



Essays in Labor Economics and Child Welfare

Citation

Grimon, Marie-Pascale. 2021. Essays in Labor Economics and Child Welfare. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368316>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of Public Policy
have examined a dissertation entitled

"Essays in Labor Economics and Child Welfare"

presented by Marie-Pascale Grimon

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature Amanda Pallais

Typed name: Professor Amanda Pallais, Chair

Signature Mark Shepard

Typed name: Professor Mark Shepard

Signature Joseph Doyle

Typed name: Professor Joseph Doyle

Signature Marcella Alsan

Typed name: Professor Marcella Alsan

Date: May 4, 2021

Essays in Labor Economics and Child Welfare

A dissertation presented

by

Marie-Pascale Grimon

to

The Harvard Kennedy School

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Public Policy

Harvard University

Cambridge, Massachusetts

May 2021

© 2021 Marie-Pascale Grimon

All rights reserved.

Dissertation Advisor:
Professor Amanda Pallais

Author:
Marie-Pascale Grimon

Essays in Labor Economics and Child Welfare

Abstract

This dissertation comprises of three independent essays on the economics of child welfare. Essays respectively examine the effects of child welfare on parents (Chapter 1), on child deaths (Chapter 2), and the effects of an algorithmic tool to improve child welfare practice (Chapter 3).

One-third of American children's homes are investigated by child welfare agencies for abuse and neglect, yet the impacts of these investigations on parents are not well studied. In the first essay, I establish that months in which a child welfare investigation begins are pivotal moments in parents' lives. I then identify causal effects of child welfare using a novel research design which combines an event study set-up with imperfect random assignments of investigators. Opening a child welfare case considerably increases mothers' enrollment in mental health and substance abuse services, suggesting that child welfare encourages parents to seek help they may need. The effects of opening a child welfare case persist even when no child is removed from the home, and the increase in use of substance abuse services persists for up to six years. In contrast, fathers are much less responsive to child welfare involvement both in the short and long run.

Despite prominent disparities in child mortality, the child protection system's impact on child death disparities has received limited attention. In my second essay, I leverage close to random assignments of workers to families in an American county to provide the first causal estimates of the child protection system on child deaths reported to the medical examiner. I find that opening a child welfare case drastically reduces such child deaths. Extrapolating from these local average treatment effects, the black-white child mortality gap

would be two to three times larger without child welfare involvement. Child welfare plays a key role in addressing mortality inequality among children.

Machine learning-based tools have drawn increasing interest from public policy practitioners, yet our understanding of the effectiveness of such tools when paired with human decision makers is limited. In the third essay which is co-authored we use a randomized control trial to evaluate the effects of an established algorithmic decision aid tool implemented by an American child welfare agency. Slightly less than halfway through the trial, this paper presents preliminary evidence on the effects of showing a child's predicted risk score on child welfare decision outcomes. Even when scores are not shown, workers target case openings and services more to high-scoring children relative to low-scoring children. Workers are thus already sensitive to underlying risk. Still, showing a score further increases the likelihood that high-scoring children have a founded allegation of abuse or neglect. Showing a score also decreases the likelihood that families not visited by child welfare are called back in. These preliminary results suggest that the tool may be improving targeting of child welfare visits.

Contents

Title Page	i
Copyright	ii
Abstract	iii
Table of Contents	v
Acknowledgments	xi
Dedication	xii
 1 Effects of the Child Protection System on Parents	 1
1.1 Introduction	1
1.2 Child Welfare in Allegheny County, PA	6
1.2.1 The Sequence of Events of a Child Welfare Referral	6
1.2.2 The Data and Sample	8
1.3 Methodology: Propensity Instruments with Panel Data	11
1.3.1 The Propensity Instrument: Investigator Assignments	11
1.3.2 Event-Study Analysis	14
1.3.3 Combining Propensity Instruments with an Event-Study	16
1.3.4 Empirical Specifications	18
1.3.5 Validity of the Identification Strategy	20
1.4 Investigators' impacts on parents	29
1.4.1 In the Months Following the Referral	29
1.4.2 Several Years after the Referral	35
1.5 Additional Analyses	38
1.5.1 Child Welfare Outcomes	38
1.5.2 Permutation Tests	39
1.5.3 Robustness Checks	41
1.5.4 Attrition	45
1.6 Conclusion	46
 2 Saved from Death? Effects of the Child Protection System on Child Deaths Reported to the Medical Examiner	 48
2.1 Introduction	48

2.2	Child Welfare and Child Deaths	52
2.2.1	Background on Child Welfare	52
2.2.2	Literature Review from Social Work	52
2.2.3	Data	54
2.2.4	Naive Descriptive Approach	55
2.2.5	Disparities in Medically Examined Deaths and Child Welfare Involvement	57
2.3	Methodology: Propensity Instruments	59
2.3.1	The Propensity Instrument: Investigator Assignments	59
2.3.2	Empirical Specifications	61
2.3.3	Validity of the Identification Strategy	63
2.4	Results	67
2.4.1	Effects of Child Welfare on Child Medically Examined Deaths	67
2.4.2	Relationship to Inequality	73
2.4.3	Robustness Checks	75
2.5	Conclusion	79
3	The Effect of Algorithmic Tools on Child Welfare Decision Making: A Partway Evaluation	81
3.1	Introduction	81
3.2	Child Welfare and Machine Learning for Prediction	83
3.2.1	Child Welfare in the Intervention County	84
3.2.2	Review of Existing Work	85
3.2.3	The Decision Aid Tool in our Context	87
3.3	Experimental Design	89
3.3.1	A Randomized Control Trial	89
3.3.2	Empirical Specification	90
3.3.3	Sample	91
3.3.4	Balance Tests	93
3.4	Results	96
3.4.1	Process Outcomes	96
3.4.2	Preliminary Results	97
3.5	Conclusion	104
	References	106
	Appendix A Appendix to Chapter 1	114
A.1	Figures	114
A.2	Tables	117

Appendix B	Appendix to Chapter 2	127
B.1	Figures	127
B.2	Tables	129
Appendix C	Appendix to Chapter 3	135
C.1	Figures	135
C.2	Tables	137

List of Tables

1.1	Descriptive Statistics on Sample	10
1.2	P-values of Pre-trends F-test	24
1.3	First Stage	26
1.4	Parents Receiving Services (%), 0-12m after referral relative to 1-6m prior . .	32
1.5	Moms Receiving Any Services 3-6 years After Referral Relative to Prior (%)	37
1.6	Child Welfare Outcomes 3-6 years After Referral Relative to Prior	39
1.7	Robustness Tests for Year Results	42
2.1	Number of Children and Rates of Medically Examined Deaths, May 2009 - Dec 2018	56
2.2	Inequality and Death	58
2.3	First Stage	65
2.4	Results and Heterogeneity	68
2.5	Robustness to Alternative Specifications	76
3.1	Descriptive Statistics	92
3.2	Test of Randomized Assignment of Treatment Status	95
3.3	Process Outcomes (First Stage)	96
3.4	Impacts of Score Availability for this Referral	99
3.5	Longer-run Outcomes	101
3.6	Re-referrals of Screened Out Cases	104
A.1	Number of Observations, May 2009 - Dec 2018	117
A.2	Randomization Test	118
A.3	Tests for Monotonicity	121
A.4	Robustness Tests for Month Results	122
A.5	Moms Receiving Any Services 3-6 years After Referral Relative to Prior (%)	126
B.1	Number of Child-Referrals and Rates of Medically Examined Deaths, May 2009 - Dec 2018	129
B.2	Probability of a Medically Examined Death Among Children Always Screened Out	130

B.3	Randomization Test	131
B.4	Tests for Monotonicity	134
C.1	Characteristics of Those with a Score Reported to be Seen	137

List of Figures

1.1	Child Welfare Flow Chart	7
1.2	Event-Study Plots Relative to the Month of the Referral	15
1.3	Means of Mothers' Outcomes by Investigator Median Propensity to Open a Case	17
1.4	Monotonicity Test	28
1.5	Assigning Mothers to Investigators One Standard Deviation More Likely to Open a Case, Month Level	31
1.6	Assigning Mothers to Investigators One Standard Deviation More Likely to Open a Case with a Child Removal, Year Level	36
1.7	Assigning Mothers to Investigators Above Median Propensity to Open a Case, Year Level	44
2.1	Effects of Having a Child Welfare Case Opened on Deaths Examined by a Medical Examiner	71
2.2	Effects of Being Removed from Home on Deaths examined by a Medical Examiner	72
3.1	Algorithm Score Distribution by Treatment Status	94
A.1	Density of Propensity Instruments	115
A.2	Selective Attrition - Mothers Not Involved in Any Program	116
B.1	Monotonicity Test	128
C.1	Placement Risk by Risk Score on Validation Holdout Sample	136

Acknowledgments

First and foremost, I would like to thank my advisors Amanda Pallais, Joseph Doyle, Mark Shepard and Marcella Alsan. Their mentorship and advice have been invaluable throughout this PhD endeavor. I am also grateful for guidance and encouragement from many others including Richard Zeckhauser, Will Dobbie, Dan Levy, and Teddy Svoronos. I would also like to express my indebtedness to Michael Boozer and Catherine Panter-Brick without whom I would probably not have started this PhD.

At Harvard, I would like to thank all the students which I have had the pleasure to teach, especially in the Economics Sophomore Tutorial (Ec 970). They have been a constant source of motivation for my research and have heightened my awareness of the academic literature. I would also like to acknowledge financial support provided by Harvard University and by the Harvard Kennedy School for my dissertation.

I am obliged to Jacques Leblanc for introducing me to this topic of child welfare through his memoirs. His many years serving foster care children in Belgium and their families in need have profoundly influenced my work and have been a constant source of inspiration.

This research of course would not have been possible without the confidence of the Allegheny County Department of Human Services, as well as that of my other local partner agency, which have both generously made their data accessible to me. I am also thankful to Rhema Vaithianathan and the Center for Social Data Analytics for their trust and the opportunity to work with them. All of these partners have provided constant feedback and advice. I have benefited immensely from their expertise and experience, from the numerous discussions with practitioners, and from field visits.

Finally, I would like to wholeheartedly thank my family and friends. It takes a village to finish a PhD, and they have been an unwavering source of help in the process. Their love and day after day encouragements have made all of the difference.

To Jacques Leblanc, the children and families he so dearly cared for

Chapter 1

Effects of the Child Protection System on Parents

1.1 Introduction

One in three American children lives in a home investigated by child protective services. Among Black children that number is 53% (Kim et al., 2017). Therefore, many parents are exposed to the scrutiny of child protection agencies which are commonly referred to as child welfare.

Notwithstanding extensive research on inequality and the importance of investing in disadvantaged children, little focus has been placed on child welfare whose mission is to protect children and to ensure their basic needs are met. Professionals, such as doctors or teachers, are mandated by law to call child welfare to report families where they suspect potential child abuse or neglect. Child welfare then investigates some of these referrals. When an investigation reveals that the child is indeed at risk, child welfare opens a case to stay involved with this family. Involving parents to address the concerns becomes critical (Sankaran et al., 2015; National Center on Substance Abuse and Child Welfare, 2019). For example, a concern for child safety might stem from parental substance abuse: parents with addictions may not be able to care for a child when intoxicated. Since the opioid epidemic,

parental drug abuse factored into the decision to remove a child for one-third of children who were removed in 2017 (Sepulveda & Williams, 2019). Parents can address this concern by participating in substance abuse treatment. This is but one example of how, in ensuring the safety of children, child welfare may have impacts on parents.

However, to the best of my knowledge, there has been no study on the causal impacts of child welfare on parents. Child welfare workers could provide the services and extrinsic motivations parents may need to facilitate a safe family environment at home. At the same time, parents' mental health may deteriorate as a result of the investigation, and parents may face the legal consequences of any reprehensible actions.

This paper examines what could have changed in parents' lives for better and worse, using unique administrative data from Allegheny county in Pennsylvania (Pittsburgh). Specifically, for all parents investigated by child welfare in this county from 2009 to 2018 for child neglect, I observe each month whether they are interacting with a variety of social services, health services, and the criminal justice system. I begin by providing descriptive statistics on the parents who interact with child welfare. I then assess the three major ways child welfare could affect parents: through investigations, opening a child welfare case, or removing a child. My two key findings are: First, child welfare investigations occur at pivotal moments in parents' lives. Second, opening a child welfare case causes meaningful changes in parents' lives by leading them to seek help for what social workers perceive to be addictions and mental illnesses.

My first key finding is on the circumstances of child welfare investigations. For background, parents in families investigated by child welfare are disproportionately racial and ethnic minorities and economically disadvantaged. Over 40% of fathers have spent some time in the local jail before the family was referred. Over one-third of mothers were teens at first birth, and three quarters of mothers had received food stamps at some point before the referral. Despite such prevalence, for both mothers and fathers, the child welfare referral and start of the investigation coincide with substantial increases in their reliance on the social safety net, their interactions with the criminal justice system, and their use of

mental health and substance abuse services. Immediately after a child welfare referral, the share of mothers using mental health services goes up by 26% for mental health services, by 85% for substance abuse services, and by 24% for housing assistance. Mothers are also twice as likely to be in jail. These changes could be caused by the child welfare referral or the incident that triggered it, but they provide a suggestive upper bound on the average effects of a child welfare investigation. The overlap between child welfare, the criminal justice system, and the safety net is particularly noteworthy. These results emphasize how pivotal are months in which families are investigated by child welfare.

I develop a new statistical method to identify the causal effects on parents of opening a child welfare case. This method estimates causal effects for families on the margin of having a child welfare case opened. Each family is investigated by a so-called investigator. Based on their personal judgment, these investigators have varying propensities to open a child welfare case. Previous literature had relied on random assignment of families to investigators and leveraged investigator propensities as an instrumental variable. In my setting, random assignment of investigators is imperfect, so propensity instruments alone cannot provide identification. My solution is to leverage panel data to control for preexisting differences of families assigned to different investigators. Past work addresses similar concerns of imperfect random assignment when random assignment is conditional on a linear function of observables (examples include Maestas et al. 2013 and Dobbie et al. 2018). My method instead assumes that families assigned to workers of varying propensities are on parallel trends in the outcomes of interest. As long as the first stage, monotonicity, the exclusion restriction and parallel pre-trends hold, my proposed methodology provides identification for local average treatment effects in settings with panel data where propensity instruments were previously not usable because of unobserved discretion in the assignments. In settings where perfect random assignment holds, this method documents the dynamics of the effects and may increase statistical power. This methodological contribution relies on some of the recent advances in econometrics on two-way fixed effects models (Borusyak & Jaravel, 2018; Goodman-Bacon, 2018; de Chaisemartin & D'Haultfœuille, 2020) to extend the

seminal work by Kling (2006) using propensity instruments in settings with panel data.

Using this identification strategy, my second key finding is that opening a case with child welfare leads mothers to increase their participation in substance abuse and mental health services. When assigned to an investigator who is one standard deviation more likely to open a child welfare case, mothers are 4.8% more likely to use mental health services and 7.6% more likely to use substance abuse services. These effects for mothers on the margin of having a child welfare case opened subsist regardless of whether a child was removed. Furthermore, some of these effects are persistent. Mothers assigned to an investigator with above median propensity to open a case are 1 percentage point (8.7%) more likely to be receiving substance abuse services up to six years after a child welfare referral. In contrast to mothers, I find only modest causal effects of increased mental health services for fathers in the short run, which could be because many of these fathers are no longer living in the same household as their biological children. These results suggest that child welfare is able to increase parents' participation in support services to address social worker concerns relating to parents' mental health and addictions. It is important to note, however, that child welfare interactions could also be increasing the need for mental health services.

I do not find statistically significant effects of having a child welfare case opened on enrollment in welfare programs nor on interactions with the criminal justice system, though some of my estimates are imprecise and cannot rule out meaningful effects. However, removing a child increases the likelihood of having a case opened in court and reduces the likelihood of receiving family assistance (TANF), at least for mothers in the short run. Both of these are effects that one would expect given that court approval is required to remove a child and that removing a child may mean the parent is no longer eligible for family assistance.

My work contributes in three ways to the growing economics literature on child welfare (Doyle, 2007, 2008, 2013; Warburton et al., 2014; Roberts, 2019; Gross, 2019; Chyn et al., 2019). First and foremost, prior literature has focused exclusively on child outcomes, so I am the first to study outcomes for parents. The most related work comes from Gross

(2019), who finds in his setting that parents are less likely to be listed again as the main perpetrator of child abuse and neglect after having been assigned to an investigator more likely to remove a child. Second, this literature has focused almost exclusively on the effects of child removals, whereas I examine the effects of investigations, opening a child welfare case and removals. Studying the range of child welfare interventions is important given that removals in fact have become more of the exception than the rule: out of all the families investigated in my sample only 10% will have a child removed, and only 43% will have a child welfare case opened. Gross (2019) does not find statistically significant effects of providing community services and targeted services to parents on child school attendance and performance. Relative to this, my results point out that opening a child welfare case leads to substantial changes for parents, even without removing children. My event-study analyses also provide a suggestive upper bound for the immediate impacts of a child welfare investigation. Third, granular data allows me to be the first to examine immediate and temporary effects of child welfare at the month level, and to study effects on an extensive range of social services such as food stamps and TANF, which has not been studied in prior literature.

Finally, some of my findings contribute to the topics of the opioid epidemic and gender disparities. Past research indicates that the opioid epidemic has had strong repercussions on the child welfare system (Prindle et al., 2018; Quast et al., 2018; Radcliff et al., 2018; Buckles et al., 2020). My results indicate that, keenly aware of this drug crisis, child welfare agencies have successfully encouraged parents to enroll in substance abuse services. As such, it complements prior economics literature on the opioid epidemic which has highlighted how supply-side and physician aimed policies can reduce the use of opioid prescriptions (Schnell & Currie, 2018; Alpert et al., 2018, 2019). With regards to the literature on gender disparities (Jolly et al., 2014; Kleven et al., 2019), my results document gender differences in child-rearing since in contrast to mothers, fathers are less responsive to child welfare investigations, if at all. Mothers may bear more of the burden of child-rearing and interacting with child welfare, but as a result, child welfare affects mostly mothers' access to services.

This paper will proceed as follows. In the first section, I provide background information on the child welfare system and my data source, followed by descriptive statistics of my sample. The second section covers the methodology for this paper and presents results from a simple event-study to motivate my methodology. Finally, the third section presents the main results on the causal effects of child welfare, and the fourth section extends these results with robustness tests and complementary analyses.

1.2 Child Welfare in Allegheny County, PA

Child welfare intervenes in cases of alleged child abuse and neglect by parents or primary caregivers. The best known intervention of child welfare is foster care which entails removal of the child from their caregivers' care to place them with kin, in a licensed foster home, or in congregate care.

However, since the early 1990's child welfare has evolved toward a "family-centered practice" which provides preventative services and aims to keep families together whenever possible.¹ Examples of preventative services include mental health and substance abuse treatment services, and in-home services to improve parenting skills.

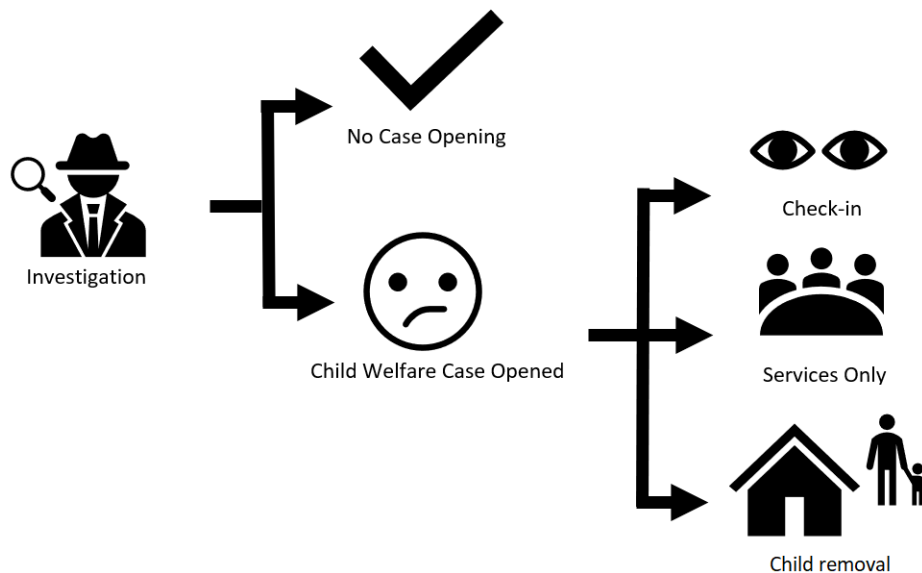
1.2.1 The Sequence of Events of a Child Welfare Referral

Individuals – such as teachers or doctors – who suspect a child of being abused or neglected refer the family to child welfare. Among these referrals, approximately half are serious enough to warrant an investigation. Investigations typically start within a few hours or a few days of the referral. In Allegheny county, based on the home address of the family and details of the call, the referrals to be investigated are sent to either the main office or one of the five regional offices. Within each of these offices, referrals are supposedly allocated to investigators in a round-robin fashion.

¹As Marc Cherna the director of the Department of Human Services of Allegheny county puts it: "I have learned that the vast majority of parents love their kids and want to do what's right for them. Problems and stresses of mental illness, poverty, and addictive diseases can get in their way, but they are not bad people" (interview with Casey Family Programs on April 8, 2019).

Investigators visit these referred families to determine whether any allegations of child abuse and neglect are indicated or substantiated. Based on the investigator's findings, they then decide whether a child welfare case should be opened for that family, as shown in the investigation flowchart of Figure 1.1. In my setting, as seen in Table 1.1, among investigated referrals 43% have a case opening.

Figure 1.1: *Child Welfare Flow Chart*



Notes: When child welfare investigates a family, a worker called an investigator visits the family and gathers the evidence to assess the family's strengths and weaknesses. Based on the findings, this investigator then decides whether there are reasons to be concerned for the child's safety and well-being at home. If there are no particular concerns, then there will be no further interaction of the family with child welfare. If however there are concerns, the investigator may believe this warrants opening a case with child welfare. A child welfare case can be opened with three purposes in mind. Child welfare may just want to check-in with the family and make sure things are going well, or they want to provide services to the family. Finally, if child safety is an immediate concern the child may be removed at least temporarily from the parents' care with court approval.

A case can be opened with several purposes in mind. The least severe reason to open a case is that the allegations are not currently of concern but that they may become so. The case is thus opened just so that a social worker checks-in with the family. The second reason to open a case is that the family could benefit from services to ensure safety of the children at home. Depending on the circumstances, participation in such services could be proposed, recommended, or even mandated by court. Finally, in cases where the child's safety cannot

be ascertained at home, child welfare may remove the child from the parents' care at least temporarily. In cases where the child is in imminent danger, the investigator may themselves remove the child. In most cases however, the decision of removing a child is made not by the investigator but, with court approval, by another social worker called the caseworker who is assigned to the open case. That being said, the results of the investigation are crucial in the caseworker's decision of removing the child. Thus investigators influence directly and indirectly whether or not a child is removed from his parents' care.

