefits while single and young and the indirect effect comes from the benefits while old (which shows up through the cutoff quality of old). Thus, all cutoff qualities increase which implies higher expected qualities conditional on a match and a higher duration of single-hood. □

Stigma

Getting the benefits could also bring about negative effects if there is stigma associated with participating in the program. In the model, we can think of this issue in two ways. First, being in the program lowers the probability of receiving an offer. Second, the distribution of offers gets worse.

In either case, the presence of the stigma makes the predictions of the model ambiguous.

Proposition 6. *If b lowers the rate of arrival of prospects λ or worsens the distribution $F(q)$ in the sense of first-order stochastic dominance (in addition to increasing the per period utility) then the sign of $\partial \mathbb{E}[q|q > q^*]/\partial b$ and $\partial D/\partial b$ becomes ambiguous.*

Proof. Lemma 1 established that the cutoff quality moved in the same direction as the benefits, the change in the probability of proposals, λ, and the distribution, $F(q)$. With a stigma effect, the program increases b but lowers λ or F. The original effect increases the cutoff but the stigma effect lowers it. It is unclear which one we should expect to dominate.
□

Work

The initial predictions are maintained when we introduce a labor decision in the model. In this extension, a woman has a probability λ_E of receiving an employment opportunity. A job offer is characterized by its wage w which is distributed $G(w)$ with support $[\underline{w}.\bar{w}]$ and $\bar{w} > b$. We assume that marriage lasts forever and that an employed woman loses her job with probability δ each period. We also assume that an employed woman can receive marriage offers at rate $\lambda_{m,e}$ and with quality distributed $\hat{F}(q)$.

In this extension, there exist three possible states: single and unemployed, single and employed, and married. The value of being single and unemployed is

$$V_{s,u} = b + \beta\lambda_m \int_q \max\{V_{m,u}(q), V_{s,u}\} dF(q) + \beta\lambda_e \int_w \max\{V_{s,e}(w), V_{s,u}\} dG(w) + \beta(1 - \lambda_m - \lambda_e)V_{s,u}.$$

The value of being married to a partner with quality q is

$$V_{m,u}(q) = q + \beta V_{m,u}(q) = q/(1-\beta).$$

The value of being employed at wage w is

$$V_{s,e}(w) = w + \beta\lambda_{m,e} \int_{q=\underline{q}}^{\bar{q}} \max\{V_{s,e}(w), V_{m,u}(q)\} \mathrm{d}\hat{F}(q) + \beta\delta V_{s,u} + \beta(1 - \lambda_{m,e} - \delta)V_{s,e}(w).$$

Letw^* be the cutoff wage and q^* be the cutoff quality for the single, unemployed woman. Then, by definition of cutoff wage and quality

$$(1-\beta)V_{s,u} = (1-\beta)V_{s,e}(w^*) = (1-\beta)V_{m,u}(q^*) = q^*.$$

Evaluating the expression above at w^*, we get

$$q^* = w^* + \frac{\beta\lambda_{m,e}}{1-\beta} \int_{q=q^*}^{\bar{q}} [q - q^*] \mathrm{d}\hat{F}(q) = w^* + \frac{\beta\lambda_{m,e}}{1-\beta} \int_{q=q^*}^{\bar{q}} [1 - \hat{F}(q)] \mathrm{d}q. \tag{C.2}$$

For each wage w, there will be a cutoff marriage quality, $q(w)$, such that all proposals with quality $q > q(w)$ will be taken. The cutoff marriage quality is implicitly defined by

$$V_{s,e}(w) = V_{m,u}(q(w)) = \frac{q(w)}{1-\beta}.$$

Then, we can write,

$$[1-\beta(1-\delta)]V_{s,e}(w) = w + \beta\lambda_{m,e}\int_{q=\underline{q}}^{\bar{q}} \max\{0, V_{m,u}(q) - V_{s,e}(w)\}\mathrm{d}\hat{F}(q) + \beta\delta V_{s,u}.$$

$$[1-\beta(1-\delta)]V_{s,e}(w) = w + \frac{\beta\lambda_{m,e}}{1-\beta}\int_{q=q(w)}^{\bar{q}} [1-\hat{F}(q)]\mathrm{d}q + \beta\delta V_{s,u}.$$

$$[1-\beta(1-\delta)][V_{s,e}(w) - V_{s,u}] = w + \frac{\beta\lambda_{m,e}}{1-\beta}\int_{q=q(w)}^{\bar{q}} [1-\hat{F}(q)]\mathrm{d}q - (1-\beta)V_{s,u}.$$

$$[1-\beta(1-\delta)][V_{s,e}(w) - V_{s,u}] = w + \frac{\beta\lambda_{m,e}}{1-\beta}\int_{q=q(w)}^{\bar{q}} [1-\hat{F}(q)]\mathrm{d}q - q^*.$$

$$[1-\beta(1-\delta)][V_{s,e}(w) - V_{s,u}] = w - q^* + \frac{\beta\lambda_{m,e}}{1-\beta}\int_{q=q(w)}^{\bar{q}} [1-\hat{F}(q)]\mathrm{d}q.$$

$$q(w) = q^* + \frac{1-\beta}{1-\beta(1-\delta)}(w - q\tilde{n}^*) + \frac{\beta\lambda_{m,e}}{1-\beta(1-\delta)}\int_{q=q(w)}^{\bar{q}} [1-\hat{F}(q)]\mathrm{d}q. \tag{C.3}$$

We can directly establish the existence and uniqueness of the solution of $q(w^*)$ (the cutoff marriage quality at the reservation wage) by evaluating this expression at $w = w^*$. The cutoff marriage quality accounts for the current wage, the search value, and the possibility of the job being lost.

Now, the value of being single and unemployed is given as before.

$$(1-\beta)V_{s,u} = b + \beta\lambda_M \int_{q=\underline{q}}^{\bar{q}} \big(\max\{V_m(q) - V_{s,u}, 0\}\big)\,\mathrm{d}F(q) + \beta\lambda_E \int_{w=\underline{w}}^{\bar{w}} \big(\max\{V_{s,e}(w) - V_{s,u}, 0\}\big)\,\mathrm{d}G(w),$$
$$(1-\beta)V_{s,u} = b + \beta\lambda_M \int_{q=q^*}^{\bar{q}} \big(V_m(q) - V_{s,u}\big)\,\mathrm{d}F(q) + \beta\lambda_E \int_{w=w^*}^{\bar{w}} \big(V_{s,e}(w) - V_{s,u}\big)\,\mathrm{d}G(w),$$
$$(1-\beta)V_{s,u} = b + \frac{\beta\lambda_M}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\,\mathrm{d}q + \frac{\beta\lambda_E}{1-\beta(1-\delta)} \int_{w=w^*}^{\bar{w}} \big(1 - G(w)\big)\,\mathrm{d}q(w),$$
$$(1-\beta)V_{s,u} = b + \frac{\beta\lambda_M}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\,\mathrm{d}q + \frac{\beta\lambda_E}{1-\beta(1-\delta)} \int_{w=w^*}^{\bar{w}} \big(1 - G(w)\big)\,\mathrm{d}q(w),$$

$$q^* = b + \frac{\beta\lambda_M}{1-\beta} \int_{q=q^*}^{\bar{q}} \big(1 - F(q)\big)\,\mathrm{d}q + \frac{\beta\lambda_E}{1-\beta(1-\delta)} \int_{w=w^*}^{\bar{w}} \big(1 - G(w)\big)\,\mathrm{d}q(w). \tag{C.4}$$

Then, we can solve for all cutoffs in the following way. We first solve for the cutoffs at the single, unemployed state. Those cutoffs are w^* and q^*. Equation (C.2) is increasing in w^* while equation (C.4) is decreasing in w^*. This means that if a solution exists, it is unique. We can also solve for the cutoff marriage quality at a job with wage w using equation (C.3). Clearly, $q(w^*) = q^*$ and $q(w)$ is a strictly increasing function.

We can now establish the comparative statics with respect to the benefits.

Proposition 7. *An increase in benefits b increases the number of periods the woman stays single and the average quality of the marriage. An increase in benefits b also increases the number of periods the woman stays unemployed and the average wages of the women that become employed.*

Proof. As before, all we need to do is establish that the increase in benefits increases the cutoff qualities and wages. For the single and unemployed cutoffs, notice that equation (C.4) is the only one affected by the change in benefits and that this equation is decreasing in w^*. Therefore, q^* and w^* must increase.

For the single and employed cutoffs, the higher benefits have an indirect effect through the single and unemployed cutoff which we already established was increasing. Intuitively, higher benefits make it better to wait before marrying even when employed because if the woman were to lose the job, she would enjoy those benefits. □

Fertility

An extra dimension that we can consider is fertility. A woman's incentives to have more children are affected by the program. We model this dimension as a binary decision that a woman makes in each period. If a woman decides to have children, she gets one next period with probability π_c. In the model, we limit the number of extra children a woman can have to one. We do this by considering a small state space. That is, a woman can be single with n children, single with $n+1$ children, or married with n and $n+1$ children. A decision to have children while married does not affect the analysis and is thus omitted.

Let us compare the decision of having children when enrolled in the program and when not. The value of being single with n children is

$$V^i_{s,n} = b^i_n + a_n + \beta \left(\lambda^i_n \int_{q=\underline{q}}^{\bar{q}} \max \left\{ V_m(q), \hat{V}^i_{s,n} \right\} \, \mathrm{d}\tilde{F}_n(q) + (1-\lambda^i_n)\hat{V}^i_{s,n} \right),$$

where the i superscript is either 0 or 1, indicating if the woman is participating in the program. $\hat{V}^i_{s,n}$ is the optimal continuation (next period) value of a single woman who has n children in this period. $\hat{V}^i_{s,n} = \max\{V^i_{s,n}, \pi_c V^i_{s,n+1} + (1-\pi_c)V^i_{s,n}\}$.

Also, a_n is the utility flow of having n children. Finally, b^i_n is the transfers that a woman who has n children receives. Some conditions change when a woman enrolls in the program. For instance, if a woman is enrolled in the program, she will receive a transfer $b^1_n > b^0_n = 0$. If $b^1_{n+1} > b^1_n$, the program provides extra incentives to have children (because$b^0_{n+1} = b^0_n$). At the same time, if $\lambda^i_{n+1} < \lambda^i_n$ and$\lambda^1_{n+1} - \lambda^1_n < \lambda^0_{n+1} - \lambda^0_n$ (the effect of an extra child on the arrival of prospects is more negative when participating in the program), there are fewer incentives to have children. When combined with the effect of the higher transfers, the overall effect of the program on fertility is ambiguous.

Proposition 8. *If b is an increasing function of the number of children then fertility will* *increase when b increases. But if more children while single lower the rate of arrival of prospects in the labor and marriage market, then the predictions about fertility become ambiguous.*

Mobility

Now, we introduce the possibility of moving to a new location. Locations are indexed by j and have different characteristics (λ_j). We consider the case where the transfer is lost upon moving to a new location. Opportunities to move to a new location arrive randomly with probability μ. We assume that a married woman does not receive moving opportunities.

$$V_s = b + \beta \left(\lambda \int_{q=\underline{q}}^{\bar{q}} \max\left\{V_m(q), V_s\right\} \; \mathrm{d}F(q) + \mu \int_j \max\left\{V_{s,j}, V_s\right\} \; \mathrm{d}H(j) + (1-\lambda-\mu)V_s \right).$$

The value of being married to a partner with quality q is

$$V_m(q) = q + \beta V_m(q) = \frac{q}{1-\beta}.$$

We take the value of being single in the new location, $V_{s,j}$, as exogenous. While we could make it endogenous, the only relevant assumption is that for each specific new location, the value of being single there is not affected by b.

The decision to migrate is governed by $\max\{V_{s,j}, V_s\}$. Define the set of locations the agent would move to as $J^* := \{j | V_{s,j} \geq V_s\}$. The probability of moving to a new location is given by $\mu H(J^*)$. The expected quality of new locations a woman moves to is given by $\mathbb{E}[V_{s,j} | j \in J^*]$.