13% of families have a child removed within 1 year of the referral (Table 1.1). Most removals end with a family reunification once the parent has made the necessary changes to provide a safe home environment for the child. By federal legislation (Adoption and Safe Families Act, 1997), a permanency plan has to be made for any child that has been removed for at least 15 out of the past 22 months according to federal legislation. If the child is not reunified with his parents within this time frame, then this permanency plan may involve guardianship by kin or termination of parental rights in the hope of a future adoption.

1.2.2 The Data and Sample

I received administrative data from the child welfare agency of Allegheny county in Pennsylvania (Pittsburgh). I focus on referrals that did not *initially* list abuse as an allegation reason, because allegations reporting sexual and physical abuse are partly expunged prior to 2017 to protect parent rights as mandated by law. Cases of child neglect constitute the vast majority of referrals made to child welfare: 75% of 2018 child welfare referrals made in Allegheny county. For cases of neglect, the data cover *all* calls to child welfare in the county starting May 2009 until December 2018.

A unique feature of my data, is that it includes information on the parents of the child. Only 1.5% of calls do not report any parent (Appendix Table A.1). 94% of referrals list a mother, and 87% of referrals list at least one father. Few child welfare agencies have such comprehensive records on both mothers and fathers. This exceptional dataset allows me to be the first to assess the impacts of child welfare on parents.

I focus on the sample of referrals who are assigned quasi-randomly to one investigator. Referrals for families that were investigated in the 90 days prior can be reassigned to their last investigator by default so I exclude them. I also exclude a fifth of referrals where the agency did not record who the investigator was or when more than one investigator was recorded (Appendix Table A.1). The latter may be because the original investigator went on leave, was promoted, left the job, etc.

For parental outcomes, my data include services received month by month from 2008 to 2018 for 97% of these parents. This includes whether these parents had received food stamps, TANF or starting in October 2013 housing assistance.² I also received health information on whether these parents were on Medicaid, and whether they were receiving mental health services or substance use disorder services paid by the county or by Medicaid managed care. Finally, I received data on whether these parents had any cases being processed by the Allegheny courts and for adults whether they were booked in the Allegheny county jail. One key limitation of this data on services is that I only have information for the above-mentioned services received *within* Allegheny county. If a family is no longer receiving services in my data, it could be that they are living in Allegheny County and are no longer receiving services, but it could also be that they have moved out of Allegheny County and are continuing to receive services. Some of my later robustness checks will address this concern.

This leaves me with a sample of approximately 28,000 referrals for child neglect with 40,500 unique parents, a third of which are listed more than once on a referral. Panel B in Table 1.1 shows the most frequently listed reasons for the referral for the calls that were investigated. In a third of cases, parents are struggling and having a hard time coping with raising a child. In 21% of cases the allegation is that the parent is struggling with drug abuse, in 18% that the parent is failing to protect the child, and in 17% that the child has a behavioral problem. It is clear from looking at these, that improving parents' behavior and wellbeing is key to a successful child welfare intervention in the sample of cases I focus on. Thus, although the goal of child welfare is to keep children safe, child welfare's impacts on

²Housing assistance includes a wide range of housing services going from Section 8 vouchers to shelter.

parents are essential in that process.

Table 1.1: *Descriptive Statistics on Sample*

<i>Panel A: Investigated Child Welfare Referrals</i>		
Opened a Child Welfare Case (%)	43	
Child Removed within 3 months (%)	10	
Child Removed within 1 year (%)	13	
Number of Referrals	28166	
<i>Panel B: Most Frequent Allegations Listed on Referrals</i>		
Parent Struggling (%)	32	
Parent Drug Abuse (%)	21	
Failure to Protect Child (%)	18	
Child Behavioral Problem (%)	17	
Inadequate Housing (%)	14	
<i>Panel C: Parent Characteristics for...</i>		
	<i>Moms</i>	<i>Dads</i>
Age	34.0	36.6
	(9.3)	(9.5)
White (%)	47	40
Race Missing (%)	17	8
Teen parent (%)	31	16
Unknown if Teen Parent (%)	26	23
Food Stamps, 3m Prior to Referral (%)	44	19
Medicaid, 3m Prior to Referral (%)	62	23
In Jail, 3m Prior to Referral (%)	1	5
Observations	26912	35274
Means - Standard Deviations in parentheses.		

In Panel C of Table 1.1, I present some descriptive statistics on this sample of parents who are interacting with child welfare. A few patterns are striking. First, this sample is disproportionately minorities. Approximately 80% of Allegheny county residents are white according to Census data whereas in my sample, depending on assumptions made with regards to parents with missing race, between 40 to 48% of fathers are white and between 47 to 64% of mothers are white. Second, at least a third of mothers had their first child as a teen mother: 31% of mothers were teen mothers, 43% were not, and for 26% teen parenthood was unknown. The actual percent of teen mothers is thus somewhere between 31 and 57%.

Third the differences between mothers and fathers are striking: three months prior to the referral mothers are more than twice as likely to be on food stamps, three times more likely to be on Medicaid and five times less likely to have been in jail. Finally, mothers seem to be highly reliant on social services given that almost two-thirds are receiving medical assistance (62%) and that nearly half (44%) are receiving food stamps. Fathers on the other hand interact substantially with the justice system since at one point in time just 3 months prior to the referral 5% of them were in Allegheny county jail. In fact, if I look at all months prior to the referral for fathers, I find that 42% of them have spent some time in Allegheny county jail in my decade of data. A notable feature of my data is also that I have a lot more fathers listed than mothers. Although I do not have any variables on family structure, this is likely due to many of these homes being non-traditional households: mothers may have had children from multiple partners and are not in a stable long-term relationship. Parents interacting with child welfare thus seem to be a particularly vulnerable sample of the US adult population.

1.3 Methodology: Propensity Instruments with Panel Data

The causal impacts of child welfare cannot be estimated by comparing families involved with child welfare to those without, since these families differ for many other reasons. I propose an identification strategy which estimates the causal effects for investigated families on the margin of having a child welfare case opened. This strategy relies on combining two well-known identification strategies: propensity instruments and event-studies.

1.3.1 The Propensity Instrument: Investigator Assignments

Prior work in child welfare relied on random assignment of cases to child welfare investigators (Doyle, 2007, 2008, 2013; Roberts, 2019; Gross, 2019; Chyn et al., 2019). Investigators may disagree on how severe they perceive a given referral to be; some investigators are more stringent than others in their investigations. The idea is thus to leverage the fact that some families were randomly assigned to stringent investigators and to compare them to families

that were assigned to investigators who were less stringent. Doing so provides causal evidence on the effects of child welfare for families on the margin of receiving child welfare services, under assumptions of random assignment, instrument relevance, monotonicity, and exclusion restriction.

Given that using investigator fixed effects leads to the problem of many weak instruments, the empirical literature's strategy has been to construct a measure of stringency called a propensity instrument (Kling, 2006; Doyle, 2007; Maestas et al., 2013; Dobbie et al., 2018). These propensities are typically constructed leaving that particular individual's referral out as a reduced-form version of the jackknife instrumental variables estimation which removes the dependence of the constructed instrument on that individual's outcomes (Angrist et al., 1999). These are also called leave-one-out instrumental variables.

Investigators' primarily influence whether a case with child welfare services should be opened. However, investigators also influence both directly and indirectly whether a child is removed (cf section 1.2.1). As a result, in my setting I construct three propensity instruments: one for case opening, and then two instruments that split this out by what follows the case opening. Case openings can lead to a removal within 3 months, or they can lead to simply a case opening where a social worker checks in with that family and provides access to services. I thus construct a propensity to remove and a propensity to open a case without a removal. These two propensities mechanically sum up to the propensity to open a case.

Looking only at the removal propensity of investigators, as the prior literature in child welfare has done, would lead to violations of the exclusion restriction given that investigators' propensity to remove is positively correlated with the propensity to open a child welfare case without removing a child.³ In that light, my work builds on the work by Mueller-Smith

³This may be less of a concern in earlier decades like the 1990's where most child welfare cases involved the removal of a child, but is definitely no longer true in my sample since three quarters of cases that are opened by child welfare do not involve a removal within the first few months. In my data, investigators that are more likely to open a case with removal seem also more likely to open a case without removal. The correlation coefficient is statistically significant but small (0.05-0.15 depending on specification), so this violation of the exclusion restriction seems unlikely to be a major concern.

(2015) but using propensity instruments instead of fixed effects similar to some of the analyses in Gross (2019).

I create the propensity instrument at the referral level since having a child removed is common to all parents in a household. Taking the example of removals, for parent i listed on referral r who was assigned to investigator j working in office o , the propensity to remove is constructed in the following way:

$$PR_{r,j,o} = \frac{\sum_{f \neq r} R_{f,j,o}}{N_{j,o} - 1}$$

where $N_{j,o}$ is the number of referrals investigated by investigator j in office o and $R_{f,j,o}$ is a dummy variable indicating whether any child was removed within three months of referral, for referral f assigned to investigator j in office o . I create similarly an instrument for the propensity to open a case $PO_{r,j,o}$ and for the propensity to open a case without any removal $PNR_{r,j,o}$.

Some studies residualize the propensity instrument on the characteristics of the referral (Maestas et al., 2013; Dobbie et al., 2018). In my setting, cases are only randomly assigned within a given office, so following Dahl et al. (2014) I residualize these constructed propensities PO , PR and PNR on a set of office-year fixed effects similarly to Gross (2019) who residualizes on a set of zipcode-year fixed effects. I do not residualize on additional characteristics of the referrals so as to be able to assess whether cases were in fact truly randomly assigned within an office-year. The residuals PO^* , PR^* and PNR^* from running the regressions of my propensity instruments on the office-year fixed effects then measure how much more/less likely the investigator was to remove or open a case relative to other referrals in that office-year. Residualizing the instruments on office-year fixed effects leaves substantial variation across investigators (Appendix Figure A.1).

I focus on 3-month removals instead of removals occurring within 1-year of the referral because investigations are required by state law to close within sixty days of the referral. By 3 months, 87% of investigations have been closed. I do run robustness checks on using removals within 1 year of the referral by which time 99% of investigations have been closed.

As reported in Table 1.1, 10% of referrals will already have a child removed within 3 months. An additional 3% of referrals will have a child removed within 1-year, but most likely that is due to the assigned caseworker later realizing that there is a concern after the case has been opened.

1.3.2 Event-Study Analysis

One of the unique features of my data is that I have access to services being received even prior to the child welfare referral and investigation. This allows me to leverage the use of panel data for identification on the topic of child welfare.

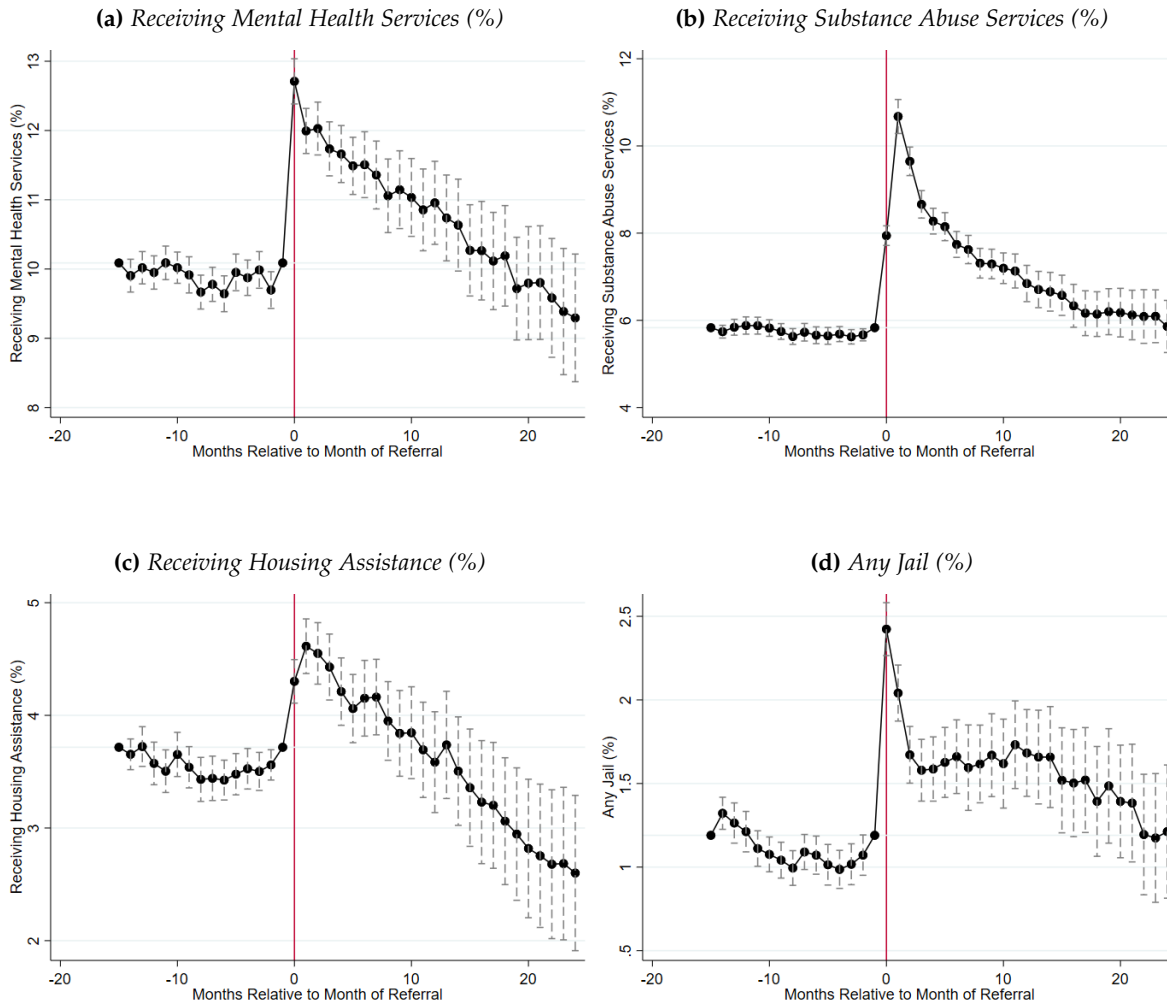
Event-study analyses estimate the effect of an event – in my context a child welfare referral and investigation⁴ – by comparing individuals who are referred at different times. These studies rely on the inclusion of both time and individual fixed effects. They also require the assumption of parallel trends prior to the referral, which is typically tested with an F-test.

In my context however, the effects of a child welfare investigation cannot be estimated using an event-study. Referrals typically happen in times of crisis for a family. It would thus be impossible to disentangle the effects of the child welfare referral from those of the event that triggered the child welfare referral. For example, if a single mother is sentenced to jail, her children would be referred to child welfare for a new home in the meantime. Jail is what triggered the child welfare referral, not the reverse. As can be seen in Figure 1.2d, mothers are two times more likely to be in jail in the month of a child welfare referral.

It is striking to see how the timing of investigated child welfare referrals coincides with sizable jumps in mothers' use of services. As shown in Figure 1.2, in the months immediately after a referral, mothers increase their use of mental health services by up to 26% (2.7 percentage points), of substance abuse by up to 83% (5 percentage points) and of housing services by up to 24% (0.9 percentage points). Note that here again these effects

⁴Note that after a child welfare referral is received, if there is an investigation it usually starts soon after the referral so the effects of the referral cannot be distinguished from those of the investigation in my sample of investigated referrals.

Figure 1.2: Event-Study Plots Relative to the Month of the Referral



Notes: Graphs represent the estimated coefficients from an event-study analysis with 90% confidence intervals including two-way clustering for investigators and parents. Fixed effects are included for each mother on a referral, for time, and – to be consistent with later analyses – for referral office-year fixed effects. The sample is mothers listed on investigated referrals and who were assigned to an investigator that investigated at least ten referrals. Reference points are the fifteenth month before the referral and the last month before the referral (see Borusyak & Jaravel 2018 for an at length explanation of why I use two time periods as reference points). The referral to child welfare and start of investigation occurs during month 0.

could originate from the incident that triggered the referral, which could even be seeing a professional to receive these services.

These descriptive analyses highlight the importance of the events around an investigated child welfare referral especially for mothers. They also provide an upper bound on the

immediate causal impacts of an investigation.

1.3.3 Combining Propensity Instruments with an Event-Study

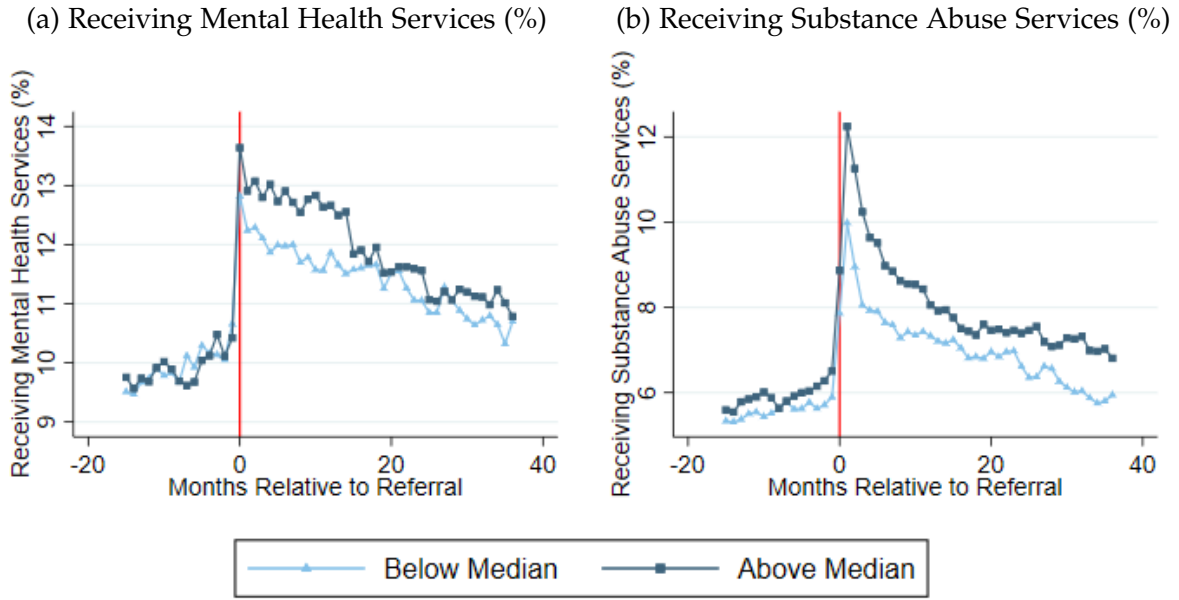
To estimate causal effects of opening a case with child welfare, I leverage the combination of propensity instruments and event-studies. This identification method's advantage is that, under the assumption of parallel trends, it can provide identification even when propensity instruments are not valid on their own.

Indeed, investigators are imperfectly randomly assigned to referrals as will be discussed at greater length in Section 1.3.5. However, parents assigned to investigators of varying propensity are on parallel trends prior to the referral as will be shown in Section 1.3.5. As a result, I can back out the effects of investigators with varying propensities by comparing the outcomes for parents assigned to different investigators after the referral, relative to the observed differences before the referral.

To better understand this, suppose that parents are assigned to two types of investigators: investigators with a propensity to open a case above the median, and those with a propensity below the median. Then, as long as the parallel trends assumption holds for parents assigned to these two groups, I can estimate the effects of being assigned an investigator with above median propensity to open a case by comparing outcomes for parents assigned to investigators in those two groups after the referral relative to how they compared before the referral. Figure 1.3 illustrates this by plotting the mean use of substance abuse and mental health services for mothers assigned to investigators above and below the median propensity to open a case. A visual inspection seems to suggest that indeed the pre-trends are similar for both groups, but that mothers assigned to an investigator more likely to open a case increase their use of mental health and substance abuse services after the referral.

Like in any difference-in-difference, this method relies on the assumption that had parents not been assigned to different investigators they would have continued on parallel trends. This could be violated if for example investigators with higher propensity were assigned to more severe incidents and that more severe incidents led to different outcomes

Figure 1.3: Means of Mothers' Outcomes by Investigator Median Propensity to Open a Case



Notes: Figures present the percent of mothers receiving mental health services (graph on the left) and substance abuse services (graph on the right), for each month relative to the month the child welfare referral occurred ($t = 0$) separately for mothers assigned to an investigator with a propensity to open a case above and below the median propensity to open a case.

for parents. In that case I would be wrongly attributing changes in parents outcomes to child welfare. As will be discussed later however, investigators highlight that it is difficult to predict severity of a case before they visit the family and I also run robustness checks to address this concern.

This setting with a binary propensity instrument (above/below median) can be generalized to the continuous version of the propensity, by looking at the coefficient on the continuous propensity in each time period. The idea is to apply event-study methods to the coefficients of the propensity interacted with the event-study dummies, as formally presented in the next section. Results are however similar if I use an above median propensity indicator variable throughout.

1.3.4 Empirical Specifications

My simplest specification is a pooled specification which I use to produce the results in table format. I compare parent outcomes Y several months after the referral and before the referral, for parents assigned to investigators with varying propensity PO^* to open a case. More specifically, for parent i listed on a referral r that is referred in month-year t_r , the date of the referral, I estimate:

$$Y_{i,r,j,t} = \beta_0 + \beta_1 \mathbb{1}_{(t \geq t_r)} + \beta_2 \mathbb{1}_{(t \geq t_r)} \cdot PO_{r,j}^* + \delta_{o,y_r} + \mu_{i,r} + \lambda_t + \epsilon_{i,r,j,t} \quad (1.1)$$

where β_1 estimates how the outcomes of parents are different after the referral ($t \geq t_r$), and β_2 estimates the effects after the referral of being assigned to an investigator with a one standard deviation higher propensity to open a case relative to the average investigator in that office year. Indeed, investigators cannot impact families prior to the child welfare referral when they are assigned to families.

I do not include the main effect for the propensity variable because I include individual-referral fixed effects $\mu_{i,r}$ which are referral-specific since parents can be listed on a child welfare referral more than once in 2009-2018. As when I constructed the residualized propensity instruments, I include office-year of referral fixed effects (δ_{o,y_r}). I also include fixed effects for the month-year at which the outcome is measured (λ_t) to control for time trends.

A growing econometrics literature on two-way fixed effects however has pointed out serious concerns with two-way fixed effects as in equation (1.1) (for example Borusyak & Jaravel, 2018; Goodman-Bacon, 2018; de Chaisemartin & D'Haultfœuille, 2020). Notably, negative weights are common in two-way fixed effects regression. A pooled specification with two-way fixed effects like the above generally overweights short run effects, and long run effects will be downweighted or even possibly have negative weights (Borusyak & Jaravel, 2018). The recommendation has been to look at effects dynamically, which helps address heterogeneity in treatment effects over time. This is the motivation for running these regressions in an event-study like specification. Using the same notation as previously,

I thus run regressions in the form:

$$Y_{i,r,j,t} = c + \sum_{d=-n}^m \left[\theta_d \mathbb{1}_{(t=t_r+d)} + \phi_d \mathbb{1}_{(t=t_r+d)} \cdot PO_{r,j}^* \right] + \delta_{o,t_r} + \mu_{i,r} + \lambda_t + \epsilon_{i,r,j,t} \quad (1.2)$$

which correspond to the equation (1.1) except instead of pooling all the months together, I look at effects month by month. I include the same set of fixed effects as previously.