Proposition 9. *If b increases, then mobility falls, and those who do migrate, move to better locations.*

Proof. By applying standard dynamic programming arguments, we can show that V_s is a strictly increasing function of b,. [First, the Bellman operator satisfies Blackwell's sufficient conditions for a contraction so there is a fixed point and it is unique. Second, the operator preserves the property of being an increasing function of b, and the operator maps weakly increasing functions of b to strictly increasing functions of b.] Since V_s is a strictly increasing function of b and each $V_{s,j}$ is constant on b, the set J^* is decreasing in b (i.e., when b increases, the set gets smaller as some locations are now excluded). Thus, the probability of moving is lower. Finally, the expected quality of a new location a woman moves to is higher when b is higher. That is because the expected quality when b is lower is a weighted average of the locations that remain when b is higher and the locations that were excluded. By construction, the latter has a lower value than any of the former which proves the result. □

C.3 Causal Forest

We implement the generalized random forest algorithm proposed by Athey et al. (2019). The algorithm, first, trains a causal forest using a full set of covariates and second, estimates conditional average treatment effects (CATE).

An individual tree in a causal forest is trained by drawing a random subsample from the dataset and the sample is split into several nodes to form a tree. Each node in a tree is split using a random subset of covariates and some value of the covariate. The GRF algorithm measures the goodness of a split using heterogeneity across nodes and maximizes the difference in treatment effects across nodes. Then, a prediction is made using a weighted average of each tree's prediction where the weight is the similarity across trees.

We make the following decisions to train a causal forest. First, we use 50% of the full dataset to grow each tree. Second, we train 100,000 trees in a causal forest. Davis and Heller (2017) use 100,000 trees and Beaman et al. (2021) use 250,000 trees but find no meaningful increase in stability when increasing the number of trees from 100,000. In training each tree, we consider $\sqrt{x} + 20$ number of variables for each tree where x is the number of variables and set 20 as the minimum number of observations in each leaf.

We estimate the average treatment effects using a doubly robust augmented-inverse-propensity weighting estimation method (Robins et al., 1994). We report the average treatment effects on the full and treated samples. We also estimate the overlap-weighted average treatment effect recommended by Li et al. (2018) that addresses an issue of estimated propensities being close to 0 or 1 and find similar results to ATE.

Bibliography

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(3):467–506, 2014.

Ran Abramitzky, Leah Platt Boustan, Katherine Eriksson, James J Feigenbaum, and Santiago Pérez. Automated linking of historical data. Technical report, National Bureau of Economic Research, 2019.

Ran Abramitzky, Leah Boustan, Elisa Jácome, and Santiago Pérez. Intergenerational mobility of immigrants in the united states over two centuries. *American Economic Review*, 111(2): 580–608, 2021.

Daron Acemoglu and Pascual Restrepo. Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 128(6):2188–2244, 2020.

Daron Acemoglu, Claire Lelarge, and Pascual Restrepo. Competing with robots: Firm-level evidence from france. In *AEA Papers and Proceedings*, volume 110, pages 383–388. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

Daron Acemoglu, Gary W Anderson, David N Beede, Cathy Buffington, Eric E Childress, Emin Dinlersoz, Lucia S Foster, Nathan Goldschlag, John C Haltiwanger, Zachary Kroff, et al. Automation and the workforce: A firm-level view from the 2019 annual business survey. Technical report, National Bureau of Economic Research, 2022.

Daron Acemoglu, Hans RA Koster, and Ceren Ozgen. Robots and workers: Evidence from the netherlands. Technical report, National Bureau of Economic Research, 2023.

ACF. Why america lost the war on poverty–and how to win it, by frank stricker, 2021.

Jérôme Adda, Brendon McConnell, and Imran Rasul. Crime and the depenalization of cannabis possession: Evidence from a policing experiment. *Journal of Political Economy*, 122(5):1130–1202, 2014.

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

Philipp Ager, Leonardo Bursztyn, Lukas Leucht, and Hans-Joachim Voth. Killer incentives: Rivalry, performance and risk-taking among German fighter pilots, 1939–45. *The Review of Economic Studies*, 89(5):2257–2292, 2022.

Philippe Aghion, Céline Antonin, Simon Bunel, and Xavier Jaravel. What are the labor and product market effects of automation? new evidence from france. 2020.

Anna Aizer, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney. The long-run impact of cash transfers to poor families. *American Economic Review*, 106(4):935–71, April 2016.

Douglas Almond and Janet Currie. Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives*, 25(3):153–72, 2011.

Douglas Almond, Janet Currie, and Valentina Duque. Childhood circumstances and adult outcomes: Act ii. Working Paper 23017, National Bureau of Economic Research, January 2017.

Desmond Ang, Panka Bencsik, Jesse M Bruhn, and Ellora Derenoncourt. Community Engagement with Law Enforcement after High-profile Acts of Police Violence. Technical report, National Bureau of Economic Research, 2024.

James E Archsmith, Anthony Heyes, Matthew J Neidell, and Bhaven N Sampat. The dynamics of inattention in the (baseball) field. Technical report, National Bureau of Economic Research, 2021.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

Kenneth J Arrow. The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173, 1962.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

J Aw. Digitally numb: Doctors are losing hands-on diagnosis skills by relying too much on technology. *National Post*, 2014.

Pierre Azoulay, Joshua S Graff Zivin, and Jialan Wang. Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589, 2010.

Bocar A Ba. Going the Extra Mile: The Cost of Complaint Filing, Accountability, and Law Enforcement Outcomes in Chicago. *Working Paper*, 2020.

A Bacher-Hicks and E de la Campa. The impact of New York City's Stop and Frisk program on crime: The case of police commanders. *Working Paper*, 2020a.

Andrew Bacher-Hicks and Elijah de la Campa. Social Costs of Proactive Policing: The Impact of NYC's Stop and Frisk Program on Educational Attainment. *Working Paper*, 2020b.

Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. How well do automated linking methods perform? lessons from us historical data. Technical report, National Bureau of Economic Research, 2017.

Martha Bailey, Connor Cole, and Catherine Massey. Simple strategies for improving inference with linked data: A case study of the 1850–1930 ipums linked representative historical samples. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53 (2):80–93, 2020.

Martha Bailey, Peter Z. Lin, A.R. Shariq Mohammed, Paul Mohnen, Jared Murray, Mengying Zhang, and Alexa Prettyman. The creation of life-m: The longitudinal, intergenerational family electronic micro-database project. Working paper, 2022.

Oriana Bandiera, Iwan Barankay, and Imran Rasul. Social incentives in the workplace. *The Review of Economic Studies*, 77(2):417–458, 2010.

Erling Barth, Marianne Roed, Pål Schøne, and Janis Umblijs. How robots change within-firm wage inequality. 2020.

Jacob Bastian. The rise of working mothers and the 1975 earned income tax credit. *American Economic Journal: Economic Policy*, 12(3):44–75, 2020.

DH Bayley. Knowledge of the Police. In Maurice Punch, editor, *Control in the Police Organization*, pages 18–35. NCJ-88943, 1983.

Lori Beaman, Dean Karlan, Bram Thuysbaert, and Christopher Udry. Self-selection into credit markets: Evidence from agriculture in mali. Technical report, National Bureau of Economic Research, 2021.

Gary S. Becker. A theory of marriage: Part i. *Journal of Political Economy*, 81(4):813–846, 1973.

Asma Benhenda. Absence, substitutability and productivity: Evidence from teachers. *Labour Economics*, 76:102167, 2022.

Morten Bennedsen, Francisco Pérez-González, and Daniel Wolfenzon. Do CEOs matter? Evidence from hospitalization events. *The Journal of Finance*, 75(4):1877–1911, 2020.

Vivek Bhattacharya and Greg Howard. Rational inattention in the infield. *American Economic Journal: Microeconomics*, 14(4):348–393, 2022.

Marianne Bitler, Jonah Geobach, Hilary Hoynes, and Madeline Zavodny. The impact of welfare reform on marriage and divorce. *Demography*, 41(2):213–36, 2004.

Marianne Bitler, Jonah Gelbach, and Hilary Hoynes. Welfare reform and children's living arrangements. *Journal of Human Resources*, 41(1):1–27, 2006.

Rebecca Blank. Evaluating welfare reform in the united states. *Journal of Economic Literature*, 40(4):1105–1166, 2002.

Christopher Blattman, Donald Green, Daniel Ortega, and Santiago Tobón. Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. Technical report, National Bureau of Economic Research, 2017.

Richard Blundell, Luigi Pistaferri, and Itay Saporta-Eksten. Consumption inequality and family labor supply. *American Economic Review*, 106(2):387–435, 2016.

Alessandra Bonfiglioli, Rosario Crino, Harald Fadinger, and Gino Gancia. Robot imports and firm-level outcomes. 2020.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, page rdae007, 2024.

Anthony A Braga, David Weisburd, and Brandon Turchan. Focused deterrence strategies and crime control: An updated systematic review and meta-analysis of the empirical evidence. *Criminology & Public Policy*, 17(1):205–250, 2018.

William Bratton and Peter Knobler. *The turnaround: How America's Top Cop Reversed the Crime Epidemic*. Random House, 2009.

Michael J. Brien, Stacy Dickert-Conlin, and David A. Weaver. Widows waiting to wed? (re)marriage and economic incentives in social security widow benefits. *Journal of Human Resources*, 39(3):585–623, 2004.

Patrick Brown. Working class americans' views on family policy. Technical report, 2021.

Ryan Brown, Verónica Montalva, Duncan Thomas, and Andrea Velásquez. Impact of violent crime on risk aversion: Evidence from the Mexican drug war. *Review of Economics and Statistics*, 101(5):892–904, 2019.

Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.

Kenneth Burdett and Kenneth L Judd. Equilibrium price dispersion. *Econometrica: Journal of the Econometric Society*, pages 955–969, 1983.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Colin Camerer, George Loewenstein, and Drazen Prelec. Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1):9–64, 2005.

David Card and Gordon B Dahl. Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics*, 126(1):103–143, 2011.

David Card, Raj Chetty, and Andrea Weber. Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *The Quarterly Journal of Economics*, 122(4):1511–1560, 2007.

Ana Rute Cardoso, Paulo Guimaraes, Pedro Portugal, and Hugo Reis. The returns to schooling unveiled. 2018.

Anne Case, Darren Lubotsky, and Christina Paxson. Economic status and health in childhood: The origins of the gradient. *American Economic Review*, 92(5):1308–1334, December 2002.

Aaron Chalfin and Justin McCrary. Are US Cities Underpoliced? Theory and Evidence. *Review of Economics and Statistics*, 100(1):167–186, 2018.

Aaron Chalfin, Michael LaForest, and Jacob Kaplan. Can precision policing reduce gun violence? evidence from “gang takedowns” in new york city. *Journal of Policy Analysis and Management*, 40(4):1047–1082, 2021a.

Aaron Chalfin, David Mitre-Becerril, and Morgan C. Williams. Evidence that curtailing proactive policing can reduce major crime. *Working Paper*, 2021b.

Aaron Chalfin, Benjamin Hansen, Emily K Weisburst, and Morgan C Williams Jr. Police force size and civilian race. *American Economic Review: Insights*, 4(2):139–58, 2022.

Andrea Cann Chandrasekher. The effect of police slowdowns on crime. *American Law and Economics Review*, 18(2):385–437, 2016.

Marisa Chappell. Self-sufficiency, welfare & employment, 2011.

Daniel L Chen, Tobias J Moskowitz, and Kelly Shue. Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242, 2016.

Jiafeng Chen and Jonathan Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 2023.

Cheng Cheng and Wei Long. The Effect of Highly Publicized Police-Related Deaths on Policing and Crime: Evidence from Large US Cities. *Working Paper*, 2018.

Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. The association between income and life expectancy in the united states, 2001-2014. *Jama*, 315(16):1750–1766, 2016.

P-A Chiappori, Steven Levitt, and Timothy Groseclose. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review*, 92(4):1138–1151, 2002.

Eric Chyn and Lawrence F Katz. Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives*, 35(4):197–222, 2021.