Furthermore, equation (1.2) assumes homogeneity of the effect of being assigned an investigator one standard deviation more likely to open a case d months after the referral. It assumes that ϕ_d is the same for all individuals i and in all years t . As shown by Borusyak & Jaravel (2018), such OLS specifications are not robust to treatment heterogeneity. They recommend instead a semi-parametric specification which is robust to these concerns. This is my preferred specification for the event-study graphical results estimating the ϕ_d 's. After checking for pretrends, this specification sets $\phi_d = 0$ for $d < 0$, and then estimates ϕ_d for $d \geq 0$ by running equation (1.2) restricting to the observations with $t = t_r + d$ or $t < t_r$ for each $d \in \{0, 1, \dots, m\}$, where t_r is the month-year of the referral.

In addition to running equations (1.1) and (1.2) at the month level, I also run them at the year level to evaluate long run effects. I use years relative to the referral month. So the year 0 is the month of the referral and eleven months after that. Year 1 is months 13-24 and so on.

Finally, as mentioned at the beginning of this section, it is interesting to distinguish the different interventions that child welfare has in a family. A case opening can lead to a removal of a child from the home within the first few months or not. These are exclusive outcomes. I thus split out the overall propensity to open a case in two and estimate separate coefficients for the propensity to open a case with removing a child within 3 months (PR^*), and the propensity to open a case without removing a child from the home at least initially (PNR^*). This can be seen as the reduced form version of multiple investigator fixed effects influencing two margins of impact, similar in spirit to Mueller-Smith (2015) where judges influence incarceration, probation length, fines, etc. For example, in the case of the pooled month specification:

$$Y_{i,r,j,t} = \gamma_0 + \gamma_1 \mathbb{1}_{(t \geq t_r)} + \gamma_2 \mathbb{1}_{(t \geq t_r)} \cdot PR_{r,j}^* + \gamma_3 \mathbb{1}_{(t \geq t_r)} \cdot PNR_{r,j}^* + \delta_{o,y_r} + \mu_{i,r} + \lambda_t + \epsilon_{i,r,j,t} \quad (1.3)$$

The coefficient γ_2 measures the effects of being assigned to an investigator more likely to remove children holding constant this investigator's propensity to open a case without a removal, and the coefficient γ_3 measures the effects of being assigned to an investigator who opens more cases without removing more children. For example, an investigator might be particularly effective at picking up on severe forms of child abuse and neglect which would warrant a removal, which sometimes families try to hide, and so they may have a high PR^* . An investigator may on the other hand perceive that more families could benefit from services or social worker visits, and so be more keen on opening a case without removing a child. In that case they would have a higher PNR^* . Hence, γ_2 can be interpreted as the marginal effect of being assigned to an investigator who opens more cases with removals, and γ_3 as the marginal effects of being assigned to an investigator who opens more cases without a child removal. This specification in equation (1.3) thus allows me to explore the mechanisms underlying the overall effect of opening more child welfare cases estimated in equation (1.1).

1.3.5 Validity of the Identification Strategy

Conditional Random Assignment

A typical propensity instrument relies on random assignment of referrals to investigators. To test this, I include all demographic characteristics listed on referrals, as well as what type of reporter was calling to report child abuse and neglect, receipt of services three months prior to the referral and receipt of services within the year prior to referral. I do not include the listed allegations because investigators can add allegations and so they are not a valid test of balance since investigators modify them.

Propensity measures are especially noisy when constructed on few referrals per investigator. This is especially true for outcomes that are less frequent like removals. Following the

previous literature (Doyle, 2007; Chyn et al., 2019), I restrict my sample to referrals assigned to investigators that investigated at least ten referrals throughout the paper. That leaves me with a sample of 460 investigators. For investigators with at least 10 referrals, the median number of investigations is 29 and the mean is 56.

I pool together mothers and fathers to maximize power in my balance test. Given some of the concerns raised by Hull (2017) on the correction for degrees of freedom in propensity regressions for the first stage, I bootstrap cluster the standard errors for the balance tests. I run 800 bootstrap clusters, which is above the recommendation by Cameron & Miller (2015).

I present full results in the Appendix table A.2 because there are thirty-five variables included in my balance test. I report the p-values of the chi-squared tests that all the coefficients are zero in the next to last row of Table 1.3.

The results of the chi-squared test are consistent with random assignment for the propensity to remove ($p\text{-value} > 0.10$). However, results are not consistent with random assignment in the case of the overall case opening propensity ($p\text{-value} < 0.05$). Since case opening is the sum of the propensity to remove and propensity to open without removal, as expected, the propensity to open a case but not remove is also inconsistent with random assignment. Investigators with a higher propensity to open a case without removal are assigned fewer white parents. The variable for missing race is picking up on the same racial imbalance. Investigators more likely to open a case are also assigned more to referrals reported by police and social workers. Finally, they may even be assigned to parents with less past year jail, less past open cases and more past re-referrals recently. This is a concern.

Although the agency has repeatedly mentioned that they assign referrals quasi-randomly, this type of issue is not unheard of in the child welfare literature. In Rhode Island, where Chyn et al. (2019) had access to a dummy variable indicating which cases were not assigned quasi-randomly, even deleting such cases, the balance tests were rejected for older children. My interactions with the South Carolina child welfare agency indicated that child welfare offices varied in how they assigned investigators to cases, some randomly and others less so, whereas Roberts (2019) relied on random assignment in all offices.

That being said, the magnitudes for the differences in all my balance tests are quite small: not more than 0.1 of a standard deviation in the propensity. For example, being white is associated with assignments to investigators with propensity to open a case lower by a tenth of a standard deviation (column 1, Appendix Table A.2).

Furthermore, investigators say that until they walk through that door, it is difficult to know what they will find and its severity. Individuals who call in to report child abuse and neglect usually have limited knowledge of what is actually happening in the home. For a subset of the data, I have an assessment of the risk based on the details of the referral and before the start of the investigation. Using this measure, I indeed find that 60% of homes where a child was removed were not flagged as high risk confirming how difficult it can be to predict severity of a case based solely on information from the referral call. Even if referrals were to be assigned non-randomly as seems to be the case, it would thus be hard to do so in a manner that would be meaningful. In my view, this is the most encouraging feature of the setting for using investigator assignments. For example, investigators can walk into a home with alleged housing concerns and find that instead the cause of concern is parental drug use and that this is a concern not for the child that was reported but for a younger sibling that the caller did not know about. Assignment of cases is unlikely to be done in a meaningful manner.

Some of the robustness checks will further address this imperfect random assignment, but this constitutes the motivation for running these analyses using the proposed methodology presented in section 1.3.4.

Testing for Pre-trends

A key requirement for any event study is the parallel trends assumptions. The F-statistic does this by testing that all the coefficients prior to the event are equal to zero. I present the p-values of an F-test for each of my outcomes and instruments in Table 1.2. Given that I look at both the short run and long run effects on parents, I run F-tests on the pre-trends both in the months leading up to the referral and on the years prior to a referral to respectively

validate the identification in the short and long run.

Out of the 96 tests, approximately 15 are rejected at the 10% significance level which is slightly more than expected though still in the range of what one could expect due to pure chance.⁵ Overall, these F-tests thus provide modest support for the parallel trends assumption. The F-tests provide support for the parallel trends assumption except for housing for mothers at the month level (column 8 Panel A) for SNAP at the year level for mothers (column 7, Panel B), for TANF for fathers at the month level (column 6 Panel C), and for courts and jail for fathers at the year level (columns 4 and 5 in Panel D). None of these relate to the key results presented later on. Parallel pre-trends are however also rejected for the propensity to open a case without a removal for substance abuse: p-value of 0.09 looking at the 15 months prior, and pvalue of 0.03 looking at the years prior. For this reason, I always report the pre-trends for my graphical results, especially for substance abuse results, so that the reader can assess these pre-trends for themselves. I inspected each of the violations of the parallel trends assumptions visually but no clear pattern emerged. Most of the rejections were due to the sample being imbalanced for a year, instead of it being consistently imbalanced across years. The one noteworthy exception is substance abuse at the year level for the propensity to open a case without removal. There seems to be a pre-trend upwards in use of substance abuse services, which drops in the two years prior to the referral, before spiking up again after the referral (results not shown). This violation of the assumption of parallel trends puts into doubt my later year-level results finding no effect on substance abuse of being assigned an investigator more likely to open a case without a removal.

That being said, overall, the results from these F-tests tests suggest that my proposed method is able to identify causal effects of investigator assignments on parents.

⁵Note that this reasoning assumes that the tests are independent, which is not the case given that I am testing for the same outcomes at the month level and year level, and that the case opening propensity is a linear combination of the other two propensities.

Table 1.2: *P-values of Pre-trends F-test*

	Health			Justice		Social Services		
	(1) Medicaid	(2) MH	(3) SA	(4) Courts	(5) Jail	(6) TANF	(7) SNAP	(8) Housing
<i>Panel A: Mothers' Pretrends of 15 Months Prior</i>								
Case Opening Propensity	.28	.32	.16	.18	1	.57	.97	.13
Propensity to Open without a Removal	.46	.19	.09	.42	.98	.32	.67	.06
Propensity to Open with a Removal	.24	.86	.46	.11	.09	.3	.09	0
<i>Panel B: Mothers' Pretrends of 11 Years Prior</i>								
Case Opening Propensity	.51	.72	.13	.37	.96	.37	.06	.29
Propensity to Open without a Removal	.39	.65	.03	.16	.92	.67	.05	.96
Propensity to Open with a Removal	.4	.98	.55	.51	.68	.38	.45	.29
<i>Panel C: Fathers' Pretrends of 15 Months Prior</i>								
Case Opening Propensity	.81	.92	.98	.3	.96	.01	.49	.55
Propensity to Open without a Removal	.64	.63	.99	.25	.89	.09	.57	.77
Propensity to Open with a Removal	.47	.95	.95	.66	.41	.09	.51	.39
<i>Panel D: Fathers' Pretrends of 11 Months Prior</i>								
Case Opening Propensity	.87	.95	.5	.01	0	.42	.43	.88
Propensity to Open without a Removal	.81	.93	.8	.01	.01	.33	.61	1
Propensity to Open with a Removal	.14	.31	.12	.78	.93	.17	.34	.2

Notes: Each cell in this table is the p-value for an F-test testing for linear and non-linear pre-trends in the propensity listed on the left-most column for the outcome listed in the columns. Panel A tests are on the sample of mothers for the 15 months prior to the referral. Panel B tests the sample of mothers for all years available prior to the referral. Panel C and D report the same tests for the sample of fathers. Note that I have data on pre-trends for all parents in the 15 months prior to the referral (jail is only available for adults though). Given that I have referral up until end of 2018, I do have up to 11 years of data on trends prior to the referral but I do not have 11 years for every individual. Furthermore, since housing records are available only starting October 2013, I only have up to 4 years of information on housing pre-trends.

Instrument Relevance

For the propensity instruments to be meaningful, they need to be predictive of what they predict (case opening, removal, opening without removal). This is assessed in Table 1.3. Overall, the propensity to open a case and to open a case without removal are quite predictive (coefficients on the order of 0.7).

The coefficients for the propensity to remove are quite predictive too ($t\text{-stat} > 5$), but the coefficient is lower than might be expected: being assigned an investigator with a propensity to remove 1 percentage points higher is associated with a 0.3 increase in the probability of having a child removed for parents (column 2). Coefficients in these regressions are expected to be on the order of 1 if the propensity were perfectly measured. This coefficient for removals is lower likely because my propensity to remove is much noisier than the other propensity instruments since removal is a rarer outcome. Indeed when I restrict to investigators at least 80 investigations, the first stage coefficient for the removal propensity goes up to 0.50 (results not shown).

Monotonicity

Furthermore, for the propensity instrument to be valid, monotonicity needs to hold. Monotonicity ensures that the estimates are positively weighted averages of effects. If some investigators are more likely to open cases or remove children from certain types of families (such as for certain types of allegations), then monotonicity would be violated. This is a considerable assumption.

Unfortunately, there is no test that proves monotonicity. The literature has proposed three tests that would raise concern if they were violated, but they are not sufficient to prove monotonicity.

The first test involves checking that the coefficient of the first stage is positive for all subgroups of variables (Bhuller et al., 2020). I do this for all the variables included in the balance tests and for all the listed allegations too. I thus run a first stage regression for 88 subgroups in my data. All the coefficients are positive and significant, except when

Table 1.3: First Stage

	(1) Case Opening	(2) Open & No Removal	(3) Open & Removal
Case Opening Propensity	0.705*** [0.030]		
Propensity to Open without a Removal		0.701*** [0.034]	0.029* [0.015]
Propensity to Open with a Removal		0.247*** [0.053]	0.309*** [0.060]
Number of investigators	460	460	460
P-value of Balance Test	0.0220	0.0148	0.164
Observations	59450	59450	59450

Notes: p<0.1, ** p<0.05, *** p<0.01. Standard errors are cluster bootstrapped by investigator, with 800 bootstraps. Year and regional office fixed effects are included throughout. The sample is restricted to investigators overseeing at least ten referrals. This table presents the results of the first stage where the outcome for a given individual is regressed on the corresponding residualized leave-referral out propensity. In the next to last row, I also present the p-values of the chi-squared test from the balance test where I regress the propensity on thirty-five individual characteristics. Details for the balance test are reported in Appendix Table A.2.

sample sizes are small. A subset of these results are shown in Appendix Table A.3 for the propensity to open a case. This first test is thus consistent with monotonicity.

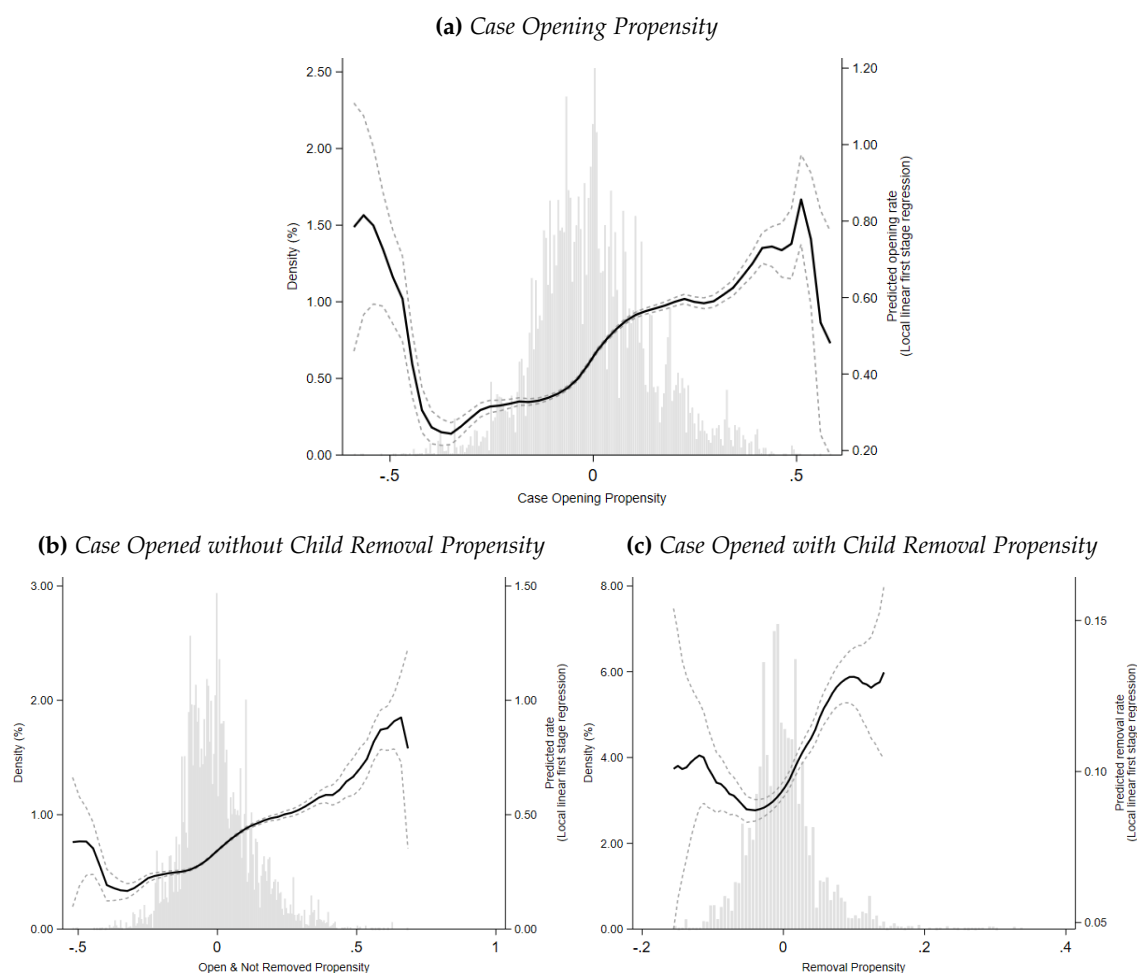
Secondly, if monotonicity holds, investigators more likely to get child welfare involved for certain types should be more likely to get involved for other types. This can be tested by constructing the propensity from a subset of cases and then looking at whether the first stage coefficient for the excluded set of cases is positive and statistically significant. I run this test for all demographic variables, reporter groups and lagged outcomes (42 subgroups). The first stage coefficients are positive and statistically significant for all subgroups. I report a subset of these results for the propensity to open a case in Appendix Table A.3.

Finally, monotonicity should hold at the individual level, but one would expect that the averages should reflect this. Hence, one would expect that investigators with higher propensities should be monotonically more likely to open a case/remove a child. Many individual characteristics may affect outcomes and each investigator only investigates a few families, so in my context this may not hold perfectly. The results for this second test for monotonicity can be assessed in Figure 1.4. I do not plot the local linear regression of the top 1% of the propensity to remove because it dramatically changes the y-axis and prevents a visual examination of the monotonicity assumption in the rest of the distribution. The plotted coefficients from the local linear regressions seem to be monotonically increasing, except at extreme values of the propensities. The bottom 5% of the propensity to open a case and the top 1%, and the top and bottom 1% of the propensities to remove and open a case without a removal seem like they could be a concern for monotonicity. Robustness checks later on will address these specific concerns. Overall, these graphs thus provide modest support for monotonicity.

Exclusion Restriction

As indicated in section 1.3.4, to leverage the specific timing of the referral in my identification strategy, I run my regressions in a reduced form. The exclusion restriction will affect how to interpret my results. If it holds, then my estimates can be interpreted as the causal effects

Figure 1.4: Monotonicity Test



Notes: Figures present the distribution of propensities on the left axis for cases assigned to an investigator with at least 10 investigations, and on the right axis the corresponding local linear predictions with 95% confidence intervals. I exclude the top 1% of the propensity to remove in the local linear predictions of panel (c) because it changes dramatically the scale and prevents visual examination of monotonicity for the rest of the distribution. These local linear predictions only include the corresponding propensity instrument as independent variable.

of being assigned an investigator more likely to open a case. Otherwise, my estimates are the effects of opening a case and any other ways that investigators can change outcomes for parents which are correlated with the propensity. It is thus important to ensure that investigators' only margin of impact is through opening a case.

Two points are worth mentioning. First, as alluded to previously, investigators decide which allegations are listed and substantiated. In fact, anecdotal evidence suggests that some

investigators may be better at picking up on certain types of allegations. Later, in robustness checks, I will present results that control for these margins of impact by constructing similarly propensities for investigators to substantiate different types of allegations.

Secondly, in the later part of my data some offices switched to a model where the investigator assigned to a given family became their caseworker if a case was opened. This is what the agency calls the one caseworker model. This only starts in late 2014. In 2016 40% of referrals are in this new system and by 2018 97% of referrals in my sample are in this new system. In the robustness checks I thus exclude this later part of my data to ensure that the effects are not driven by effects of investigators after the close of the investigation.

1.4 Investigators' impacts on parents

1.4.1 In the Months Following the Referral

The first step in examining the effects of interactions with child welfare investigators on parents is to look at what happens immediately following the referral. About half of cases that are opened are closed within one year of opening, and federal legislation requires a permanency plan within two years. Effects of child welfare interactions on parents should thus be particularly strong in the first year following a referral to child welfare.

Figure 1.5 confirms this for the sample of mothers assigned to investigators with varying propensities to open a child welfare case, in the two years before and after the referral. Figure 1.5a and Figure 1.5b show the pre-trends for mental health and substance abuse services, whereas Figures 1.5c and 1.5d show the preferred estimates which address the concerns of negative weighting and set the pre-trends to zero for efficiency. In the month of the referral (month zero), mental health services spikes up for mothers by approximately 0.5 percentage points when assigned to an investigator one standard deviation more likely to open a case. Use of mental health services remains elevated for approximately 1 year from the referral. Substance abuse take-up also goes up substantially after a referral by about 0.5 percentage points, though it spikes only two months after the referral. A delayed spike

could be due to spots for substance abuse treatment being sometimes hard to find. Similarly to mental health services though, within a year substance abuse services at the month level seem to have resumed their prior levels.

Results for all services being received in the months following the referral are presented in Table 1.4. These results pool together the twelve months immediately following the referral relative to the six months prior to the referral, using the specification explained in equation (1.1) for Panels A and B, and equation (1.3) for Panel C.

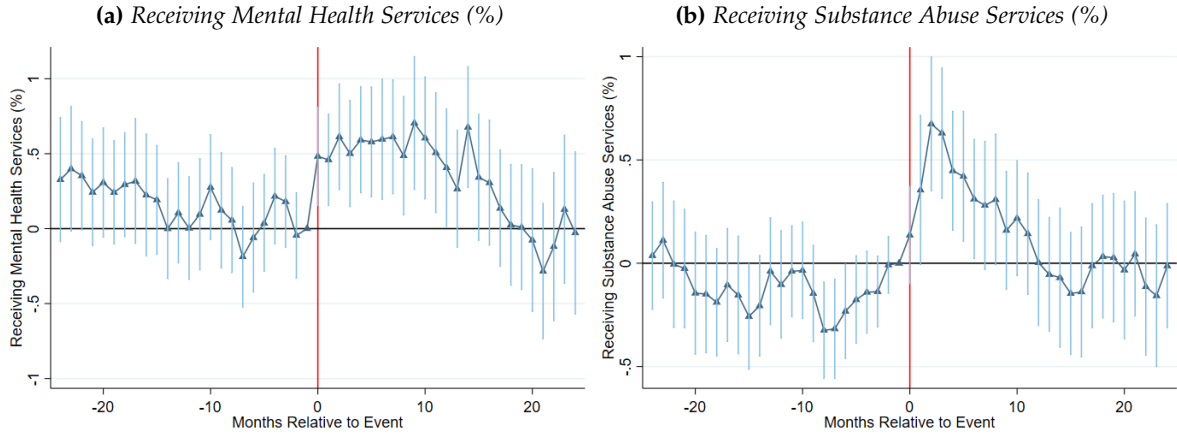
Panel A indeed shows that mothers who are assigned to an investigator that is one standard deviation more likely to open a child welfare case than average are 0.49percentage points more likely to be using mental health services in the 12 months following the referral. This corresponds to a 4.8% increase in mental health services relative to the year prior to the referral (means in the 12 months prior to the referral are reported at the bottom of the table). They are in addition 0.45pp more likely to be using substance abuse services, which represents a 7.6% increase in the use of such services. The percentage increases in use of these services are substantial. Finally, mothers assigned to an investigator one standard deviation more likely to open a case are 0.32pp less likely to be on TANF (1.8% decrease).