Jonathan P Cohen, Andrew C Johnston, and Attila S Lindner. Skill depreciation during unemployment: Evidence from panel data. Technical report, National Bureau of Economic Research, 2023.

Alain Cohn, Jan Engelmann, Ernst Fehr, and Michel André Maréchal. Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2):860–885, 2015.

Thomas Cornelissen, Christian Dustmann, and Uta Schönberg. Peer effects in the workplace. *American Economic Review*, 107(2):425–456, 2017.

Jacqueline Craig, Katherine Eriksson, and Gregory T. Niemesh. Marriage and the intergenerational mobility of women: Evidence from marriage certificates 1850-1910. Technical report, 2021.

David Cutler, Angus Deaton, and Adriana Lleras-Muney. The determinants of mortality. *Journal of Economic Perspectives*, 20(3):97–120, 2006.

Wolfgang Dauth, Sebastian Findeisen, Jens Suedekum, and Nicole Woessner. The adjustment of labor markets to robots. *Journal of the European Economic Association*, 19(6):3104–3153, 2021.

Jonathan Davis and Sara B Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50, 2017.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

Tanaya Devi and Roland G Fryer Jr. Policing the Police: The Impact of "Pattern-or-Practice" Investigations on Crime. *National Bureau of Economic Research*, 2020.

Stacy Dickert-Conlin and Scott Houser. Eitc and marriage. *National Tax Journal*, pages 25–40, 2002.

Marcus Dillinder. Social security and divorce. *The B.E. Journal of Economic Analysis and Policy*, 16(2):931–971, 2016.

Michael Dinerstein, Rigissa Megalokonomou, and Constantine Yannelis. Human capital depreciation and returns to experience. *American Economic Review*, 112(11):3725–62, 2022.

Jay Dixon, Bryan Hong, and Lynn Wu. The robot revolution: Managerial and employment consequences for firms. *Management Science*, 67(9):5586–5605, 2021.

Patricio Dominguez, Magnus Lofstrom, and Steven Raphael. The Effect of Sentencing Reform on Crime Rates: Evidence from California's Proposition 47. *Institute of Labor Economics (IZA)*, 2019.

Cheryl Doss. Intrahousehold bargaining and resource allocation in developing countries. *The World Bank Research Observer*, 28(1):52–78, 2013.

Mirko Draca, Stephen Machin, and Robert Witt. Panic on the streets of London: Police, crime, and the July 2005 terror attacks. *American Economic Review*, 101(5):2157–81, 2011.

Mark Duggan and Steven D Levitt. Winning isn't everything: Corruption in sumo wrestling. *American Economic Review*, 92(5):1594–1605, 2002.

Darren Duxbury, Tommy Gärling, Amelie Gamble, and Vian Klass. How emotions influence behavior in financial markets: A conceptual analysis and emotion-based account of buy-sell preferences. *The European Journal of Finance*, 26(14):1417–1438, 2020.

Per-Anders Edin and Magnus Gustavsson. Time out of work and skill depreciation. *ILR Review*, 61(2):163–180, 2008.

Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

Shari Eli and Laura Salisbury. Patronage politics and the development of the welfare state: Confederate pensions in the american south. *The Journal of Economic History*, 76(4): 1078–1112, 2016.

Shari Eli, Price Fishback, Adriana Lleras-Muney, and James Uguccioni. The emergence of the modern welfare system: Evidence from the mothers' pension program. Technical report, Working Paper, 2022.

Ingrid Gould Ellen and Katherine O'Regan. Crime and US cities: Recent patterns and implications. *The Annals of the American Academy of Political and Social Science*, 626(1): 22–38, 2009.

Ozkan Eren and Naci Mocan. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3):171–205, 2018.

William N Evans and Emily G Owens. COPS and Crime. *Journal of Public Economics*, 91 (1-2):181–201, 2007.

Armin Falk and Andrea Ichino. Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57, 2006.

Henry S Farber. Is tomorrow another day? the labor supply of new york city cabdrivers. *Journal of Political Economy*, 113(1):46–82, 2005.

Eric Fesselmeyer. The impact of temperature on labor quality: Umpire accuracy in major league baseball. *Southern Economic Journal*, 88(2):545–567, 2021.

John Fitzgerald and David Ribar. Welfare reform and female headship. *Demography*, 41: 189–212, 06 2004.

Edward G Fox. Do taxes affect marriage? lessons from history. 2017.

Michael Friedson and Patrick Sharkey. Neighborhood inequality after the crime decline. *Annals of the American Academy of Political and Social Science*, 660(1):341–58, 2015.

Shigeru Fujita and Giuseppe Moscarini. Recall and unemployment. *American Economic Review*, 107(12):3875–3916, 2017.

Claudia Goldin. The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, 96(2):1–21, 2006.

Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

Felipe Gonçalves. Do police unions increase officer misconduct? Technical report, Working paper, 2021.

Felipe M Gonçalves and Steven Mello. Police discretion and public safety. Technical report, National Bureau of Economic Research, 2023.

Bo Gong, James P Nugent, William Guest, William Parker, Paul J Chang, Faisal Khosa, and Savvas Nicolaou. Influence of artificial intelligence on canadian medical students' preference for radiology specialty: A national survey study. *Academic radiology*, 26(4):566–577, 2019.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Kathryn Graddy and Carl Lieberman. Death, bereavement, and creativity. *Management Science*, 64(10):4505–4514, 2018.

J Grogger and LA Karoly. Employment, labor supply, and earnings. *Welfare reform: Effects of a Decade of Change*, pages 134–154, 2005.

Jeff Grogger and Stephen G Bronars. The effect of welfare payments on the marriage and fertility behavior of unwed mothers: Results from a twins experiment. *Journal of Political Economy*, 109(3):529–545, 2001.

André Grow and Jan Van Bavel. Assortative mating and the reversal of gender inequality in education in europe: An agent-based model. *PloS one*, 10(6), 2015.

José R Guardado and Nicolas R Ziebarth. Worker investments in safety, workplace accidents, and compensating wage differentials. *International Economic Review*, 60(1):133–155, 2019.