Overall these results present both positive and negative effects on mothers. The decrease in TANF receipt most likely comes from mothers losing eligibility for TANF when their child was removed.⁶ If that is indeed the case, then the TANF decrease might be considered a negative effect because it is an income reduction. However, the net effect on parental income is unclear, because these mothers are no longer facing the costs of raising the child that was removed. The increase in substance abuse services for its part is a positive impact on parents provided one believes that substance abuse services are helpful. How to interpret the results showing an increase in mental health services is less clear. On the one hand, interactions with child welfare could lead to increased anxiety or depression by parents which could explain the increase in the use of these services. They could however on the other hand be giving access or encouraging parents who need the extra mental health

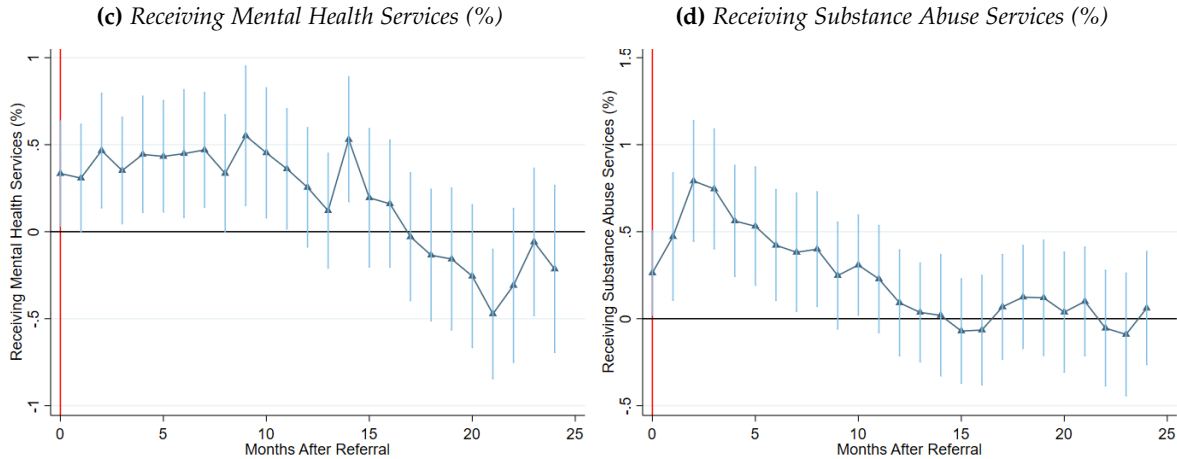
⁶It could also be due to increased income, but I do not have access to income data for these households.

Figure 1.5: *Assigning Mothers to Investigators One Standard Deviation More Likely to Open a Case, Month Level*

Event-Study Method



Preferred Alternative Specification



Notes: Figures present the effects for mothers of being assigned to an investigator one standard deviation more likely to open a case using two different specifications. Coefficients are reported along with 90% confidence intervals, with two-way clustering for investigators and parents. Panels (a) and (b) plot the coefficients for the ϕ_d from equation (1.2) for the 24 months before and after a child welfare referral, respectively for use of mental health and substance abuse services. Panels (c) and (d) are the corresponding plots from my preferred specification which differs from panels (a) and (b) in two ways. First, the coefficients are set to zero before the referral for efficiency purposes since there does not seem to be pre-trends. Second, as recommended by Borusyak & Jaravel (2018), I run a separate regression for each month relative to all the data prior to the referral to address heterogeneity concerns. Panels (c) and (d) thus plot the results from 24 different regressions. As an example, the coefficient for month 10 in panel (c) is obtained by restricting the sample to the sample of mothers in the 24 months before a referral or in the tenth month after a referral, and then plotting the coefficient for the effect of the propensity in this tenth month.

Table 1.4: Parents Receiving Services (%), 0-12m after referral relative to 1-6m prior

	Health			Justice		Social Services		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Medicaid	MH	SA	Courts	Jail	TANF	SNAP	Housing
<i>Panel A: Mothers - Average Opening Effect</i>								
Case Opening Propensity	0.29 [0.18]	0.49*** [0.14]	0.45*** [0.13]	0.17 [0.17]	0.08 [0.06]	-0.32* [0.17]	0.14 [0.21]	0.03 [0.13]
After referral	2.51*** [0.17]	2.71*** [0.16]	4.24*** [0.18]	2.08*** [0.19]	0.95*** [0.08]	1.81*** [0.17]	1.08*** [0.23]	1.20*** [0.15]
<i>Panel B: Fathers - Average Opening Effect</i>								
Case Opening Propensity	-0.06 [0.14]	0.15** [0.07]	-0.04 [0.06]	0.03 [0.12]	0.05 [0.11]	0.05 [0.06]	-0.03 [0.15]	0.07 [0.05]
After referral	0.24* [0.13]	0.48*** [0.08]	0.38*** [0.07]	0.94*** [0.14]	1.06*** [0.12]	0.10 [0.06]	0.26* [0.15]	0.26*** [0.07]
<i>Panel C: Mothers - Separating Out Effects</i>								
Opened no Removal Propensity	0.23 [0.18]	0.30** [0.14]	0.38*** [0.13]	0.01 [0.17]	0.07 [0.06]	-0.18 [0.17]	0.30 [0.21]	0.04 [0.12]
Opened and Removal Propensity	0.20 [0.18]	0.58*** [0.16]	0.29** [0.14]	0.46** [0.22]	0.05 [0.07]	-0.43** [0.17]	-0.37 [0.23]	-0.01 [0.14]
Investigators	460	460	460	460	460	460	460	378
Prior Mean for Mothers	61.5	10.2	5.93	29.3	1.13	18.2	44.2	3.72
Prior Mean for Fathers	22.3	4.13	2.46	36.7	4.86	2.08	19.1	1.07
Number of Mother-months	471042	471042	471042	471042	468746	470928	471042	283376
Number of Father-months	620425	620425	620425	620425	619078	619881	620425	359289

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors are clustered two-way by investigator and individual. Individual-referral, time, and office-year fixed effects included throughout. Propensity is standardized and outcomes are reported in percentage points (%). The variables labeled as the propensities are the interaction of the propensity and being after the referral. Columns list which type of services are received by the parent, from left to right: enrolled in Medicaid, receiving mental health services, receiving substance abuse services, case open in court, in jail, TANF recipient, receiving food stamps, receiving housing assistance. Panel A and C present results for mothers, and Panel B for fathers. Note that the dummy variable for after the referral is not reported in Panel C because the results correspond exactly those in Panel A for that same variable. At the bottom of the table I report the mean outcomes for mothers and fathers in the twelve months prior to the referral.

support to be receiving these services.

Panel C explores which child welfare interventions seem to be driving each of these results. The case opening propensity instrument is split by whether or not a child was removed within the first 3 months following the referral.⁷ Both opening a case without removal and opening a case with removal seem to increase use of mental health and substance abuse services. Being assigned to an investigator with one standard deviation higher propensity to remove a child is associated with twice the level of increase in mental health services relative to being assigned an investigator one standard deviation more likely to open a case without removal (0.58 relative to 0.30). One might indeed expect that a removal could be more traumatic than simply having a child welfare case opened, or alternatively that children may be more likely to be removed from parents with more severe unmet mental health needs. Removal of a child from his/her parents' care requires court approval, and if the child is removed for a duration that is expected to exceed 180 days then the parent will lose TANF benefits. One would thus expect effects of removals on both court involvement and TANF receipt. Results in Panel C show that this is indeed the case. Being assigned an investigator one standard deviation more likely to remove leads to a 0.46pp increase in court involvement and to a 0.43pp decrease in TANF receipt. The negative impacts on parents thus seem to be particularly salient for parents assigned to an investigator more likely to remove. Importantly, these results show that child welfare has impacts on parents even when a child is not removed. The magnitudes of the effects of the propensity to open a case without removal could not be explained by later removals: only 10% of cases that are opened without a removal in the first three months lead to a removal 4-12 months after the referral.

The estimates for how receipt of services changes after the referral is also of interest (variable "After referral" in Table 1.4). They are averaging the effects observed in Figure

⁷For the rest of the paper, I will refer to the propensity to open a case without a removal within the first 3 months to the propensity to open a case without removal. However, it is possible that some of these cases would lead to a later removal of the child. Table 1.1 showed that 3 percent of referrals had a removal within 1 year of the referral that did not occur within the first 3 months.

1.2 for the first twelve months. These estimates could be driven by several factors. First, they could be driven by the event that triggers the referral for child welfare. This seems particularly likely for example for the coefficient showing an 86% increase in jail (0.95 relative to a mean of 1.13), as was discussed previously. Second, they could be driven by the effects of being assigned an investigator with average propensity to open a case. This would include both effects for parents who will have a case opened regardless of which investigator they are assigned to and more generally any effects that are common across all investigations such as the effects of being investigated. It is not possible to separate out these possible channels, but the magnitudes of changes are stunning for some of these variables.

The magnitude of the coefficients for the propensity to open a case may appear small relative to the coefficients for being after the referral. Two reasons explain this. First, as previously mentioned, the incidents that trigger a child welfare referral are substantial and these incidents in themselves – regardless of child welfare – have notable repercussions. Second, the effects of opening a case are measured per standard deviation increase in the propensity to open a case. So, being assigned an investigator in the top quartile compared to the bottom quartile increases use of mental health services by 1.1pp (10.8% increase). For substance abuse, it corresponds to a 1pp (19.9%) increase in enrollment. Clearly, the effects of opening a child welfare case are sizable.

Relative to mothers, impacts of child welfare interactions on fathers seem small as shown in Panel B of Table 1.4. The estimates of the increases following the referral are several factors smaller than for mothers, except for jail where the magnitudes are comparable (approximately 1pp) but the percent of fathers in jail is higher in the year prior to the referral so this also translates into a smaller percentage increase. The only effect on fathers of being assigned an investigator with a one standard deviation higher marginal propensity to accept a family is a 0.15pp increase in use of mental health services by fathers. The standard errors of the propensity coefficient also suggest that these estimates can rule out effects that are of meaningful magnitude for all but possibly jail.

Given that in Figure 1.5, the effects of investigators seemed to fade off, it seems worthwhile to explore whether these effects are temporary or persistent.

1.4.2 Several Years after the Referral

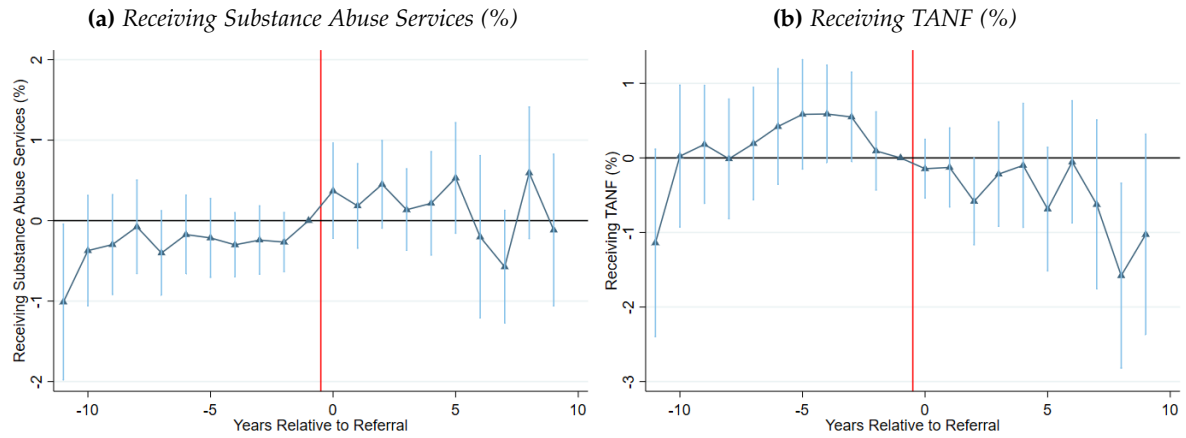
As would be expected, since the effects for fathers were already small when looking in the months immediately following a referral, there are no detectable effects on health, justice or service outcomes for fathers in the years following the referral, when excluding the first two years immediately following the referral which I covered in the previous section (results not shown).

Mothers for their part seem to still be affected much after their interaction with the investigator comes to an end. Results are shown in Table 1.5 and Figure 1.6. As shown in Figure 1.6, significantly more mothers are using substance abuse services for up to the sixth year inclusive of the referral year. In the three years prior to a referral, on average 11.5% of mothers use substance abuse services in a year. Relative to that, mothers assigned to an investigator with a one standard deviation higher propensity to remove a child are 0.62pp more likely to be using substance abuse services each year, which corresponds to a 5% increase in number of mothers using such services in a year (Table 1.5). As shown in Figure 1.6c, these effects seem to persist in magnitude for up to the sixth year following the referral. In addition, these mothers assigned to an investigator one standard deviation more likely to remove are also -0.69pp less likely to be receiving TANF which corresponds to a 2.8% decrease relative to prior years. These persistent effects are statistically significant only for removals, but the confidence intervals for the propensity to open a case without removal are large enough that these results cannot rule out persistent effects for mothers that had an open case but no removal.

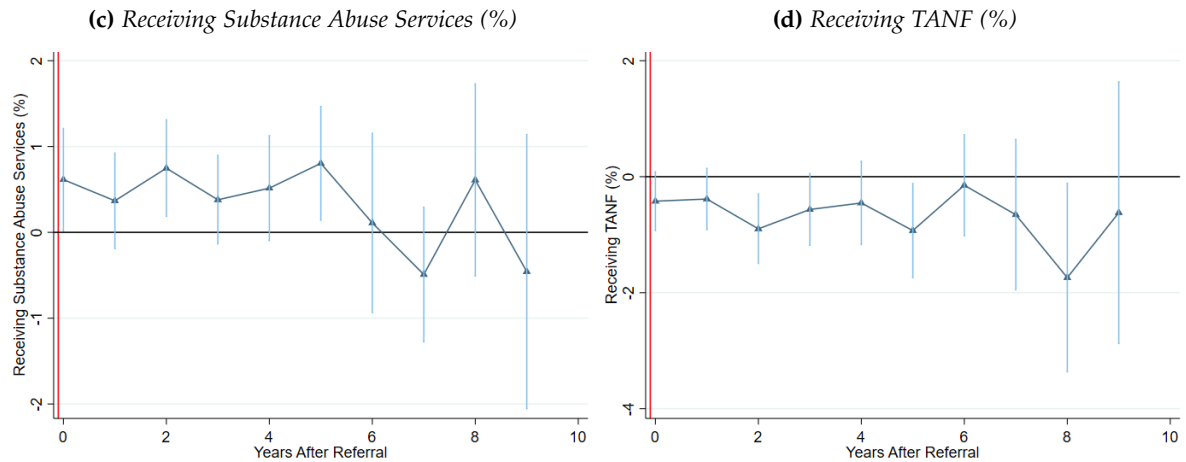
Hence, when looking at the year level instead of at the month level, one can find persistent effects of investigators. Month level analyses were not able to pick up on the smaller but still significant effects of opening a case. Figure 1.7 shows that indeed the effects of opening a case are largest in the year of the referral, but effects persist for several years.

Figure 1.6: Assigning Mothers to Investigators One Standard Deviation More Likely to Open a Case with a Child Removal, Year Level

Event-Study Method



Preferred Alternative Specification



Notes: Figures present the year-level effects for mothers of being assigned to an investigator one standard deviation more likely to open a case using two different specifications. Coefficients are reported along with 90% confidence intervals, with two-way clustering for investigators and parents. Panels (a) and (b) plot the coefficients for the ϕ_d from equation (1.2) for the 11 years before and 10 years after a child welfare referral, respectively for use of mental health and substance abuse services. Panels (c) and (d) are the corresponding plots from my preferred specification which differs from panels (a) and (b) in two ways. First, the coefficients are set to zero before the referral for efficiency purposes since there does not seem to be pre-trends. Second, as recommended by Borusyak & Jaravel (2018), I run a separate regression for each year relative to all the data prior to the referral to address heterogeneity concerns. Panels (c) and (d) thus plot the results from 10 different regressions. As an example, the coefficient for year 3 in panel (c) is obtained by restricting the sample to mothers up to 11 years before a referral or in the fourth year after a referral, and then plotting the coefficient for the effect of the propensity in this fourth year. The referral occurs in the first month of year 0, so year 0 is the first year after a referral, year 1 the second year after a referral, etc.

Table 1.5: Moms Receiving Any Services 3-6 years After Referral Relative to Prior (%)

	Health			Justice		Social Services		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Medicaid	MH	SA	Courts	Jail	TANF	SNAP	Housing
Opened no Removal Propensity	-0.14 [0.35]	-0.23 [0.29]	-0.15 [0.25]	-0.09 [0.32]	-0.07 [0.17]	-0.45 [0.36]	-0.17 [0.39]	0.58 [0.37]
Permutation Test <i>P-value</i>	0.22	0.10	0.15	0.34	0.27	0.01	0.22	<0.01
Opened and Removal Propensity	0.33 [0.42]	0.32 [0.32]	0.62** [0.28]	0.28 [0.42]	0.11 [0.20]	-0.69** [0.33]	-0.08 [0.35]	-0.16 [0.37]
Permutation Test <i>P-value</i>	0.05	0.08	<0.01	0.11	0.23	<0.01	0.38	0.21
Investigators	460	460	460	460	460	460	460	443
Prior Mean	65.6	23.9	11.5	32.8	5.62	24.4	58.6	7.32
Observations	180597	180597	180597	180597	177670	180366	180597	79494

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors in brackets are clustered two-way by investigator and individual. Individual-referral, time, and office-year fixed effects are included throughout. The propensity is standardized and outcomes are reported in percentage points (%). The variables labeled as the propensities are the interaction of the propensity and being after the referral. Columns list which type of services are received by the parent, from left to right: enrolled in Medicaid, receiving mental health services, receiving substance abuse services, case open in court, being in jail, TANF recipient, receiving food stamps, receiving housing assistance. The numbers below the standard errors are the p-values based on a permutation test for which years are after the referral. For each one of my thousand permutation samples, I randomly change which years are after the referral and run the exact same specification as above. Across all the samples, I then look at the probability of being further away from zero relative to the estimated coefficient for each propensity. Those are the p-values that I report from these one-sided permutation test.

1.5 Additional Analyses

1.5.1 Child Welfare Outcomes

Gross (2019) found a decrease in re-referrals of child welfare perpetrators that was statistically significant starting 5 years after the referral and interpreted it as improved parenting by parents who were assigned to an investigator more likely to remove.

My results seem to suggest differently, as shown in Table 1.6. I do not find any statistically significant effects of my propensity instruments on opening another case, but the confidence intervals cannot rule out meaningful estimates. However, once a referral has been made, fathers especially seem more likely to be re-referred when assigned to a higher removal propensity investigator: their likelihood of being referred goes up by 0.73pp 3-6 years after the referral.⁸ This represents a small 2% increase in re-referrals for being assigned an investigator one standard deviation more likely to remove.

Fathers who are assigned to an investigator that is one standard deviation more likely to remove children are 0.87pp more likely to be investigated in subsequent years. This effect corresponds to an approximately 5% increase in investigations. Part of this may come from the increase in referrals but given that typically only half of re-referrals are investigated, it could not explain the full magnitude of the effect which means that conditional on being reported fathers are more likely to be investigated. This result makes sense if one considers that having removed a child from his/her parents care is an indicator of a past abusive or neglectful environment. In light of that, one would expect child welfare to be vigilant about re-referrals for children of parents who previously had children removed from their care.

Estimates for mothers are smaller in magnitude but are in the same direction: mothers assigned in the past to an investigator one standard deviation more likely to remove a child are 0.59pp more likely to be investigated. The impact for fathers is thus 50% higher, with an effect of 0.87pp. These gender differences suggest that child welfare may be more

⁸It is important to exclude the first years after the referral because having just investigated a home may mechanically make it less likely to investigate it again.

Table 1.6: Child Welfare Outcomes 3-6 years After Referral Relative to Prior

	(1) Referral	(2) Investigation	(3) Case Opening
<i>Panel A: Mothers</i>			
Opened no Removal Propensity	0.22 [0.35]	-0.05 [0.27]	0.25 [0.20]
Opened and Removal Propensity	0.36 [0.38]	0.59* [0.31]	0.18 [0.24]
<i>Panel B: Fathers</i>			
Opened no Removal Propensity	0.12 [0.40]	-0.17 [0.30]	0.14 [0.20]
Opened and Removal Propensity	0.73* [0.42]	0.87*** [0.33]	0.25 [0.24]
Prior Parents' Mean	29.3	17.9	8.26

Notes: $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered two-way by investigator and individual. Individual-referral, time, and referral year-office fixed effects included throughout. The propensity is standardized and outcomes are reported in percentage points (%). The variables labeled as the propensities are the interaction of the propensity and being after the referral. This table reports how likely mothers (in Panel A) and fathers (in Panel B) are to be listed as a parent on a new child welfare referral in column (1), to be investigated in column (2) and to have a case opened in column (3). The bottom row reports the means for both mothers and fathers in the three years prior to the referral.

circumspect with regards to fathers' history with child welfare though the confidence intervals for the point estimates are wide enough that these effects not statistically different from each other.

Overall, these results indicate that more stringent past child welfare interactions seem to be associated with heightened future child welfare investigations.

1.5.2 Permutation Tests

I run an additional test for the long term effects of Table 1.5 which relies on a permutation test. The idea is to leverage the fact that effects of being assigned to a particular investigator should only appear *after* the referral. I thus run 1,000 permutations where for each permutation I randomly change which observations are after and before the referral and run the same specification as in Table 1.5. I then calculate the probability of seeing a value

higher than the estimated coefficient reported in Table 1.5 if the coefficient is positive, and lower if the estimated coefficient is negative. The p-values from these one-sided permutation tests are reported in the second row below each coefficient in Table 1.5. First of all, they confirm that the increase in substance abuse services and the decrease in TANF for removals seem to be specific to the precise timing of the referral with no other occurrence of estimates as extreme in the 1,000 permutation samples ($p\text{-value} < 0.001$).

This permutation test also suggests that child welfare interactions may have even more of an impact on mothers than Table 1.5 suggested. Notably, higher removal propensity investigators seem to increase enrollment in Medicaid and investigators with a higher propensity to open a case without removal increase housing assistance received by mothers substantially (8% increase based on estimates in Table 1.5). Mothers seem to be persistently receiving additional support even several years after a referral to child welfare.

The results from the permutation test for mental health reported in column (2) of Table 1.5 are particularly striking. Remember that during the first year immediately following a referral both mothers assigned to an investigator with higher propensity to open a case without a removal and mothers assigned to an investigator with higher propensity remove were more likely to be receiving mental health services. Insignificant estimates based on the standard errors in brackets in Table 1.5 seemed to suggest that several years after the referral mothers who had been assigned an investigator more likely to open a case without removal were less likely to be using mental health services, and mothers who were assigned to an investigator more likely to remove a child were more likely to use mental health services. The p-values from the permutation tests indicate that these may in fact be statistically significant effects. Mothers assigned to an investigator that opens a case without removal thus seem more likely to receive mental health services in the short run but use them less in the long run. On the contrary, removals increase mental health needs both in the short run and in the long run: removing or threatening to remove a child from mother's care may cause an additional mental health burden for mothers.

The one result that is surprising from the permutation tests is the reduction in TANF

associated with being assigned an investigator more likely to open a case without removal (column 7 in Table 1.5). The magnitude of the effect is too large to be explained by later removals that did not occur within 3 months. A possible reason for this is that an individual can only receive cash assistance for up to five years in their lifetime. Given that a substantial number of these parents were on TANF even prior to the referral, 3-6 years after the referral it may well be that these parents exhausted their TANF benefits earlier as a result of the child welfare investigation. Mothers did seem to have somewhat higher rates of TANF receipt in the second year following referral when assigned to an investigator that was more likely to open a case without removal (results not shown).

Overall, the permutation tests thus confirm the long term impacts, and even paint a more nuanced picture of child welfare effects on parents.

1.5.3 Robustness Checks

I run a number of additional robustness checks for both short and long term effects. Given that the long term effects are of greater interest, I present these robustness checks in Table 1.7. The month-level robustness checks are available in Appendix Table A.4. For ease of comparison, I report in the first row of Table 1.7 my main results from Table 1.5: the coefficient on the propensity to remove for the sample of mothers. Each additional row in this table reports the same coefficient on this same sample from a different regression model, where the left-most column indicates what changed in this specification relative to the original specification in row 1.

Concerns with regards to monotonicity and the exclusion restriction warranted controlling for the allegation propensities in row 2. Allegation propensities are noisy and allegations are imperfectly reported, but the test does provide support for the substance abuse results. The effects are consistent for substance abuse services. Estimates for TANF are lower and insignificant but still within the range of estimates in row 1.

Relative to some prior research in child welfare, I did not exclude individuals that were re-referred as long as they were not mechanically reassigned to the same investigator. My

Table 1.7: Robustness Tests for Year Results

	Health			Justice		Social Services		
	(1) Medicaid	(2) MH	(3) SA	(4) Courts	(5) Jail	(6) TANF	(7) SNAP	(8) Housing
1. Original Specification	0.33 [0.42]	0.32 [0.32]	0.62** [0.28]	0.28 [0.42]	0.11 [0.20]	-0.79** [0.36]	-0.08 [0.35]	-0.16 [0.37]
2. Controlling for Allegation Propensities	0.49 [0.44]	0.29 [0.40]	0.62** [0.30]	0.23 [0.50]	0.17 [0.23]	-0.32 [0.38]	0.26 [0.39]	-0.32 [0.44]
3. Keeping only First Referral	0.21 [0.37]	0.49 [0.36]	0.76** [0.30]	0.25 [0.43]	0.18 [0.21]	-1.24*** [0.41]	-0.13 [0.44]	-0.68* [0.37]
4. Before One Caseworker Model	0.42 [0.47]	0.40 [0.35]	0.72** [0.29]	0.19 [0.46]	0.14 [0.22]	-0.89** [0.41]	-0.18 [0.37]	-0.85 [0.62]
5. Control for Imbalanced Characteristics After	0.35 [0.42]	0.32 [0.32]	0.63** [0.28]	0.27 [0.42]	0.10 [0.20]	-0.78** [0.36]	-0.04 [0.34]	-0.15 [0.37]
6. Control for Risk and Safety Levels After	0.30 [0.45]	0.22 [0.34]	0.52* [0.30]	0.18 [0.43]	0.00 [0.21]	-0.92** [0.37]	-0.12 [0.37]	-0.71 [0.55]
7. Truncate Propensity Values	-0.27 [0.39]	0.14 [0.36]	0.28 [0.29]	0.41 [0.49]	0.07 [0.21]	-0.99** [0.42]	-0.27 [0.41]	-0.32 [0.41]
8. Excluding you always in Propensity	0.33 [0.42]	0.34 [0.32]	0.63** [0.29]	0.20 [0.42]	0.11 [0.20]	-0.68* [0.38]	0.06 [0.36]	-0.14 [0.37]
9. Excluding your year in Propensity	0.21 [0.33]	0.19 [0.27]	0.17 [0.25]	-0.01 [0.37]	0.02 [0.18]	-0.28 [0.35]	0.11 [0.32]	-0.32 [0.36]
10. Propensity Residualized on Past Year Outcomes	0.36 [0.42]	0.35 [0.32]	0.68** [0.29]	0.23 [0.42]	0.10 [0.20]	-0.63* [0.38]	0.02 [0.37]	-0.11 [0.37]

Notes: p<0.1. ** p<0.05, *** p<0.01. Table presents results for robustness checks for the coefficient on the removal propensity using the same specification as in Table 1.5. Each row shows the estimated coefficients from separate regressions of the removal propensity interacted with being after the referral. The propensity is standardized and outcomes are reported in percentage points (%). Columns list which type of services are received by the parent, from left to right: enrolled in Medicaid, receiving mental health services, receiving substance abuse services, case open in court, in jail, TANF recipient, receiving food stamps, receiving housing assistance. Row 1 is the original specification from Table 1.5 for comparability (coefficients for being after the referral and the propensity to open a case without removal are not reported). Row 2 controls for the propensity instruments for substantiating different types of allegations. Row 3 keeps only the first referral for each parent. Row 4 excludes all the later data when the investigator assigned becomes the caseworker for the family if the case is opened. Row 5 includes controls for outcomes being different after the referral for whites, those with missing race, referrals reported by the police, and referrals reported by a social worker (characteristics that were imbalanced in balance tests). Row 6 allows for referrals assessed as moderate and high risk or safety before the start of the investigation to have divergent paths. Row 7 truncates the propensity excluding the top and bottom 1% of propensities where there might be concern with regards to monotonicity. Row 8 constructs the propensity by excluding you not only in this current investigation but in all future investigations with that investigator. Row 9 constructs the propensity excluding all investigations done by your investigator that year. Row 10 residualizes the propensity on all characteristics included in the balance test (individual characteristics, referral characteristics, lagged 3m outcomes, lagged outcomes for the past year).

results do not seem to be affected by this decision as shown in row 3, and if anything might even be stronger when I keep parents only the first time I see them referred and investigated in my sample.

Concerns with the exclusion restriction in the later part of my data when some investigators continue working with families even after the investigation, justified running the analyses excluding this later part of my data in row 4. Results seem perfectly consistent.

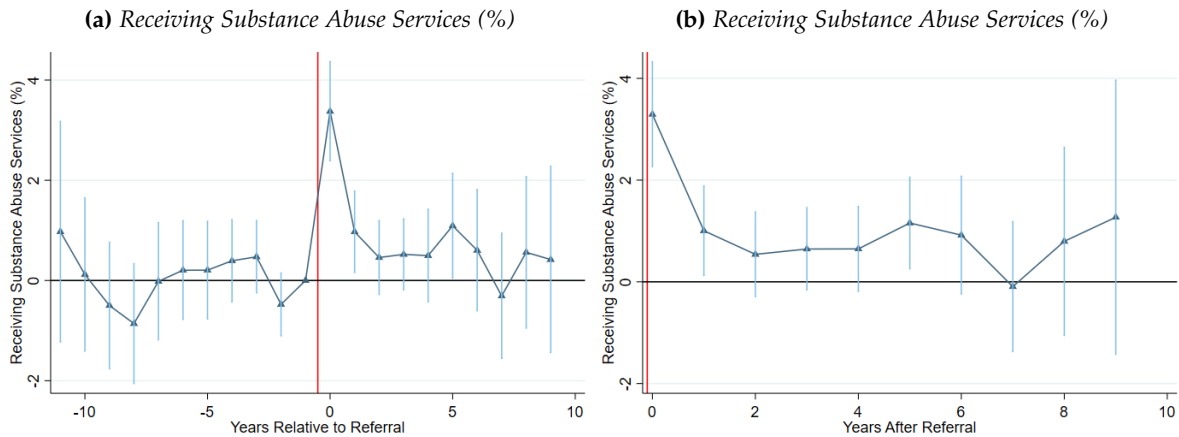
One way to address the concerns of the imperfect random assignment of investigators is to allow for the imbalanced time-invariant characteristics to have a different effect after the investigation. Time invariant characteristics that were imbalanced were race, missing race, referrals reported by the police and referrals reported by a social worker. I thus include a dummy variable for each of these variables after the referral. So, in the example of imbalance by race, this entails allowing whites to be differently impacted after the referral. Results in row 5 are robust to the inclusion of these variables, suggesting that my identification method was able to address these concerns of non-random assignment.

If investigators with higher propensity were assigned to more severe cases, then the severity of the case could be what explains the divergence in trajectories for parents. For a subset of my sample, there was a risk and safety assessment (low/moderate/high) made based on the details of the alleged neglect before the start of the investigation. In row 6, I thus allow for parent's trajectories to diverge based on this assessment of a child's safety and risk. Here again the results seem to hold up.

The graphs examining monotonicity (Figure 1.4) cannot rule out violations of monotonicity at extreme values of the propensity instruments. I thus ran a robustness check in row 7 excluding individuals assigned to the top and bottom 1% of the propensities. Results are consistent for TANF, but the coefficient on substance abuse is halved and no longer statistically significant. This was definitely a substantial concern, and so I ran another robustness check to explore precisely this concern. The results of this additional robustness check are presented in Figure 1.7. The figure presents results from regressions comparing at the year level outcomes for mothers assigned an investigator assigned to above median

propensity to open a case.⁹ The key point is that these results are run excluding mothers assigned to an investigator with a propensity to open a case in the top or bottom 5% because monotonicity seemed to hold from the 5th percentile to the 99th percentile for the propensity to open a case. Results definitely seem consistent: mothers assigned to an investigator with above median propensity to open a case are still approximately one percentage point more likely to be using substance abuse services relative to their peers assigned to an investigator with below median propensity.

Figure 1.7: *Assigning Mothers to Investigators Above Median Propensity to Open a Case, Year Level*



Notes: These figures plot the effect for mothers of being assigned an investigator with above median propensity to open a case. For the purpose of this robustness check, mothers assigned to an investigator with a propensity to open a case in the top and bottom 5% are excluded, and as usual the sample is restricted to mothers assigned to an investigator overseeing at least 10 referrals. Coefficients are reported along with 90% confidence intervals, with two-way clustering for investigators and parents. Panel (a) uses the same specification as Panel (b) of Figure 1.5, except that the propensity is a binary indicator for being above median. The p-value of the F-test for pretrends in Panel (a) is 0.66 which justifies running my preferred specification in Panel (b). Panel (b) specification is comparable to Panel (d) of Figure 1.5.

Finally, I report three additional robustness checks in Table 1.7 which are variations on how to construct the propensity instrument. In row 8, I exclude the individual being considered not only in this referral but in all other referrals investigated by this investigator. It is thus a variation similar in spirit to row 3 except that instead of changing the sample, it changes the propensity. Results are robust to this change.

⁹The binary version of the propensity to remove was not able to pick up such effects, likely because the propensity to remove is a noisier instrument than the propensity to open a case.

In row 9, I construct the propensity instruments excluding all referrals assigned to that investigator in the year of your referral. Point estimates are much smaller though the standard errors are unaffected. It is unclear why results do not hold in this specification, but propensity instruments excluding all referrals in your referral-year are less predictive in the first stage, based on a smaller sample of data, and have fewer investigator clusters (many investigators work less than a year). Results for substance abuse at the month level for this construction of the propensity instrument did not hold either, as shown in Appendix Table A.4.

In row 10, I residualize my propensity instruments based on all the characteristics included in my balance test. This a method used by other researchers for conditional random assignments. Results are consistent and statistically significant for both substance abuse and TANF.

I ran additional robustness checks not presented in Table 1.7. Many of these involved constructing the propensity instrument in varying ways: residualizing the propensity only for office fixed instead of office-year fixed effects, using removals within 1y instead of 3m, constructing the propensity at the individual adult level instead of at the referral level, residualizing the propensity as in row 10 of Table 1.7 except excluding lagged year outcomes. Contrary to Doyle (2007), I cannot construct the propensity of an investigator based on cases that he investigated in other offices, because only 21 investigators worked in more than one office (these are individuals that changed jobs). Finally, clustering at the referral level did not change results.

1.5.4 Attrition

Attrition could cause some of the long-term estimates to be biased towards zero as some families may persist in receiving services after moving outside of the county. Unfortunately, I was not able to find any satisfying proxy for leaving the county. I am thus unable to provide any precise estimate of what portion of my sample has left the county. That being said, 36% of my sample is not interacting with any of the programs in the year of the referral,

and that number goes up to 43% in the fifth year after referral year. So, even five years out at least 57% (and likely much more) of the sample must still be in Allegheny county.

Attrition could also be affected by the child welfare interactions. Anecdotally, some families flee when notified that a child welfare report has been made, or move out of the state after being reunified with their children out of fear of having their children removed from their care again (child welfare registries are state-based). The first type of moving would not bias my estimates, but the second type of moving could possibly bias the long term results in either direction. However, there was no evidence that families were moving more out of county when assigned an investigator more likely to open a case when I examined the effects of my instruments on the probability of not receiving any services. Nor is there any evidence of selective attrition 3-6years out for mothers assigned to an investigator more likely to open and remove a child (results shown in Appendix Figure A.2).

The more time has elapsed since the referral, the more likely a given family will have moved out of county. The long term results of child welfare interactions are thus more likely to be affected by attrition. So, a last check that I do with regards to attrition is to run long term results restricting in each year to the sample of parents who are receiving any type of services. The results on substance abuse persist (results not shown).

Thus, attrition does not seem to be of particular concern with regards to my results.

1.6 Conclusion

This paper shows that child welfare meaningfully affects parents on the margin of having a child welfare case opened, and that these impacts differ for mothers and fathers. Child welfare's impact on mothers is not limited to removing children, contrary to what one could have thought. On the contrary, child welfare seems to be providing extensive support notably for mental health and substance abuse to mothers and even to mothers who did not have a child removed. Future work will want to explore how child welfare is able to affect parents without removing children.

These results do not contribute to the ideological and legal debates on whose rights

come first between children and their parents. However, they do suggest that child welfare, even when focused on the safety and rights of children, also gets parents to address alleged issues in their lives. In that light this paper highlights the plausible value of increasing funds for preventative services in child welfare, at least in this one county. More work of course will be needed to understand whether such effects of child welfare are consistent across other child welfare agencies. Even so, these results are particularly timely since the Family First Prevention Services Act in 2018 dramatically increased availability of federal funds for evidence-based preventative services in child welfare.

Helping parents get mental health or substance abuse treatment may be in itself worthwhile from a societal point of view, but it also could have ripple effects on current or future children that these parents may have. Baranov et al. (2020) indeed suggests that getting depressive pregnant mothers mental health services improved children's outcomes years later. In some sense child welfare is addressing home environment concerns in the households that are most likely to have intergenerational effects because they are homes with children. Future work may want to consider whether these effects on parents translate into safer and more nurturing homes for children.

Chapter 2

Saved from Death? Effects of the Child Protection System on Child Deaths Reported to the Medical Examiner

2.1 Introduction

Inequalities in age at death, cause, and place of death are widely acknowledged in many developed countries (Cutler et al., 2006; Mackenbach et al., 2008; Tuljapurkar, 2011; Oliviere et al., 2011). For example, infant mortality is twice higher for Black Americans than for non-Hispanic American whites (CDC, 2020). These disparities are consequential since a death during childhood reduces the most years of life.

The child protection system – also referred to as child welfare – aspires to address disparities in childhood mortality. Child welfare’s goal is to address gross injustice among children, so that each child’s basic needs are provided for: food, health, shelter, and no abuse. First and foremost, this means averting child deaths due to caregiver neglect or abuse.

Surprisingly, to the best of my knowledge, there is no causal evidence that child welfare reduces child mortality. In fact, what has been shown is rather the opposite: children in

foster care have 42% higher mortality rates than the rest of the population (Chaiyachati et al., 2020). This naturally casts some doubt as to whether child welfare keeps children safe from harm. Recurrent incidents publicized by the press of children dying even when known by the child protection system also convey similar skepticism regarding the effectiveness of child welfare.

This paper is the first to apply causal identification methods to assess effects of child welfare on child deaths reported to the medical examiner (a medically trained coroner). Apart from child deaths with a clear natural cause attended to by a medical professional, all child deaths should be reported to the medical examiner. I focus on the two main child welfare interventions: opening a child welfare case and removing a child from their home. To estimate causal effects, I rely on imperfect but close to random assignment of child welfare investigators to families. Investigators indeed vary in their propensity to recommend case opening and removal of the child from the home. Using administrative data from one county in Pennsylvania, I thus compare the frequency of child deaths by investigators' propensity to recommend child welfare involvement. Three results stand out.

First, I find that child welfare substantially reduces child fatalities that required a medical examination. For children where investigators would disagree regarding whether to open a case, opening a child welfare case reduces the child mortality rate by a stunning 3.65% using reduced form estimates, or 5.3% deaths using 2SLS estimates. More experienced investigators seem the most effective at inducing these reductions in mortality and immediate effects are particularly pronounced among infants and toddlers, Black American children, and economically disadvantaged children. Importantly, these effects of child welfare need not exclusively be through preventing deaths from abuse or neglect. Child welfare could also be improving parenting practices, such as through parent adoption of safer sleeping practices for their children which reduce sudden infant deaths.

Second, I find suggestive evidence that in cases where investigators would disagree as to whether removing a child from their home is necessary, removing a child dramatically reduces child deaths reported to the medical examiner a few years after the referral. However,

the point estimates of the main specification are imprecisely estimated, suggesting that I lack statistically power to provide conclusive evidence on the ways through which opening a case reduces child deaths.

Third, as a result of these effects, child welfare interventions dramatically reduce disparities in childhood mortality. Indeed, children with the largest reductions in death are those with the highest predicted likelihood of child death. Furthermore, prevalence of child welfare interventions is much higher among disadvantaged children. For instance, given that child welfare interventions are more prevalent among Black children, the gap in mortality rates for children would be two to three times higher without such disparate child welfare involvement. Of course, these estimates warrant caution given the surprising magnitude of these effects.

This work contributes in two ways to the Economics literature. Previous work evaluating the effects of child welfare for children (Doyle, 2007, 2008; Warburton et al., 2014; Roberts, 2019; Gross, 2019; Chyn et al., 2019) had examined important child outcomes but these were outcomes outside the immediate scope of child welfare's mission (future income, test scores at school, receipt of social services, teen pregnancy, juvenile justice involvement). My contribution relative to this literature is to be the first to examine the effects on child deaths, and one of the first to examine target outcomes of the child protection system. Doyle (2013) had indeed found that removing children increased their use of emergency care in the year following the referral using data from Illinois in the 1990s, which is an outcome of direct interest to child welfare agencies. Three reasons may explain why my results indicate that children are protected by child welfare interventions, whereas the results of Doyle (2013) suggested otherwise. The first potential reason is that Doyle focused on a sample of children 5-15 years old whereas the strongest effects that I find are for children younger than 4 years old. And in fact, previous work Warburton et al. (2014) and Chyn et al. (2019) seems to suggest different effects of child welfare based on the age of the child. A second potential explanation is that Doyle's results would also be compatible with a reduction in child fatality if, as a result of a child welfare intervention, children were being taken more

to the hospital when in need of it or out of precaution. Finally, it is also possible that effects are different over time and across child welfare agencies. Studies in North Carolina, Rhode Island and Michigan in the 2000s and 2010s indeed seem to find positive effects for children (Roberts, 2019; Gross, 2019; Chyn et al., 2019) which contrast with the negative effects found by Doyle in Illinois during the 1990s (Doyle, 2007, 2008).

The second main contribution of this paper is to the large literature on health disparities and inequality (Peltzman, 2009; Vogl, 2018; Deaton, 2013). Prior work has emphasized how health insurance, nutrition programs, reductions in domestic violence, reductions in environmental pollution, and nurse home visiting programs improve health outcomes for young children (see Aizer & Currie 2014 for a review). My work highlights the importance of an essential program improving child health outcomes: child welfare. Part of the reductions in infant and toddler mortality, and reductions in racial mortality gaps for these children documented by Currie & Schwandt (2016a) and Currie & Schwandt (2016b) might even arise from child welfare interventions. Although often overlooked, the child protection system plays a key role in addressing child mortality inequality. Finally, in a growing literature highlighting how structural racism has lead to worse health outcomes for Black Americans (Bailey et al., 2021; Alsan & Wanamaker, 2018; Smedley et al., 2003), this paper brings some hope by providing suggestive evidence that current higher prevalence of child welfare interventions among Black families is saving more Black American children's lives.

The rest of this paper will proceed as follows: the first section will present some background about the setting and relevant work from the social work literature, before describing the data including advantages and limitations of using medical examiner records to track child deaths. The second section will then cover the methodology used in this paper and assess its validity. Finally, the third section will provide the results, heterogeneity analyses and robustness tests before concluding.

2.2 Child Welfare and Child Deaths

2.2.1 Background on Child Welfare

The foremost purpose of the child protection system is to protect children from harm. Death is an extreme form of harm. Eliminating or at least reducing preventable deaths is thus a key part of child welfare's mission statement. This is a salient goal to child welfare employees since they may face legal consequences in cases where they were negligent in their job, as evidenced by the case of the death Gabriel Fernandez which received substantial publicity with a Netflix documentary series in 2020. Some journalistic work has even shown that when a child dies from maltreatment and makes the headlines, workers in that area remove more children from their caregivers' care (Williams, 2018).

Child welfare hears of potential maltreatment or neglect through calls. Neighbors, family, friends as well as mandated reporters like teachers or doctors call child welfare to report instances where they suspect a child may be abused or neglected. Based on the allegations, child welfare then decides to investigate some of the families reported to their attention through these calls. The investigation is typically carried out by a social worker who will visit the family and make an assessment of whether there is any reason to be concerned for the child's well-being or safety. If there is any potential concern, a case may be opened with child welfare. This may be done with the purpose of checking in with the family for a few more weeks, of providing services, or – in the most severe cases – of removing a child from their caregivers' care when the child's safety cannot be ascertained at home.