Jonathan Guryan, Kory Kroft, and Matthew J Notowidigdo. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4):34–68, 2009.

Robert Paul Hartley, Carlos Lamarche, and James P. Ziliak. Welfare reform and the intergenerational transmission of dependence. *Journal of Political Economy*, 130(3):523–565, 2022.

MD Hayward and BK Gorman. The long arm of childhood: the influence of early life social conditions on men's mortality. *Demography*, 41(1):87–107, 2004.

Paul Heaton. Understanding the Effects of Antiprofiling Policies. *The Journal of Law and Economics*, 53(1):29–64, 2010.

Nathaniel Hendren and Ben Sprung-Keyser. A unified welfare analysis of government policies. *The Quarterly Journal of Economics*, 135(3):1209–1318, 2020.

Vincent E Henry. *Death work: Police, trauma, and the psychology of survival.* Oxford University Press, 2004.

Chris M Herbst. The impact of the earned income tax credit on marriage and divorce: Evidence from flow data. *Population Research and Policy Review*, 30(1):101–128, 2011.

Paul J Hirschfield. Exceptionally lethal: American police killings in a comparative perspective. *Annual Review of Criminology*, 6:471–498, 2023.

Johannes Hirvonen, Aapo Stenhammar, and Joonas Tuhkuri. New evidence on the effect of technology on employment and skill demand. *Available at SSRN 4081625*, 2022.

Günter J Hitsch, Ali Hortaçsu, and Dan Ariely. What makes you click?—mate preferences in online dating. *Quantitative Marketing and Economics*, 8(4):393–427, 2010.

Howard Hogan and Gregg Robinson. What the census bureau's coverage evaluation programs tell us about differential undercount. *Washington, DC, Census Bureau*, 1993.

Justin E Holz, Roman G Rivera, and Bocar A Ba. Peer effects in police use of force. *American Economic Journal: Economic Policy*, 15(2):256–291, 2023.

Long Hong and Salvatore Lattanzio. The peer effect on future wages in the workplace. *Available at SSRN 4052587*, 2022.

V Joseph Hotz and John Karl Scholz. *3. The Earned Income Tax Credit.* University of Chicago Press, 2007.

Hilary Hoynes, Diane Whitmore Schanzenbach, and Douglas Almond. Long-run impacts of childhood access to the safety net. *American Economic Review*, 106(4):903–34, April 2016.

Hilary Williamson Hoynes. Work, welfare, and family structure: What have we learned? Working Paper 5644, National Bureau of Economic Research, July 1996.

Anders Humlum. *Robot Adoption and Labor Market Dynamics.* Rockwool Foundation Research Unit, 2022.

Robert M. Hutchens. Welfare, remarriage, and marital search. *The American Economic Review*, 69(3):369–379, 1979. ISSN 00028282.

Louis S Jacobson, Robert J LaLonde, and Daniel G Sullivan. Earnings losses of displaced workers. *The American Economic Review*, pages 685–709, 1993.

Simon Jäger and Jörg Heining. How substitutable are workers? evidence from worker deaths. Technical report, National Bureau of Economic Research, 2022.

Xavier Jaravel, Neviana Petkova, and Alex Bell. Team-specific capital and innovation. *American Economic Review*, 108(4-5):1034–1073, 2018.

Gregor Jarosch. Searching for job security and the consequences of job loss. *Econometrica*, 91(3):903–942, 2023.

Benjamin F Jones and Benjamin A Olken. Do leaders matter? national leadership and growth since world war ii. *The Quarterly Journal of Economics*, 120(3):835–864, 2005.

Charles I Jones and Peter J Klenow. Beyond gdp? welfare across countries and time. *American Economic Review*, 106(9):2426–57, 2016.

Lawrence M Kahn. The sports business as a labor market laboratory. *Journal of Economic Perspectives*, 14(3):75–94, 2000.

Kyogo Kanazawa, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe. Ai, skill, and productivity: The case of taxi drivers. Technical report, National Bureau of Economic Research, 2022.

Jacob Kaplan. Jacob Kaplan's Concatenated Files: Uniform Crime Reporting Program Data: Offenses Known and Clearances by Arrest, 1960-2018. *Inter-university Consortium for Political and Social Research (ICPSR)*, 2020a.

Jacob Kaplan. Jacob Kaplan's Concatenated Files: Uniform Crime Reporting Program Data: Law Enforcement Officers Killed and Assaulted (LEOKA) 1960-2018. *Inter-university Consortium for Political and Social Research (ICPSR)*, 2020b.

Jacob Kaplan. Jacob Kaplan's Concatenated Files: Uniform Crime Reporting (UCR) Program Data: Supplementary Homicide Reports, 1976-2019. *Inter-university Consortium for Political and Social Research (ICPSR)*, 2020c.

Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, et al. Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175, 2018.

Melissa Schettini Kearney. Is there an effect of incremental welfare benefits on fertility behavior? a look at the family cap. *Journal of Human Resources*, 39(2):295–325, 2004.

Jerry W Kim and Brayden G King. Seeing stars: Matthew effects and status bias in major league baseball umpiring. *Management Science*, 60(11):2619–2644, 2014.

Jeffrey R Kling, Jeffrey B Liebman, and Lawrence F Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

Michael Koch, Ilya Manuylov, and Marcel Smolka. Robots and firms. *The Economic Journal*, 131(638):2553–2584, 2021.

Issa Kohler-Hausmann. *Misdemeanorland: Criminal Courts and Social Control in an Age of Broken Windows Policing.* Princeton University Press, 2018.

Jacob Kohlhepp and Robert McDonough. Workplace injury and labor supply within an organization. *Working Paper*, 2022.

Kory Kroft, Fabian Lange, and Matthew J Notowidigdo. Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics*, 128(3):1123–1167, 2013.

Rafael Lalive. Unemployment benefits, unemployment duration, and post-unemployment jobs: A regression discontinuity approach. *American Economic Review*, 97(2):108–112, 2007.

Wang-Sheng Lee and Terra McKinnish. The marital satisfaction of differently aged couples. *Journal of population economics*, 31(2):337–362, 2018.