2.2.2 Literature Review from Social Work

The literature on social work has extensively studied both the topics of the child protection system, and child mortality. Much work has focused on protective or risk factors associated with child maltreatment fatalities, and policies that can help reduce maltreatment or fatalities (Putnam-Hornstein et al., 2013; Barth et al., 2016; Olds et al., 2014; Vaithianathan et al., 2018). Of particular interest for this study are a few studies that have examined what types of

child welfare interventions were associated with a child fatality among the most severe cases of abuse and neglect (Chance & Scannapieco, 2002; Thurston et al., 2017), and among children with past child welfare involvement (Douglas & Mohn, 2013; Douglas, 2016). These studies rely on multivariate regression models and use mostly retrospective data. Their results suggest that more services and more child welfare involvement may be protective from fatality among children involved with child welfare. However, all of these studies are limited by their methodology which is effective at picking correlations but unable to conclude to causal effects. As they focus on children who are already interacting with child welfare to some extent, they are also unable to speak to the effects of child welfare on child fatalities for families without prior child welfare involvement.

A second large strand of research in social work has highlighted measurement issues in child deaths from maltreatment. It is widely acknowledged in the literature that deaths from child maltreatment and neglect are undercounted by 50 to 90% in death records and official maltreatment reports from child welfare agency records (Crume et al., 2002; Schnitzer et al., 2008; Herman-Giddens et al., 1999). Indeed, diagnosing the cause of death, especially among children, can be difficult if not impossible. Sudden and Unexpected Infant Death (SUID) are often misclassified, Shaken Baby Syndrome may be hard to notice (Newton & Vandeven, 2006)... Examinations can often only conclude that maltreatment was probable or possible. In this light, child death review panels are sometimes viewed as the most promising surveillance approach (Covington, 2019), but J. W. Parrish et al. (2017) shows that there is still substantial disagreement across different child death review panels as to which deaths involved maltreatment or neglect. Given this particular difficulty in counting number of child deaths because of abuse or neglect, this paper will contribute to the literature by showing that the effects of the child protection system on child mortality can be examined while staying agnostic with regards to the results and conclusions medical-legal investigations or child death reviews. Examining effects on child deaths concluded to be because of child abuse and neglect would limit statistical power, and could introduce bias.

2.2.3 Data

This project makes use of data from the child protection system in Allegheny county, Pennsylvania in the United States (city of Pittsburgh). It includes information on all children living in the county that were ever called in to child welfare from May 2009 to December 2018 for child neglect (which represent three quarters of calls to child welfare), as well as information on whether the family was investigated, results of the investigation and whether child welfare decided to open a case and remove a child. For the years of 2017 and 2018, it also includes information on all calls made for physical or sexual abuse which are all required to be investigated by law in this county.

This child welfare data is merged to records from the Allegheny County Medical Examiner's Office (ACOME). The Medical Examiner investigates *all* deaths in the county except those that were clearly natural and for which a physician issued a death certificate. Any violent, accidental, sudden unexpected, or medically unattended deaths within the county ought to be reported to this office. So, in some sense, deaths reported to the medical examiner should be the child deaths purged of natural deaths that are not of interest for this paper. In that light one might privilege the use of medical examiner records over vital statistics data.

The data I have access to indicates which of the children referred to child welfare had a death reported to the medical examiner's office, as well as the month and year that this occurred at. It includes all child deaths reported to the medical examiner *regardless* of cause. Medical examiner records are typically preferred to vital statistics data for their more reliable information on cause of death (G. Parrish, 1995; Herman-Giddens et al., 1999). However, I do not have cause of death in the data I have access to, and in fact a key advantage of my methodology is that it precisely does not rely on determined cause of death. This is important, because – even with an autopsy – it is often hard to determine the manner of death: in ACOME's annual report, for example, manner of death could not be determined for 13 of the 22 infants that had a pathological examination in 2019. My methodology thus includes cases where cause of death may have been unclear, as well as deaths where the

role of maltreatment might have been missed by the medical examiner.

The caveat of using medical examiner records is that I do not know of all child deaths for children in my sample. In 2018, half of child deaths in Allegheny county were reported to the medical examiner. That number differs substantially by age group: 83% of children who died between the ages of 5 and 19 were reported to the medical examiner, compared to only 30% for 0-4 year olds (author's calculations using the medical examiner's report for 2018 and Pennsylvania vital statistics data). The lower rates of coverage for young children could be due to deaths from developmental and genetic conditions that were present at birth, which would most likely not be reported to the medical examiner. Still, these rates are high since only approximately 20% of deaths in the US are certified by a medical examiner (G. Parrish, 1995).

One might be notably concerned with using medical examiner records to track child deaths if doctors and police use discretion in what deaths they report to the medical examiner's office (the difference between what ought to be reported and what is reported). Although I am not aware of any evidence of this, it is definitely plausible. In general though, if there were to be discretion in what deaths are reported to the medical examiner, one would expect that children with more child welfare involvement prior to their death would be considered as more "suspicious" deaths and so more likely to be reported to the medical examiner. This would bias me in the opposite direction of the results that I find.

2.2.4 Naïve Descriptive Approach

In Table 2.1, I present the sample size and frequency of deaths by type of child welfare involvement.¹ In total over this time period close to 88,000 unique children were referred to child welfare as potentially subject to abuse or neglect. Of these, I observe a medically examined death for 382 of them between 2008 and 2018, which means that 4.3‰ of these

¹The full dataset which will be used in the analyses includes all referrals for each child from 2009-2018: a given child may be referred several times over this period. Table 2.1 includes each child only once instead of each time they are referred. The table for the referral-specific children sample is presented in Appendix Table B.1.

children were medically examined over the course of this decade. This mortality rate already seems extraordinarily high compared to the Pennsylvania 0.57‰ death rate for 0-19 year olds over the full duration of childhood for deaths from all causes (2010 US Census).

Table 2.1: *Number of Children and Rates of Medically Examined Deaths, May 2009 - Dec 2018*

Sample	(1) Nb Unique Children	(2) Nb Examined Deaths	(3) Rate (‰)
Referred	87849 (100%)	382	4.3
Ever Investigated	56875 (65%)	260	4.6
Ever Had a Case Opened	29433 (34%)	170	5.8
Ever Removed	5115 (6%)	28	5.5

Notes: This table reports in the first column the number of unique children that, from May 2009 to December 2018, were referred to child welfare (row 1), investigated by child welfare (row 2), had a case opened with child welfare (row 3), and were removed from their home (row 4). The percent of referred children that each row represents is reported below the raw counts. The second column reports the number of children that died among each of these groups of children, and the third column reports the rate of medically examined deaths children per thousand children.

Of these children approximately two-thirds (65%) will be investigated at some point between May 2009 and December 2018, 34% will have a child welfare case opened at some point and 6% will be removed from their home by child welfare at some point (placed in out of home care).

Interestingly, the rate of medically examined deaths is higher for children where child welfare intervened. A naive approach examining whether receiving child welfare interventions (opening a case, removing a child) would thus conclude that child welfare seems associated with higher rates of medically examined child deaths, and that child welfare may be doing more harm than good.

However, a more careful analysis will quickly point out that child welfare should be intervening in the families where the child was most likely to die. The rate of medically examined deaths could thus be higher precisely because child welfare was trying to intervene in the families that most needed it. This raises the need for research that is able to identify

causal effects, instead of just examining associations and correlations using multivariate analyses.

2.2.5 Disparities in Medically Examined Deaths and Child Welfare Involvement

Before moving on to present my identification strategy, it seems worthwhile to describe the children being investigated by child welfare. Column 1 of Table 2.2 presents the full sample of children investigated by child welfare for whom I could find the investigator in the child welfare records (not just those investigated by an investigator overseeing at least 50 investigations, relative to column 3). A few patterns are striking.

Although the percent of girls and Hispanics are relatively comparable to the population at large, children investigated by child welfare are disproportionately Black American. 18% of children in Allegheny county are Black American according to the Department of Human Services, whereas at least 52.8% of investigated children are Black American and race is unknown for an additional 7.7%, so the percent of Black American kids among those who are investigated could be as high as $52.8 + 7.7 = 60.5\%$. These numbers of course are stunning. Children investigated by child welfare are also younger: $8.3 + 15.9 = 24.2\%$ of investigated children are below the age of 4 whereas according to US Census data in 2010 only 10.5% of children in this county are below this age. This speaks to the fact that young children are a more vulnerable age group. Families investigated by child welfare are also much more likely to be disadvantaged: 76.8% of children had been on Medicaid at some point before the referral, and they are also frequent beneficiaries of TANF and SNAP (food stamps).

In column 2 of Table 2.2, I present the relative likelihood of having a medically examined deaths for this same sample of investigated children. Instead of using realized death rates for my sample which would conflate the effects of child welfare with child inherent risk, I predict investigated children's likelihood of having a medically examined death using medically examined deaths that occurred for the sample of children referred to child welfare but who were never investigated (probit regression model is presented in Appendix Table

Table 2.2: Inequality and Death

	(1) Percent (%)	(2) Relative likelihood of death	(3) Relative likelihood of being a complier
Girls	49.5	0.66	0.99
Black American	52.8	1.18	0.96
Missing Race	7.7	0.77	1.03
Hispanic	1.2	0	0.62
Missing Ethnicity	25.1	1.42	1.05
Infant (<1y)	8.3	3.6	0.99
Toddler (1-3y)	15.9	1.22	0.95
TANF Recipient Before	29.7	0.59	1.01
SNAP Recipient Before	48.8	0.70	0.98
Medicaid Recipient Before	76.8	0.99	0.96
Received Mental Health Services Before	27.5	0.97	0.96
Number of Observations	101623	101623	75428

Notes: Column 1 reports the percent of *investigated* children with each of characteristics in the leftmost column. The sample size is larger here than in Table 2.1 because the same child can be investigated multiple times. Column 2 reports the relative likelihood of predicted medically examined deaths for each of these subgroups relative to the mean. Predicted likelihood of a medically examined death is estimated from the sample of children referred to child welfare but never investigated in this decade (Appendix Table B.2). So, for example girls my sample are predicted to be 33% less likely to have a medically examined death relative to average and Black Americans are 15% more likely to have medically examined death among investigated children. Column 3 reports – for the sample of children assigned to an investigator who oversaw at least 50 investigations (hence the reduced sample size relative to columns 1 and 2) – the likelihood of being a complier for each subgroup which is estimated by taking the ratio of the first stage coefficient for each subgroup relative to the full sample first stage coefficient (see Maestas et al. 2013 for more details).

B.2). As seen in column 2 of Table 2.2 for the sample of investigated children, girls are predicted to be 33% less likely to have a medically examined death relative to boys, and Black American children are 18% more likely to have a medically examined death. Unsurprisingly, infants are much more likely to have a medically examined death, +360% than average. This includes of course sudden infant deaths. But young children more generally are more likely to have a medically examined death: +22% relative to the average for children 1-3 years old.

These descriptive statistics highlight thus two facts that will be particularly important in interpreting my results later on. First, child welfare intervenes much more frequently in Black American and disadvantaged households. Second, child welfare investigations also disproportionately target young children who are at particularly high risk of having a medically examined death.

2.3 Methodology: Propensity Instruments

To identify causal effects, this paper relies on the use of a particular type of instrumental variables referred to as “propensity instruments” in the Economics literature. Initially developed by Kling (2006) for judges, it has been widely applied in other settings including to child welfare investigators (Doyle, 2007, 2008, 2013; Roberts, 2019; Gross, 2019; Chyn et al., 2019; Grimon, 2020). In a first subsection, I will present a brief overview of this methodology before describing my empirical specifications and reviewing the validity of this identification strategy in my setting.

2.3.1 The Propensity Instrument: Investigator Assignments

The basic intuition behind this methodology is that some investigators are more likely to recommend child welfare involvement: they have a higher *propensity* of getting child welfare involved. If investigators are quasi-randomly assigned to families, then one can compare the outcomes of similar families that were assigned to investigators with different propensities. The differences in the outcomes for the children are then attributable to child

welfare involvement. This methodology relies on a few critical assumptions that I will examine at greater length in section 2.3.3, notably that:

1. children assigned to different types of investigators are similar (*quasi-random assignment* of investigators)
2. I can tell who are investigators more likely to recommend child welfare involvement (*first stage*)
3. investigators' influence on these families is exclusively through whether or not they recommend child welfare involvement (*exclusion restriction*)
4. investigators would agree in ranking families by need of child welfare involvement (*monotonicity* assumption).

The key variable behind this methodology is of course a measure of the propensity for each investigator to get child welfare involved. To construct this variable, the idea is to use all of the children that this investigator examined and count what percent child welfare got involved with. However each individual's own outcomes are not independent of whether child welfare got involved in their own case leading to bias towards OLS estimates (see Angrist et al. 1999). The idea is thus to build the propensity for each investigator using all children this investigator examined, *except* those in your family. Under quasi-random assignment of investigators, the propensity measure thus constructed should be independent of your own case and a good measure of how likely your investigator was to get child welfare involved. Such instrumental variables have also been called leave-one-out instruments or jackknife instrumental variables as a result (Angrist et al., 1999).

In practice, the propensity PO to open a child welfare case of investigator j assigned to child i listed on referral r is constructed in the following manner:

$$PO_{i,r,j} = \frac{\sum_{c,s \neq r} O_{c,s,j}}{N_j - n_r}$$

where the numerator sums the number of children c assigned to investigator j but not in referral r who had a child welfare case opened, with $O_{c,s,j}$ is an indicator for whether child c in referral s assigned to investigator j had a case opened. This is then divided in the

denominator by the number of children that investigator j investigated N_j subtracted by the number of children in the referral r , n_r .

One minor complication to this set up in my setting is that investigators are quasi-randomly assigned to cases only within the office that they belong to. There are six different offices. In addition, you can only be assigned to investigators that were working the year you were referred in. So, throughout the paper, any time I will be using my propensity instruments, I will be including office year fixed effects to only leverage variation across investigators *within* the same referral year and office.

In addition to measuring how likely investigators are to open a child welfare case, I construct an equivalent measure for how likely investigators are to remove a child, $PR_{i,r,j,o}$ (see Grimon 2020 for more details on constructing multiple propensity instruments). This is constructed by counting the number of children who were removed from home in the three months following the referral.

Such measures of propensity are only as good as what I can observe. Notably, for investigators that investigated few cases, these propensity measures are going to be noisy and of poor quality. Given the rare nature of the outcome of interest (death), I thus restrict the below analyses to investigators that oversaw at least 50 investigations to reduce noise in the propensity instruments (comparable to Gross, 2019). That leaves me with 226 investigators over the course of this decade, and approximately three-quarters of the full sample of child-referrals.

2.3.2 Empirical Specifications

My main empirical specification will thus be to simply regress whether a child i on referral r died in circumstances that lead to a medical examination in the years after referral r ($Y_{i,r}$) on the propensity measure:

$$Y_{i,r} = \beta_0 + \beta_1 PO_{i,r,j,o} + \delta_{o,y_r} + \epsilon_{i,r} \quad (2.1)$$

where δ_{o,y_r} are a set of office (o) and referral year (y_r) fixed effects and ϵ is the error term. β_1

will thus measure the effect of being assigned to an investigator whose *measured* propensity to open cases is 1 and opened all cases, relative to being assigned to an investigator whose measured propensity to open cases is 0 and opened no cases. Remember that this is an instrumental variable, so to back out the effects of being accepted by child welfare, I will need to divide this coefficient by the corresponding first stage coefficient in Table 2.3. As children can be referred multiple times across this decade and that the propensity measures are highly correlated among children assigned to the same investigator, I cluster standard errors two-way at the child and investigator level.

A unique feature also in my setting is that children can be listed on a referral even after their death if the person making the referral did not know the child had passed away or if the family has had a past referral (the new referral would be linked to the past referral). This allows me to run a different specification, where instead of relying on random assignment of investigators, I rely on the fact that they were assigned to families that were evolving in the same way (parallel trends in the rates of medically examined death rates), as in Grimon (2020). The idea is then to compare how the gap in medically examined death rates changed after the children were assigned to an investigator more likely to get child welfare involved, relative to observed gaps before, and this can be done for every year before and after the child welfare referral. So, if $W_{i,r,t}$ measures whether the child had a medically examined death by year t , then:

$$W_{i,r,t} = \alpha + \sum_{d=-n}^m [\phi_d \mathbb{1}_d \cdot PO_{i,r,j,o} + \delta_{o,y_r,d}] + \epsilon_{i,r,j,t} \quad (2.2)$$

where $\mathbb{1}_d$ indicates being d years relative to the referral, and $\delta_{o,y_r,d}$ are year of referral and office fixed effects for each relative time period d . In this specification the key coefficients of interest are thus the ϕ_d which indicate how deaths are different when assigned to an investigator with measured propensity 1, relative to the year prior to the referral, when controlling for differences across office years.

2.3.3 Validity of the Identification Strategy

Remember that this identification strategy relies on four assumptions: random assignment of investigators to children, a first stage, monotonicity and the exclusion restriction.

Imperfect Random Assignment

The first assumption behind using these propensity instruments is thus that investigators are quasi-randomly assigned to children and their families. The child welfare agency says it assigns cases randomly. In addition, for most referrals, assignment of the investigator will take place before child welfare actually knows what is happening in the home: the only known information is what the reporter told child welfare. Social workers say that it is *ex ante* often hard to predict how difficult a case will be based solely on this information. Assignments of investigators to specific types of children thus seems unlikely to be done in particularly meaningful ways.

However, a balance test suggests that assignment of investigators is not perfectly random. This is a rather frequent concern in the Economics child welfare literature, as referenced in Grimon (2020). I include 43 variables in these balance tests, so results are presented in Appendix Table B.3. The p-value of the F-test for random assignment is <0.01 for the propensity to open a case, and is 0.06 for the propensity to remove a child. However, a closer look at the variables showing up as statistically significant indicates that the imbalance is statistically significant but rather small in magnitude. For example, being Black American is associated with being assigned to an investigator that is 0.8pp more likely to open a case (relative to a mean propensity of 42pp). Investigators that open more cases are also assigned to slightly younger children, slightly fewer calls from mandated reporters, but more calls reported by the courts, police and social workers, as well as fewer non-urgent referrals (fewer required response times of 3-10 days), and are more likely to be assigned to children with a court case in the past year. The imbalance for all of these on the order of magnitude of 1pp, except for non-urgent referrals which are assigned to investigators that are 3.7-3.8pp less likely to open a case on average. For removals, children reported to be

in present danger are assigned to investigators 0.1pp more likely remove. Referrals where the youngest child is younger, children with a response time of 3-7 days or within 24h, and children with recent increases in court involvement are all assigned to investigators more likely to remove a child by at most a 1pp. Removals are much rarer, so the magnitude of these differences are more of concern. Still given the small imbalance in terms of magnitude especially for the propensity to open a case, it seems reasonable to consider this as close to random assignment. Analyses presented later in the paper will attempt to address these concerns by relying on parallel trends instead of random assignment as in Grimon (2020) and by controlling for this predicted likelihood to die to give a better sense of how large bias may be.

In this particular instance though, the imbalance from these imperfect random assignments seems to bias my results in the opposite direction of the effects that I will find. Indeed, using the demographic characteristics of children, I predict the likelihood that a child would have a medically examined death for the sample of children referred to child welfare but who are never investigated by child welfare in this decade (see Appendix Table B.2 for results). Children with a higher likelihood of a medically examined death are assigned to investigators with a propensity to open more cases (correlation coefficient $r = 0.02$, $p_{value} < 0.01$). A visual examination of the years prior to the referral in Figure 2.1a also seems to confirm this. Investigators with higher propensity of getting child welfare involved were more likely to be assigned to children at higher risk of having a medically examined death, whereas I will find effects in the opposite direction. This suggests that my estimates will be underestimates of the true effects of getting child welfare involved with a family.

Relevance, Monotonicity and Exclusion Restriction

I present results from the first stage in Table 2.3, with cluster bootstrapped standard errors given potential concerns in the calculation of the analytic standard errors in first stage regressions as highlighted by Hull (2017). As one would expect, given that my estimated

measures of investigator propensity are noisy, the first stage coefficients are biased towards zero instead of being equal to 1. A child assigned to an investigator with a measured propensity of 1 to open a case, has a 0.697 chance of having a case opened (column 1, Table 2.3). The coefficient is highly statistically significant with a t-statistic of 18.

Table 2.3: First Stage

	(1) Case Opening	(2) Removal
Propensity to Open a Case	0.697*** [0.039]	0.020 [0.013]
Propensity to Remove a Child		0.352*** [0.074]
Number of investigators	226	226
Observations	75428	75428

Notes: $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are cluster bootstrapped by investigator, with 800 bootstraps. Office-year fixed effects are included throughout. The sample is restricted to investigators overseeing at least fifty referrals. This table reports the first stage coefficients of my propensity instruments. The propensity is individual specific as it is constructed using all referrals investigated by the investigator, excluding this specific referral the child is listed on (leave-referral out). In the first column is the specification predicting case opening based on your investigator's propensity to open a case. In the second column is the specification predicting whether the child is removed based on their investigator's propensity to remove, but holding constant the investigator's propensity to open a case.

Removals are a much rarer outcome, with only 3% of child-referrals being removed within three months of a referral (Appendix Table B.1). So, for investigators overseeing at least 50 investigations to which I have restricted to, I would expect to see only a few child removals. Consequently, the removal propensity is much noisier than the propensity to open a case which explains why the first stage coefficient is so much lower at 0.35 though it is still quite statistically significant (column 2, Table 2.3). These results highlight that the first stage is highly statistically significant; the propensity measures constructed in my setting are quite predictive of child welfare involvement.

With regards to monotonicity, it also seems reasonable to assume that investigators would in general agree when ranking families based on whether a case should be opened. There is bound to be some disagreement, but some families clearly need child welfare involvement

and others do not. I run three sets of tests for monotonicity which are common in the literature. First stage coefficients are monotonically increasing throughout the distribution for both propensity instruments as shown in Appendix Figure B.1 (with a possible exception at the right tail extreme value of the propensity to remove likely due to mismeasurement in the construction of the propensity). I also check that the first stage coefficient is positive and statistically significant for all the subgroups using the characteristics in the balance test, in addition to the listed allegations (106 subgroups). That is true for both propensity instruments, except in a few instances when the sample sizes of the subgroup is too small. Finally, I run a third check for monotonicity for the propensity to open a case (removals are too rare to construct valid propensity instruments on split samples). I construct the propensity to open a case based on the reverse sample and check that the propensity constructed in this way also leads to a positive and statistically significant first stage. If monotonicity holds, investigators more likely to open a case for one type of individuals should also be more likely to open a case for the others. So, for example, in the case of Blacks, I construct the propensity to open a case using the sample of non-Black children and then examine the first stage coefficient of this propensity measure on the sample of Black children. The reverse sample first stage coefficients are all positive and statistically significant, except when the sample size to construct the propensity is too limited. There is also one additional exception which is surprising: when using the sample of abuse cases to calculate the propensity, the first stage coefficient of that propensity on the sample of neglect cases is positive but not statistically significant (coefficient of 0.079 with a standard error of 0.059, see Appendix Table B.4). The reverse sample first stage coefficient however for the sample of abuse cases is however positive and statistically significant, which suggests that the concern is rather that the propensity instrument on the abuse sample is too noisy. Indeed, recall that all referrals for abuse need to be investigated by law, so although the sample size is still large, the rate of case opening is much lower so the propensity instruments constructed from the abuse sample will be much noisier. This one exception thus does not seem to be of particular concern, though I will examine heterogeneity for abuse and neglect

cases later on. I present a subset of the results from these two sets of regression tests for monotonicity in Appendix Table B.4. Overall, all three sets of test thus seem consistent with the assumption of monotonicity in this setting.