Mark H Leff. Consensus for reform: The mothers' pension movement in the progressive era. *Social Service Review*, 47(3):397–417, 1973.

Joscha Legewie. Racial profiling and use of force in police stops: How local events trigger periods of increased discrimination. *American Journal of Sociology*, 122(2):379–424, 2016.

Steven D Levitt. Using electoral cycles in police hiring to estimate the effects of police on crime. *American Economic Review*, 87(3):270–290, 1997.

Steven D Levitt. The relationship between crime reporting and police: Implications for the use of Uniform Crime Reports. *Journal of Quantitative Criminology*, 14(1):61–81, 1998.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Lee A Lillard, Michael J Brien, and Linda J Waite. Premarital cohabitation and subsequent marital. *Demography*, 32(3):437–457, 1995.

Andrew W Lo, Dmitry V Repin, and Brett N Steenbarger. Fear and greed in financial markets: A clinical study of day-traders. *American Economic Review*, 95(2):352–359, 2005.

Colin Loftin, David McDowall, and Min Xie. Underreporting of homicides by police in the United States, 1976-2013. *Homicide Studies*, 21(2):159–174, 2017.

Hamish Low, Costas Meghir, Luigi Pistaferri, and Alessandra Voena. Marriage, labor supply and the dynamics of the social safety net. Working Paper 24356, National Bureau of Economic Research, February 2018.

Robert E Lucas Jr. On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42, 1988.

Shelly Lundberg, Robert A Pollak, and Jenna Stearns. Family inequality: Diverging patterns in marriage, cohabitation, and childbearing. *Journal of Economic Perspectives*, 30(2): 79–102, 2016.

Stephen Machin and Alan Manning. The causes and consequences of long-term unemployment in europe. *Handbook of Labor Economics*, 3:3085–3139, 1999.

Nicole Maestas, Kathleen J Mullen, Alexander Strand, et al. Does delay cause decay? the effect of administrative decision time on the labor force participation and earnings of disability applicants. Technical report, National Bureau of Economic Research, 2015.

David A Malueg and Andrew J Yates. Testing contest theory: Evidence from best-of-three tennis matches. *The Review of Economics and Statistics*, 92(3):689–692, 2010.

Riccardo Marchingiglio and Michael Poyker. The economics of gender-specific minimum wage legislation. Technical report, 2021.

Otwin Marenin. Cheapening Death: Danger, Police Street Culture, and the Use of Deadly Force. *Police Quarterly*, 19(4):461–487, 2016.

D Mark Anderson, Benjamin Hansen, and Daniel I Rees. Medical marijuana laws, traffic fatalities, and alcohol consumption. *The Journal of Law and Economics*, 56(2):333–369, 2013.

Thomas B Marvell and Carlisle E Moody. Specification problems, police levels, and crime rates. *Criminology*, 34(4):609–646, 1996.

Alexandre Mas. Pay, Reference points, and Police Performance. *The Quarterly Journal of Economics*, 121(3):783–821, 2006.

Alexandre Mas and Enrico Moretti. Peers at work. *American Economic Review*, 99(1): 112–145, 2009.

J. J. McCall. Economics of information and job search. *The Quarterly Journal of Economics*, 84(1):113–126, 1970.

Justin McCrary. The effect of court-ordered hiring quotas on the composition and quality of police. *American Economic Review*, 97(1):318–353, 2007.

Steven Mello. Speed Trap or Poverty Trap? Fines, Fees, and Financial Wellbeing. *Working Paper*, 2018.

Steven Mello. More COPS, Less Crime. *Journal of Public Economics*, 172:174–200, 2019.

Katherine Michelmore. The earned income tax credit and union formation: The impact of expected spouse earnings. *Review of Economics of the Household*, pages 1–30, 2016.

Brian M Mills. Technological innovations in monitoring and evaluation: Evidence of performance impacts among major league baseball umpires. *Labour Economics*, 46:189–199, 2017.

Robert Moffitt. Incentive effects of the u.s. welfare system: A review. *Journal of Economic Literature*, 30(1):1–61, 1992. ISSN 00220515.

Robert Moffitt et al. An economic model of welfare stigma. *American Economic Review*, 73 (5):1023–1035, 1983.

Robert A. Moffitt, Brian J. Phelan, and Anne E. Winkler. Welfare rules, incentives, and family structure. Working Paper 21257, National Bureau of Economic Research, June 2015.

Clayton J Mosher, Terance D Miethe, and Timothy C Hart. *The mismeasure of crime*. Sage Publications, 2010.

Tobias Moskowitz and L Jon Wertheim. *Scorecasting: The Hidden Influences Behind How Sports are Played and Games are Won*. Crown Archetype, 2011.

John Mullahy and Edward C Norton. Why transform Y? A critical assessment of dependent-variable transformations in regression models for skewed and sometimes-zero outcomes. Technical report, National Bureau of Economic Research, 2022.

Christopher A Neilson and Seth D Zimmerman. The Effect of School Construction on Test Scores, School Enrollment, and Home Prices. *Journal of Public Economics*, 120:18–31, 2014.

Whitney K Newey. Two-step series estimation of sample selection models. *The Econometrics Journal*, 12:S217–S229, 2009.

Jack Nicas and Zach Wichter. A worry for some pilots: Their hands-on flying skills are lacking. *New York Times. https://www. nytimes. com/2019/03/14/business/automated-planes. html*, 2019.

Austin Nichols and Jesse Rothstein. The earned income tax credit? in economics of means-tested transfer programs in the united states, 2016.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*, 2023.

Claudia Olivetti and M. Daniele Paserman. In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940. *American Economic Review*, 105(8):2695–2724, August 2015. doi: 10.1257/aer.20130821.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

Emily Owens, David Weisburd, Karen L Amendola, and Geoffrey P Alpert. Can you build a better cop? Experimental evidence on supervision, training, and policing in the community. *Criminology & Public Policy*, 17(1):41–87, 2018.

Emily Greene Owens. Cops and Cuffs. In *Lessons from the economics of crime: What reduces offending?*, page 17. MIT Press Cambridge, 2013.

Christopher A Parsons, Johan Sulaeman, Michael C Yates, and Daniel S Hamermesh. Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, 101(4): 1410–1435, 2011.