Finally, for most observations in my sample (over three quarters), investigators only interact with families at the investigation stage. After the investigator has closed the investigation, if a case is opened then another child welfare employee will be assigned to work with the family. During the investigation, the investigator's chief role is to assess the situation so as to make the most informed report that will determine whether child welfare should get involved. Investigators thus only influence families through their assessment: the exclusion restriction seems to hold. I do also run a robustness check later to exclude some of the later data where investigators could be assigned to the family when a case was opened.

2.4 Results

2.4.1 Effects of Child Welfare on Child Medically Examined Deaths

In Panel A of Table 2.4, I present the main results of this study using the specification presented in equation (2.1). Being assigned to an investigator with a measured propensity of 1 to open a case leads to a decrease in the per thousand rate of medically examined child deaths by 3.65 percentage points which is statistically significant at the 10% significance level. Given that the first stage coefficient is 0.697 (from Table 2.3), this means that for children where investigators would disagree about opening a case opening a case reduces child mortality for medically examined deaths by $-3.65/0.697 = -5.24$ per thousand children. The effect size is stunning.

Table 2.4: Results and Heterogeneity

<i>Panel A: Main Results</i>	
Propensity to Open a Case	-3.65*
	[1.96]
Propensity to Remove	-13.20
	[8.57]
Propensity to Remove, excluding year 0	-18.36***
	[7.03]
<i>Panel B: Heterogeneity Analysis of Opening a Case</i>	
Investigator Experience: 10-64	0.39
	[2.40]
65+	-6.35***
	[2.11]
Referred for: <i>Neglect</i>	-3.85*
	[2.13]
<i>Abuse</i>	-2.31
	[1.96]
Age of child: < 4 years	-7.88**
	[3.97]
≥ 4 years	-2.33
	[2.18]
Race: <i>Black American</i>	-5.43**
	[2.72]
<i>All others</i>	-2.27
	[2.89]
Ever on Food Stamps Prior: <i>Yes</i>	-6.23**
	[3.09]
<i>No</i>	-0.76
	[3.03]
Predicted Likelihood of Death: <i>Above Median</i>	-6.09*
	[3.15]
<i>Below Median</i>	-1.97
	[1.73]

Notes: p<0.1, ** p<0.05, *** p<0.01. Standard errors are clustered two-way for investigators and individual children. This table presents in Panel A the main results from the regression presented in equation (2.1) in the first row. The second row looks at the effects of being assigned to an investigator more likely to remove holding constant the investigator's propensity to remove. The third row runs then the same specification but excluding medically examined deaths that occur in the month of the referral or 11 months following. Panel B then presents heterogeneity analyses for the results in the first row of Panel A, restricting each time the sample of children as indicated on the left: assigned to investigators overseeing 10-64 investigations in this decade, assigned to investigators investigating more than 65 investigations, initially referred for neglect, initially referred for abuse, children younger than 4 years, children 4 years and above, children identified as Black American, all other children, children that received food stamps at some point before the referral, children that did not receive food stamps prior to the referral, children that had a predicted likelihood of death above and below median as based on predictions from Appendix Table B.2.

Holding constant the propensity to open a case of your investigator, the effect on child mortality of being assigned to an investigator that is more likely to remove a child is not statistically significant, but the coefficient is particularly large (-13.20). The size of the standard error suggests that I lack power to be able to detect the effects of removing children on child medically examined deaths. An alternative specification excluding the year of the referral seems to suggest however that removals do seem lead to a highly statistically significant reduction in child medically examined deaths in the longer run.² The size of the coefficient size is colossal: -18.36. When converting these estimates using 2SLS, the effect of removing a child is even more astonishing: -52.2‰ reduction in child deaths that were examined by a medical examiner. It is important however to remember, that children are precisely removed from their home when the child's safety cannot be ascertained at home. In that light, one would thus expect effects of removals to be especially large on child deaths reported to the medical examiner, which they seem to be.

Remember that a case must be opened to remove a child. So the effects of child removals are also included in the effects of opening a child welfare case presented in the first row of Table 2.4. One might thus wonder whether opening a case without initially removing a child would also have an effect on child deaths. Results not presented here seem to suggest so (point estimate of -2.5‰), though the effect is no longer statistically significant. Given that I thus seem to run into issues with statistical power to separate the effects of removals from other child welfare interventions, I will primarily focus on the propensity to open a case for the rest of this paper.

A graphical analysis also provides support for these effects in Figures 2.1 and 2.2. Figures 2.1a and 2.2a plot the cumulative number of child deaths that were medically examined

²The motivation for running this analysis was based on seeing Figure 2.2 (a) where effects of removals seemed to take a few years before they were noticeable. It is true that some children who are removed will never be reunified with their family, and so one could expect the cumulative benefits to increase over time. Yet, the vast majority of children who are removed from their home will be reunified with their family, so I would have on the contrary expected the effects of removing a child to be strongest in the first year for two reasons. First, when children are removed it is because they are in immediate danger. Second, most removals are short and only last a few months and I would have expected the reduction in child deaths to be strongest while the child was removed.

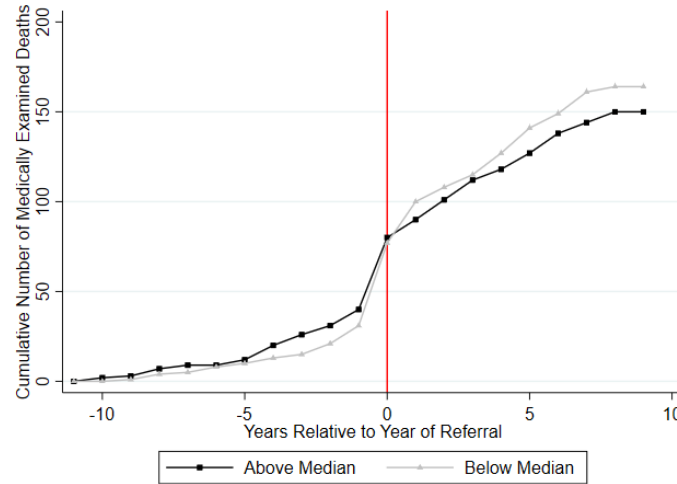
relative to the year of the referral for investigators that oversaw at least 50 investigations. The trends are drawn separately for two groups of investigators: those with above median propensity, and those with below median propensity.³ Figure 2.1 is for the propensity to open a case and Figure 2.2 for the propensity to remove. These simple descriptive raw graphs tell a compelling story. Investigators with above median propensity and below median propensity seem to be seeing approximately similar trends in number of medically examined deaths, with possibly investigators above median being slightly more likely to be assigned to families that have had children with medically examined deaths. However after the referral, this reverses. In Figure 2.1a, both groups of investigators see a similar increase of medically examined deaths, but by the following year the group assigned to investigators more likely to open a case has fewer medically examined deaths. This gap persists over time suggesting that opening a child welfare case leads to fewer child medically examined deaths: opening a case seems to be truly preventing deaths, and not just delaying them by a few months or years. In Figure 2.2a for child removals, the story seems similar before the referral, but the reversal in the gap across the two groups of investigators only seems to appear a few years after the referral (2-3 years after the referral). This seems consistent with a story where removals may prevent child deaths in the longer run.

Given these observed slight gaps in number of medically examined deaths for investigators with different propensities, I thus run an additional specification developed in Grimon (2020). This methodology no longer relies on perfect random assignment of investigators, but instead relies on children assigned to different investigators being on parallel trends. At the bottom of the Figures 2.1b and 2.2b, I present the p-value of the F-test testing for parallel trends before the referral. The p-values for both tests are quite high – 0.906 and 0.471 respectively – suggesting that parallel trends indeed do seem to hold. The years after the referral for their part indicate that children assigned to investigators more likely to get child welfare involved are less likely to have had a medically examined death. This

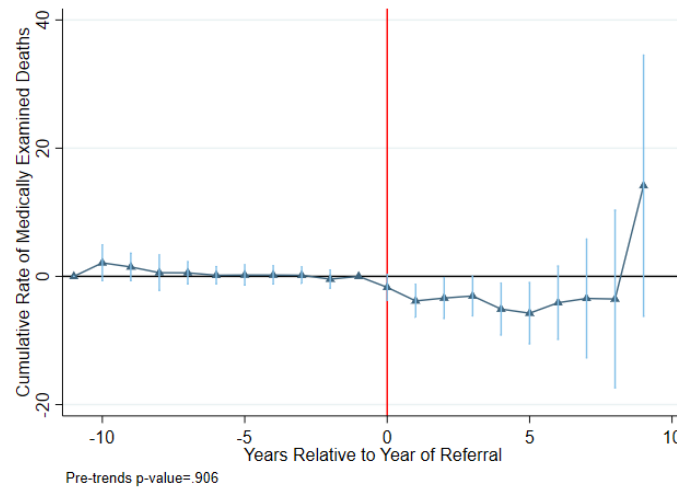
³Given large differences across offices and years, I split investigators depending on whether they were above median propensity after accounting for these differences by office-year.

Figure 2.1: *Effects of Having a Child Welfare Case Opened on Deaths Examined by a Medical Examiner*

(a) *Cumulative number by median of investigator propensity to open a Case*

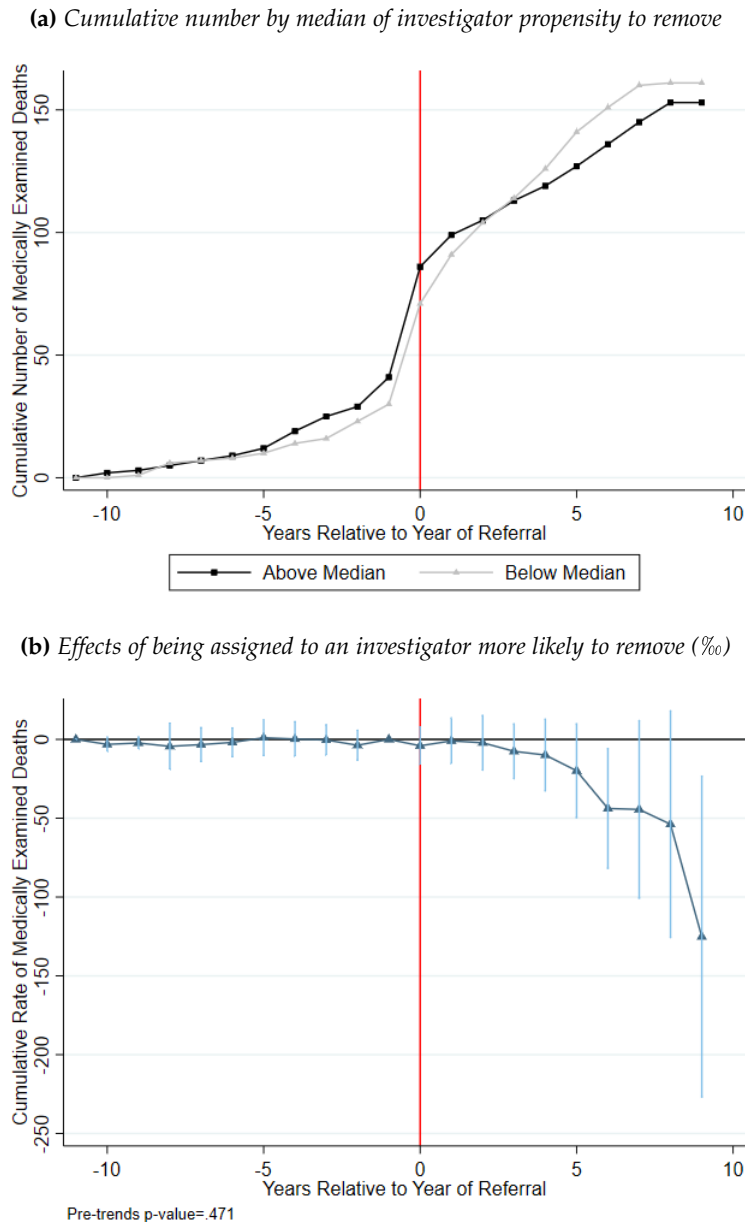


(b) *Effects of being assigned to an investigator more likely to open a case (‰)*



Notes: A child can be listed on a referral after their death, if their siblings are referred to child welfare. The red line in Figure 1a and Figure 1b is year 0 which is the month of the referral and 11 months following. Year 1 is the 12 months following that, and Year -1 the twelve months prior to the referral, etc. Figure 1a presents the number of children who died and had a medical examination at death. The black line with squares is the cumulative number of medically examined deaths for children assigned to an investigator with above median propensity to open a child welfare case relative other investigators working in that office year. The light gray line with triangles is for children assigned to investigators with below median propensity to open a case. Figure 1b plots the effect of being assigned to an investigator with measured propensity of 1, relative to the observed gap in death rates in the year prior the referral (year -1). These results also control for differences in death rates by office referral-year for each relative year before and after the referral. Coefficients are reported along with 90% confidence intervals, with two-way clustering for investigators and children. In the bottom left of Figure 1b is reported the p-value of the F-test that all the coefficients prior to year 0 are equal to zero (validation test for parallel trends).

Figure 2.2: Effects of Being Removed from Home on Deaths examined by a Medical Examiner



Notes: Figure 2 is similar to Figure 1 except that it examines the effects of being assigned to an investigator more likely to *remove* children, holding constant the investigator's propensity to open a case. A child can be listed on a referral after their death, if their siblings are referred to child welfare. The red line in Figures 2a and 2b is the month of the referral and 11 months following. Figure 2a plots the number of medically examined deaths for children assigned to an investigator with above (in black with squares) and below (gray with triangles) median propensity to remove a child, relative to that office-year. Figure 2b plots the effect of being assigned to an investigator with a (hypothetical) measured propensity to remove of 1, relative to the observed gap in death rates in the year prior to the referral (year -1). These results also control for your investigator's propensity to open a case, for office referral-year fixed effects for each year relative before and after the referral. Coefficients are reported along with 90% confidence intervals, with two-way clustering for investigators and children. In the bottom left of Figure 2b is reported the p-value of the F-test that all the coefficients prior to year 0 are equal to zero (validation test for parallel trends).

seems rather persistent for the propensity to open a case, as the later years involve very few clusters (only 52 investigators 9 years out for example). Again as in Figure 2.2a, the effects of being assigned to an investigator who is more likely to remove seem to accentuate over time and especially several years out in Figure 2.2b. The effects however for being assigned to an investigator more likely to remove are rather noisy (note the difference in scale for the vertical axis relative to Figure 2.1b), suggesting that I may not have sufficient power to detect the effect of being assigned to an investigator more likely to remove.

Overall, these tables and figures provide compelling evidence that child welfare involvement leads to fewer medically examined deaths for children. This raises the natural question of who benefits from these effects and what this may mean for mortality inequality.

2.4.2 Relationship to Inequality

In Panel B of Table 2.4, I present heterogeneity of these results by investigator experience, type of referral and several child characteristics. The first feature to note is that all of the effects in reductions of child medically examined deaths come from more experienced investigators who have investigated 65 or more households. Experienced investigators are the ones effective at reducing child mortality. Of all the heterogeneous effects presented in this table, this is the only statistically significant difference. Additional analyses not presented here suggest that although experienced investigators are possibly assigned to children at higher predicted risk of dying, such differences cannot explain why they are so effective at reducing child deaths. Experienced investigators do not seem to be opening more cases for children at higher predicted risk of dying either. Future ethnographic and qualitative work may thus want to explore why for the same families, experienced investigators are most effective at reducing child deaths when they recommend a case opening.

Effects are statistically significant for the subsample of cases that were initially referred for neglect, but the coefficient for the subsample of cases that were initially referred for abuse is also quite large and negative. The sample of cases referred for abuse is substantially

smaller, so statistical power is lower in that subsample. Reductions in medically examined deaths thus do not seem to be more pronounced in the sample initially referred for abuse as opposed to neglect.

Among investigated children, young children were the ones most at risk of having a medically examined death, as presented in Table 2.2. Correspondingly, it seems that the effects of being assigned to an investigator who is more likely to open a case reduces child mortality especially in this young group of children (<4 years old): a coefficient of -7.88 relative to -2.33. Effects are also particularly strong among children with a higher predicted likelihood of death (last rows in Table 2.2).

With regards to the effects of child welfare interventions on inequality, the two most interesting tests for heterogeneity are however for race and being a food stamp beneficiary prior to the referral. Effects of opening a case seem particularly strong among Black American children: -5.43 relative to -2.27, although this difference is not statistically significant. Similarly children who received food stamps at some point prior to the referral seem to be the group of children most likely to see a reduction in rates of medically examined death: -6.23 vs -0.76. Again, this difference is not statistically significant but nonetheless provides suggestive evidence that child welfare could be particularly effective at reducing child mortality among Black American children and children from disadvantaged households.

A simple probabilistic model may be helpful at this stage to help contextualize what these effects would translate into in terms of racial gaps in mortality rates. The change induced in black (B) death rate by child welfare (CW) is the percent of blacks who interact with child welfare $\mathbb{P}(CW|B)$ multiplied by the effect that child welfare has in reducing deaths:

$$\Delta \mathbb{P}(\text{death}|B) = \mathbb{P}(CW|B) \cdot \mathbb{P}(\text{death}|CW, B)$$

Here the effects of opening a child welfare case were estimated for the group of children where investigators would disagree as to whether or not to open a case: the estimated effects are a local average treatment effect (LATE). In column (3) of Table 2.2, I present

which groups are more likely to be compliers so for who the estimated effects are estimated for. Black Americans here seem slightly less likely to be a complier (-4%), but overall no particular group seems to be particularly more or less likely to be a complier except for Hispanics.

However, looking more generally in the population, 12,440 unique Black American children will have a child welfare case opened over this decade in this county, relative to 10,475 non-Black children. According US Census data, there were 45,552 Black American children in this county in 2010 and 196,111 non-Black children. Assuming away mobility concerns, that means that respectively 27.3% of Black and 5.3% of non-Black children will have a child welfare case opened over the course of this decade.

Although, the effect of child welfare on Black children's medically examined deaths $\mathbb{P}(death|CW, B)$ seems qualitatively different from the effect for non-Black children, as shown in Table 2.4, this difference is not statistically significant. So, using the main estimates from Panel A in Table 2.4 and the first stage estimates from Table 2.3 to extrapolate the local average treatment effects to both compliers and always takers⁴, the change in child medically examined deaths for Blacks as a result of opening a child welfare case would be $0.273 \cdot \frac{-3.65}{0.697}$ and for non-Blacks $0.053 \cdot \frac{-3.65}{0.697}$, which means that the gap in medically examined deaths would be 1.15‰ higher between Black children and other children without child welfare intervention. The death rate among all Black American children is 0.83‰ and for non-Blacks 0.41‰ so the racial mortality gap among children would be $1.15 / (0.83 - 0.41) = 274\%$ higher than it is without child welfare involvement. Obviously this number is staggering, and warrants caution as the estimates presented here are only based on one county and extrapolated from a local average treatment effect. They do however highlight the importance that child welfare seems to play in reducing racial mortality differences among children.

2.4.3 Robustness Checks

⁴The lowest investigator propensity to open a case in my sample for neglect cases is 0.061 so it suggests that only 6% of cases are always takers.

Table 2.5: Robustness to Alternative Specifications

	(1)
1. Base Specification	-3.65*
	[1.96]
<i>Permutation Test P-value</i>	.021
2. Controlling for Propensity to Die	-4.13**
	[1.96]
3. Effects After Referral, controlling for before	-3.26
	[2.12]
4. Residualized Propensity	-3.98**
	[1.98]
5. Single Investigator Assignment	-3.00
	[1.98]
6. No Recently Closed Investigation	-7.08**
	[3.19]
7. First Referral Only	-4.44*
	[2.42]
8. Investigator involved only during investigation	-3.92
	[2.81]
9. Addressing Lag in Death Reporting	-3.71*
	[1.95]
10. Excluding Referral Month	-2.70
	[1.79]

Notes: $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered two-way for investigators and individual children. The first row presents the base specification from Table 2.4 Panel A, with in addition a one-sided permutation test p-value. This p-value is obtained by generating 1,000 different datasets where I randomly reassign investigators to families, and then by calculating the percent of instances where I observe a reduction in mortality in each of these datasets as large as the one observed in the true data. The second specification in this table controls for the predicted propensity to die. The third specification uses data on medically examined deaths prior to the referral to examine how the effects of being assigned to an investigator more likely to open a case changed after the referral relative to before. Note that the office-year fixed effects in this specification are allowed to be different before and after the referral, so as to only take into account variation in the propensity that is net of office-year differences. The fourth specification uses a propensity instrument residualized on all the variables from the balance tests in Appendix Table B.3. The fifth row restricts to the sample of referrals that were assigned only one investigator. The sixth row restricts to the set of children that did not have an investigation closed in the 90 days prior to the referral. The seventh row restricts the sample to the first time a child is referred and assigned to an investigator seeing at least 50 investigations. The eighth row restricts to the sample of referrals where the investigator was only involved with the family at the investigation stage (before investigators became assigned to families if a case was opened). The ninth specification drops the last few months of data where coverage of medically examined deaths may not be complete. The tenth specification excludes medically examined deaths that occurred in the month of the referral.

In Table 2.5, I present robustness checks for the effects of being assigned to an investigator who opens more child welfare cases.

The first row presents the base specification presented in Panel A of Table 2.4. Below it I present the first robustness check: a permutation test. Instead of relying on the analytical standard errors, I randomly reassign investigators to families generating 1,000 similar datasets each with different investigator assignments. Then, I calculate the percent of datasets where I observed a reduction in mortality as large as the one I observed in the true data: only 2.1% of cases. The permutation test p-value thus provides additional evidence that this effect is robust.

In the second row I attempt to address the concern that investigators with higher propensity are more likely to be assigned to children who are more likely to die. Indeed, when I control for the predicted likelihood of having a medically examined death, the effect of being assigned to an investigator who opens more cases is even larger and more statistically significant: -4.13 instead of -3.65. Of course, it is hard to predict whether a child will have a medically examined death (the pseudo- R^2 of the probit regression in Appendix Table B.2 is low), but these results provide comforting evidence that these effects are not driven by the imperfect random assignment. They also give an indication as to the potential magnitude of the bias from imperfect random assignment of investigators.