Petra Persson. Social insurance and the marriage market. 2017.

Deepak Premkumar. Intensified Scrutiny and Bureaucratic Effort: Evidence from Policing and Crime After High-Profile, Officer-Involved Fatalities. *Working Paper*, 2020.

Canice Prendergast. Selection and Oversight in the Public Sector, with the Los Angeles Police Department as an Example. *National Bureau of Economic Research*, 2001.

Canice Prendergast. 'drive and wave': The response to lapd police reforms after rampart. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2021-25), 2021.

David J Price and Jae Song. The long-term effects of cash assistance. 2018.

Joseph Price and Justin Wolfers. Racial discrimination among nba referees. *The Quarterly Journal of Economics*, 125(4):1859–1887, 2010.

Joseph Price, Lars Lefgren, and Henry Tappen. Interracial workplace cooperation: Evidence from the nba. *Economic Inquiry*, 51(1):1026–1034, 2013.

Joseph Price, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. Combining family history and machine learning to link historical records. Technical report, National Bureau of Economic Research, 2019.

Matthew L Renner. Using multiple flawed measures to construct valid and reliable rates of homicide by police. *Homicide studies*, 23(1):20–40, 2019.

Jason L. Riley. Good Policing Saves Black Lives. *Wall Street Journal*, 2020. URL https://www.wsj.com/articles/good-policing-saves-black-lives-11591052916.

Roman Rivera and Bocar A. Ba. The Effect of Police Oversight on Crime and Allegations of Misconduct: Evidence from Chicago. *U of Penn, Inst for Law & Econ Research Paper*, (19-42), 2019.

Dorothy E Roberts. The value of black mothers' work. *Conn. L. Rev.*, 26:871, 1993.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Laura Salisbury. Women's income and marriage markets in the us: Evidence from civil war pensions. *Journal of Economic History*, 7(1):1–38, 2017.

Danielle H Sandler and Ryan Sandler. Multiple Event Studies in Public Finance and Labor Economics: A Simulation Study with Applications. *Journal of Economic and Social Measurement*, 39(1-2):31–57, 2014.

Johannes F Schmieder, Till von Wachter, and Stefan Bender. The effect of unemployment benefits and nonemployment durations on wages. *American Economic Review*, 106(3): 739–777, 2016.

Lan Shi. The Limit of Oversight in Policing: Evidence from the 2001 Cincinnati Riot. *Journal of Public Economics*, 93(1-2):99–113, 2009.

Michael Sierra-Arévalo. The Commemoration of Death, Organizational Memory, and Police Culture. *Criminology*, 57(4):632–658, 2019.

Michael Sierra-Arévalo. American policing and the danger imperative. *Law & Society Review*, 55(1):70–103, 2021.

Christianna Silva. Law Professor On Misdemeanor Offenses And Racism In The Criminal System. *NPR*, 2020. URL https://www.npr.org/sections/live-updates-protests-for-racial-justice/2020/06/12/876221163/law-professor-on-how-misdemeanors-sweep-blacks-into-the-criminal-system.

Theda Skocpol. *Protecting Soldiers and Mothers*. Harvard University Press, 1995.

CarlyWill Sloan. The Effect of Violence Against Police on Police Behavior. *Working Paper*, 2019.

Tymon Sloczynski. Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *The Review of Economics and Statistics*, 104(3):501–509, 2022.

Alice Speri. Police Make More than 10 Million Arrests a Year, But That Doesn't Mean They're Solving Crimes. *The Intercept*, 2020. URL https://theintercept.com/2019/01/31/arrests-policing-vera-institute-of-justice/.

Christopher M Sullivan and Zachary P O'Keeffe. Evidence that curtailing proactive policing can reduce major crime. *Nature Human Behaviour*, 1(10):730–737, 2017.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

Elena Ashtari Tafti. *Technology, Skills, and Performance: The Case of Robots in Surgery*. Institute for Fiscal Studies, 2022.

Julien O Teitler, Nancy E Reichman, Lenna Nepomnyaschy, and Irwin Garfinkel. Effects of welfare participation on marriage. *Journal of Marriage and Family*, 71(4):878–891, 2009.

Peter Thompson. Learning by doing. *Handbook of the Economics of Innovation*, 1:429–476, 2010.

Jan C Van Ours and Milan Vodopivec. Does reducing unemployment insurance generosity reduce job match quality? *Journal of Public Economics*, 92(3-4):684–695, 2008.

W Kip Viscusi and Joseph E Aldy. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, 27:5–76, 2003.

Maarten J Voors, Eleonora E M Nillesen, Philip Verwimp, Erwin H Bulte, Robert Lensink, and Daan P Van Soest. Violent conflict and behavior: A field experiment in Burundi. *American Economic Review*, 102(2):941–964, 2012.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Deborah E Ward. *The White Welfare State: The Racialization of US Welfare Policy*. University of Michigan Press, 2009.

Emily K Weisburst. Safety in Police Numbers: Evidence of Police Effectiveness from Federal COPS Grant Applications. *American Law and Economics Review*, 21(1):81–109, 2019.

Emily K Weisburst. "Whose help is on the way?" The importance of individual police officers in law enforcement outcomes. *Journal of Human Resources*, 59(4):0720–11019R2, 2024.

Leslie A. Whittington and James Alm. 'til death or taxes do us part: The effect of income taxation on divorce. *Journal of Human Resources*, 32(2):388–412, 1997.

Mark T Williams. Mlb umpires missed 34,294 ball-strike calls in 2018. *Bring on Robo-Umps*, 2019.

Jennifer Withrow. The farm woman's problems: Fram crisis in the us south and migration to the city, 1920-1940. Technical report, Census Bureau, 2021.

Xiaokai Yang and Jeff Borland. A microeconomic mechanism for economic growth. *Journal of Political Economy*, 99(3):460–482, 1991.

Jeffrey E Zabel. An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. *Journal of Human Resources*, pages 479–506, 1998.

Franklin E Zimring. *The City that Became Safe: New York's Lessons for Urban Crime and its Control*. Oxford University Press, 2011.