The third row controls for the gaps observed in medically examined deaths before the referral for different investigators, and presents the effect of being assigned to an investigator more likely to open cases in any of the years after the referral (similar to Figure 2.1b). The magnitude of the coefficient is slightly smaller (-3.26) and the standard errors slightly larger so the effect is no longer statistically significant. The results nonetheless still seem consistent in terms of magnitude and direction of the effect.

In the fourth row, I address imbalance in assignment of investigators by using a propensity instrument which is residualized on individual and referral characteristics, as well as all the other variables included in the balance test (Appendix Table B.3). Constructing the propensity in this way means that the instrumental variable is orthogonal to individual characteristics: the effects of the residualized propensity instrument are net of imperfect random assignment. The effect size is slightly larger (-3.98) but comparable to the estimate

in the base specification.

In the fifth row, I then restrict the sample to referrals where there was only one investigator assigned (approximately 9% of cases have more than one investigator assignment), since random assignment is more likely to hold in this sample. This reduces the point estimate slightly (-3.00) and so the effect is no longer statistically significant, but the magnitude is still consistent.

In the sixth row, I restrict to the three-quarters of the sample where the child did not have an investigation that was closed within the 3 months prior to the referral. Families may indeed be reassigned to the same investigator if it has been less than 3 months since the close of a previous investigation, which could be yet another concern relating to imperfect random assignment. The effect size almost doubles (-7.08)! Effects are thus particularly pronounced for the sample of children that were not investigated recently.

In the seventh row, I restrict to the first investigated referral for each child to address potential concerns that the same child death may be associated with multiple referrals, and the magnitude of the effect seems even larger (-4.44). Sample size is reduced so the standard errors increase, but the effect is still statistically significant.

In the eighth row, I restrict to the sample of referrals where the investigator was only involved with the family during the investigation. Indeed, in some of later years for certain offices (a fifth of my sample), investigators became the social worker of the families they investigated if the decision was made to open a case. This is to address potential concerns in the exclusion restriction. The magnitude of the coefficient is slightly larger than in the main specification (-3.92), but the standard errors increase by a lot so the effect is no longer statistically significant. I nonetheless still interpret this as evidence that the effect is consistent in this specification since the magnitude of the effect size stays about the same.

In the ninth row, I address a potential concern which is that information about whether a child had a medically examined death may not be complete in the last few months of 2018. I restrict my sample to observations where information on medical examinations is complete, and the effect again seems consistent (-3.71) and is statistically significant.

In the tenth row, I present a final test which excludes any medical examinations that occurred in the month of the referral. The effect here is reduced in magnitude (-2.70) and no longer statistically significant, though the direction is still consistent with previous results. This suggests that opening a child welfare case seems to have effects starting in the month of the referral, though the statistically insignificant point estimate suggests that the effect is not limited to the first month.

As the number of investigators is rather small (226), I also do a jackknife estimation, where I re-run the base specification excluding each time one investigator. The effects of opening a child welfare case are statistically significant at the 10% level for all jackknife samples and 89% of the jackknife samples have a statistically significant effect at the 5% significance level. My results thus do not seem to be driven by any one investigator having particularly large effects.

Overall, these robustness checks provide compelling evidence that opening a child welfare case reduces the number of medically examined deaths among children.

2.5 Conclusion

The press often calls attention to children who died from maltreatment, portraying in so doing the child protection system under a negative light. It is much harder to spot the cases of children that would have died, but did not because child welfare intervened. Those cases are less “visible”. Using causal identification methods, this paper shows that child welfare in fact saves quite a few children’s lives by opening child welfare cases, and possibly also by removing children. In doing so, child welfare reduces inequality in child mortality.

Of course, this paper warrants additional research as this is only the effects of one leading child welfare agency in the US. Future research will want to explore how widespread such effects are across different child welfare agencies. It also highlights the need for understanding better what services or help being provided leads to such reductions in medically examined deaths. The sample size of this one county was too small to explore what specific programs and interventions were most effective at reducing deaths.

As a consequence of this work, evaluations of child welfare programs and interventions ought to pay attention to differential attrition due to mortality. If indeed some of the most disadvantaged children are saved from death when provided with services, then evaluations of child welfare programs could be biased downwards.

Chapter 3

The Effect of Algorithmic Tools on Child Welfare Decision Making: A Partway Evaluation¹

3.1 Introduction

Eight million children in the U.S. are referred to child protective services every year, but fewer than half of these are visited by child welfare. Time-constrained agents often must decide who to visit in a handful of minutes (Kelly et al., 2019). Accordingly, child welfare practitioners from around the world have expressed interest in developing algorithmic tools to guide such consequential decisions.

In the context of the criminal justice system, seminal work by Kleinberg et al. (2018) has indeed established the hypothetical potential of improving bail decisions using machine learning to reduce both crime and jailing rates. Contrary to this work, however, machine learning algorithms are typically implemented as an aid to human decision making: the human decision is not replaced by an algorithm's prediction, but rather complemented by it. This raises the natural question of whether the benefits of machine learning tools actually

¹Co-authored with Christopher Mills

carry over when the end decision is made by a human with the help of the tool.

The empirical evidence is rather mixed. On the one hand, decision makers often override recommendations in ways that introduce new inefficiencies. Stevenson & Doleac (2019) find that the introduction of a risk score in felony sentencing did not reduce recidivism rates in part because judges were likely to override risk score recommendations for young, high-risk defendants. Judges have also been shown to override risk scores in ways that may systematically favor or disadvantage particular groups, for example being less lenient toward Black defendants in bail decisions (Albright, 2019). On the other hand, De-Arteaga, Fogliato, & Chouldechova (2020) show that decision makers deviate from algorithmic recommendations in the case of algorithm glitches, highlighting a potential complementarity between algorithmic tools and human oversight.

This paper is among the first to assess the effectiveness of a machine learning decision aid tool in any domain using a randomized control trial. Specifically, it evaluates the effectiveness of a machine learning based risk score predicting child removals which was developed by leading experts in the development of machine learning tools for child welfare. Relying on the roll-out of this tool in the child welfare agency of an American county, we randomly assign for which families child welfare could see a score when deciding whether to visit a family. We then assess the effects this has on outcomes for the children and child welfare involvement. The results of this paper are based on the first five months of the roll-out of this trial and are entirely preliminary given the small sample size and lack of follow-up.

Our results provide suggestive evidence that the decision tool improves targeting of child welfare visits. We find that workers have consulted the tool for at least two-thirds of instances where the score is available. We also find that, even when the score is not seen, workers are more likely to investigate, open cases for, and provide services to families with a high risk score, helping to validate the risk model and showing that workers are sensitive to underlying risk. Even still, showing a score increases the number of referrals with a founded investigation for high risk cases, and decreases the likelihood of being called back

in to child welfare for the families where child welfare decided not to intervene.

To the best of our knowledge, this is the first randomized control trial to document potentially beneficial effects of an algorithmic tool implemented in practice on the partner agency’s target outcomes. This is our main contribution to the research literature. Our study is most similar to a recent study by Fitzpatrick & Wildeman (2020) who find increases in investigations for high scores and decreases in investigations for low scores, but are unable to validate risk scores on other meaningful child welfare outcomes due to constraints in experimental design. Our intervention takes place in a high-stakes real-life setting, in contrast to lab-like experiments by Green & Chen (2019) on Amazon Mechanical Turk. Imai et al. (2020) find limited, mixed effects when evaluating one judge’s use of the Public Safety Safety using an randomized control trial. The decision aid tool in our context, in contrast, has greater independent support from the scientific community and our evaluation studies decision making across many individuals.

The rest of this paper proceeds as follows. Section 1 provides background on the child welfare system in our context, reviews the literature regarding machine learning tools in child welfare, and describes the tool being evaluated in this study. Section 2 describes our empirical design, descriptive statistics of our sample, and provides evidence that the randomization was effective in our context. The third and final section includes preliminary results from the ongoing evaluation.

3.2 Child Welfare and Machine Learning for Prediction

Child welfare tries to prevent child maltreatment. Its aim is thus closely related to a prediction problem: who are the children we need to intervene for (Russell, 2015)? Machine learning tools are precisely reputable for predictions (Kleinberg et al., 2015). In fact, among child welfare practitioners, these methods are commonly referred to as “predictive analytics”.² It thus seems natural that machine learning tools would be implemented in this

²The child welfare literature also refers to linear regression models as predictive analytics whereas the economics literature distinguishes them from machine learning methods.

context. This section will thus begin with a brief overview of how child welfare works in this setting, followed by a detailed review of the literature on machine learning in child welfare, and finally present the specific tool which is the focus of this study.

3.2.1 Child Welfare in the Intervention County

Child welfare – also referred to as child protective services (CPS) – aims to keep children safe from abuse and neglect. In the United States, the way child welfare learns about children at risk is through individuals who typically call a local hotline to report instances where they suspect a child may be abused or neglected. Indeed some professionals such as doctors, teachers, or police are mandated by law to report to child welfare any instances where a child may be abused or neglected. Beyond mandated reporters, any individual who notices behavior toward children that could be suspect is encouraged to report to child welfare. Calls made to child welfare are called referrals.

Among these referrals, only a subset will be checked out in person by child welfare. A visit by child welfare is called a screen in. In this county, screen ins can be of two types: a formal investigation, or a less formal visit. Referrals that are not screened in are called “screened out” and filed for future reference by the child welfare agency. Screened out families will have no direct interaction with child welfare as a result of this referral. In most cases, these screened out families will not even know that a report to child welfare was made.

The decision to screen in or out a family can be made at two stages. A decision can be made immediately when a call is received. This typically concerns cases which require immediate intervention, for example if a child is alleged to be in imminent danger. The family will thus be immediately screened in and someone from child welfare will be sent to visit this home within a few hours regardless of time of day.

Most decisions of whether or not to screen in a family in this county are however made later than at call intake. Non-urgent referrals in this county are reviewed by teams of social workers and caseworkers each led by a supervisor. These teams meet every workday

morning to review referrals that arrived since their last meeting, and there are typically three different teams each reviewing a subset of cases at the same time. For example, referrals arriving on Thursday afternoon would be reviewed by a team on Friday morning. Referrals arriving after the Friday meeting and during the weekend would be reviewed on Monday morning. Each referral would only be sent to one of the teams.

After reviewing and discussing the details of the allegations listed on the referral, checking for past family involvement with child welfare, and checking for alleged perpetrators' criminal records in the state courts system, each team of workers comes to a decision regarding whether or not to screen in the referrals they were assigned. When a team decides to screen in a family, they also decide if they are going to screen it in for assessment on one of two tracks: a low risk track or a high risk track. The low risk track is part of a differential response framework by child welfare, where the child welfare involvement is lower and where findings of the assessment do not need to be recorded. Families screened in on low risk tracks can be switched to a high risk track within the first 45 days, but the reverse is not true. At the same time that this team of child welfare workers decides what track to send a screened in family on, they also determine how fast to send someone to visit this family: immediately, 3 day response, or 5 day response.

3.2.2 Review of Existing Work

Some early work showed the potential of neural networks to retrospectively predict abuse of children among a sample of investigated children using a few basic characteristics (Zandi, 2001; Schwartz et al., 2004). Flaherty & Patterson (2003) used them to predict repeat substantiated child physical abuse. Sledjeski et al. (2008) instead used classification and regression trees (CART) to predict repeat referrals to child welfare for a small subset of investigated families. Slightly more recently Jolley (2012) used a combination of CART and random forest analysis with neural networks to predict child likelihood of repeat child welfare involvement. Of lesser interest for this study but concurrent to this early literature on the topic, machine learning methods were also applied in child welfare to better understand

how social workers weigh factors in their decisions (Marshall & English, 2000; Johnson et al., 2002).

It was only several years later that child welfare agencies expressed interest in implementing such tools to improve their practice. Packard (2016) provides a good – though now slightly outdated – overview of such attempts. Predictive risk tools in child welfare were implemented either by private companies, or by academic researchers.

Tools developed by private firms are proprietary which makes it hard to assess the quality of the methodology developed by for-profit vendors. For example, the nonprofit firm Eckerd Connects partnered with a private for-profit company called Mindshare to develop a tool predicting child death from maltreatment in Hillsborough county, Florida (*Eckerd Rapid Safety Feedback*, 2021). It is unclear whether any machine learning methodologies were used in this context. What is clear is that it resulted in the development of a 14 point checklist of items which determine the risk of a child: the resulting output of this methodology seems far away from the richness of machine learning tool predictions... This tool was later implemented in other states (currently in Indiana, Maine, New Hampshire, Ohio and Oklahoma), but with various degrees of success. Indeed, Illinois resigned its contract with these partners as the tool was not flagging as high risk the children who were dying from maltreatment (Jackson & Marx, 2017). To the best of our knowledge, no external independent review has been conducted on this tool. In 2014, the statistical analytics company SAS also developed a model to predict child abuse for the Los Angeles child welfare agency using historical data. However, given that the tool was flagging as high risk children so many children who were not high risk, in addition to the lack of transparency about how the algorithm worked, the Los Angeles child welfare agency decided to never roll it out (Nash, 2016).

Relative to this work by for-profit vendors which lacked transparency, tools developed by academic researchers Professors Rhema Vaithianathan, Emily Putnam-Hornstein and the Center for social data analytics (CSDA) of the New Zealand Auckland University of Technology have received noteworthy attention (Hurley, 2018). After much work using

retrospective data highlighting the potential of these tools (Vaithianathan et al., 2012; Cuccaro-Alamin et al., 2017), their first major implementation of a machine learning tool for a day-to-day use in a child welfare agency was rolled out for Allegheny County in Pennsylvania (Pittsburgh) in 2016. The currently running version of the tool in Allegheny County assesses risk for a child to be removed in the two years following a referral to guide social workers deciding which families to investigate. The tool is based on a LASSO regression model using comprehensive administrative data. In building the tool, several alternative, more unstructured machine learning methodologies were explored including random forests, XG boost and SVM. Given that these led to only minor improvements, the LASSO model was chosen over these alternatives. The tool was externally validated using hospital data (Vaithianathan et al., 2020), and was also externally evaluated using interrupted time series analyses and generalized linear models (see details of the methodology and evaluations in the report of DHS, 2019). As a result of this successful collaboration, the researchers developed another tool for Douglas County in Colorado which was rolled out in 2019 to similarly predict children at high risk of being removed. An evaluation of the roll-out of this tool in Douglas County is also currently being conducted, with midway results (Fitzpatrick & Wildeman, 2020).

3.2.3 The Decision Aid Tool in our Context

Evaluations in Allegheny County and Douglas County have provided interesting insights indicating better targeting of high risk families as assessed by the risk score, and even in the case of Allegheny County reductions in racial disparities for high risk families. Yet, neither have yet provided convincing answers regarding effects this tool has had on child wellbeing, on targeting of preventative services, and more generally the insights that have been drawn so far on human and machine learning interactions are still limited.

The decision aid tool used in this study was also developed by Rhema Vaithianathan and her colleagues at the CSDA. Like the two previous iterations of this tool in Colorado and Pennsylvania, the tool predicts a child's likelihood of being removed from home

by child welfare within the two years following the referral using a LASSO regularized logistic regression. Other more flexible and unstructured methodologies (notably random forests and XGBoost) were considered, but the LASSO model was preferred for running time purposes. The model was trained on a dataset of past child welfare referrals in this county and other neighboring counties using the same data system (close to half a million child-referrals). It uses several hundred features relating to demographics (excluding race and disability), past child welfare involvement of household members, as well as some case management data with social services. The initial choice of which variables to include for the model was made in agreement with local partners. The area under the receiver operating characteristic curve (AUC) of this model is 76.2% which indicates good discrimination.

Given tool users' unease with probability, the team designing the tool decided to bin the predictions from the model into 20 equally sized groups, as was previously done in Allegheny county and Douglas county. The lowest 5% of children thus score a 1 and the highest 5% score a 20 for the full sample on which the model was trained (so not just the county of this study).

In order to generate a score for a child, the child's referral needs to have been copied over into the state central data system. This happens automatically at the end of each work day: the process starts around 5pm but typically takes about 12 to 14 hours to complete. Cases that are *immediately* screened in will thus not have had the time to be copied over and have a score generated. Furthermore, cases arriving between approximately 5pm and 8am the next day will also be reviewed by a team before their information is copied over into the main data system, so they will not have a score generated. On certain days, the main data system is down or the copying over took longer than usual and so the team started reviewing the referrals before a score could be generated. Hence, scores will not be generated for every referral. In our setting, approximately a quarter of child-referrals remain unscored.

When a child is scored, for a random subset the score will then be made available to the teams of workers deciding which families to screen in. The score's sole purpose is to

guide the decision making of which families to screen in. The score is purposely not made available to any other worker at any other time apart from these team meetings. The teams of workers were however asked to make note of the score in the notes from their discussion, so it cannot be ruled out that a worker visiting a family for an assessment would know the score if he or she consulted the discussion notes or randomly happened to be in the meeting that discussed the referral.

A final point to note about the model and the scores it generates is that the likelihood of a child being removed goes up exponentially at the top of the score distribution. Appendix Figure C.1 is shown to the workers every time they look up the score for a referral to remind them that below a 15 the difference in likelihood of placement is not so high, but above that each additional score indicates a substantially increased likelihood of removal.

3.3 Experimental Design

3.3.1 A Randomized Control Trial

To assess the causal effects of this decision aid tool on child welfare decisions and child outcomes, we set up a randomized control trial. Randomization happens at the household level so that all children in the same household will either each have a score that is shown (treatment) or none of them will have a score that is shown (control). This design is unique relative to prior studies: the Allegheny evaluation was not a randomized control trial, and the Douglas evaluation randomized at the team-day level which means that the same child referred multiple times over a few days could have the score shown sometimes and other times not.

This unique design randomizing at the household level allows us to look at effects of showing a score beyond just this immediate referral (relative to Douglas county), and to do so more convincingly than using a pre-post design that could be confounded by time trends (relative to Allegheny county).

3.3.2 Empirical Specification

For all children referred in the sample time period, we first examine whether the team reviewing child i listed on its first scored referral r reported seeing a score for this referral (W_r):

$$W_r = \beta_0 + \beta_1 T_{i,r} + \beta_2 \text{Flag}_{i,r} + \epsilon_{i,r} \quad (3.1)$$

where $T_{i,r}$ indicates that child i on referral r was treated which is to say that the score was accessible to the team. $\text{Flag}_{i,r}$ is a flag indicating whether the child had an open case, an ongoing investigation or an investigation closed within the past 60 days. This variable is not essential in this specification, but is included for consistency with the second specification. Finally, the standard errors are clustered at the household level given that randomization is at the household level.

For other outcomes, the effect of seeing the score should be different depending on the level of the score. Statistical power is still rather limited as data collection is only approximately halfway. For that reason, and given that the score is most informative for high scores as seen in Appendix Figure C.1, we decided to run the specification by comparing the outcomes for children scoring in the top tercile relative to those scoring in the bottom two terciles. The top tercile corresponds to children scoring a score of 16 or above the first time they were referred.

For outcome Y_i of child i first listed and scored on referral r , we estimate:

$$Y_i = \beta_0 + \beta_1 T_{i,r} + \beta_2 \mathbb{1}(\text{Score}_{i,r} \geq 16) + \beta_3 T_{i,r} \cdot \mathbb{1}(\text{Score}_{i,r} \geq 16) + \beta_4 \text{Flag}_{i,r} + \epsilon_{i,r} \quad (3.2)$$

where $T_{i,r}$ indicates that the score was made accessible to the team during their discussion of referral r (treatment), $\mathbb{1}(\text{Score}_{i,r} \geq 16)$ indicates that the child i scored 16 or above on their first referral r , and $\text{Flag}_{i,r}$ controls for the fact that child welfare outcomes are different for referrals coming in that are likely related to an ongoing or recent episode examined by child welfare. Here again, since randomization is at the household level, standard errors are clustered at that level. Importantly outcomes Y_i can be of two types: they can relate to the

specific outcomes of that first referral r (length of discussion, whether it was investigated, results of the investigation) or it can relate to outcomes that occur after that referral. Those outcomes could be because of that specific referral or of another later referral, and include for example removals, case openings, provision of services, etc.

3.3.3 Sample

All the current data and outcomes come from the child welfare agency's administrative records. Any child referred to the child welfare agency of this county from November 2, 2020 to March 30, 2021 would be included in our sample.

In Panel A of Table 3.1 we present some descriptive statistics on the children being referred to child welfare in this county for all children referred in this county in the first column, and for all children for whom a score was generated in the second column. Children being referred are approximately gender-balanced, and predominantly white: at least 72.8% of children are white and the remainder are either black (2.7%) or information on race is missing (25.3%). At least a fifth of children identify as Hispanic (19.4% are identified as Hispanics, but since ethnicity is missing for 29.7% of the sample that number could be as high as half of the sample). Over the course of these five months, the average child is referred 1.5 times but the median child is referred only once. 2,534 children are referred in this time period, and of these 1,932 were scored at some point in this time frame.

Panel B of the same Table 3.1 presents some descriptive statistics on the referrals coming in to child welfare. On average there are 4.6 individuals listed on a referral with two of them being children. 73.7% of referrals coming in are from mandated reporters. Reporters indicate that the child is in present or imminent danger in only 2.5% of cases, and the three most frequent allegations are injurious environment (62.8% of referrals), physical abuse (21.9% of referrals) and parental substance abuse (16.3% of referrals). A given family can be referred with multiple allegations of course. From November 2020 to March 2021, child welfare was notified 2,091 times and scores could be generated in 1,580 of these referrals.

As can be seen by comparing the two columns of Table 3.1, children being scored do

Table 3.1: Descriptive Statistics

	All	Scored
<i>Panel A: Child Characteristics</i>		
Female (%)	51.8	52.2
White (%)	72.8	74.1
Black (%)	2.7	2.6
Missing Race (%)	25.3	24.3
Hispanic (%)	19.4	19.6
Missing Ethnicity (%)	29.7	29.7
Number of Times Referred	1.5	1.5
Number of Unique Children	2534	1932
<i>Panel B: Referral Characteristics</i>		
Number of Persons Listed on Referral	4.6	4.6
Number of Children	2.0	2.0
Mandated Reporter (%)	73.7	73.6
Child in Danger (%)	2.5	2.1
Alleged Injurious Environment (%)	62.8	64.7
Alleged Physical Abuse (%)	21.9	21.8
Alleged Parental Substance Abuse (%)	16.3	14.8
Number of Referrals	2091	1580

Notes: This table reports descriptive statistics on the full sample of children referred to child welfare in the county from November 2, 2020 to March 30, 2021. Panel A provides descriptive statistics for the children referred to child welfare by gender, race (Black, white, missing), ethnicity (Hispanic, missing), and the number of times the child was referred during the sample period. Column 1 provides details of all the children referred, and column 2 for the subset of children that were scored in this time period. In Panel B, we provide descriptive statistics on the referrals coming in: number of persons listed on the referral, number of children listed on the referral, whether the reporter was a mandated reporter, whether the child was reported to be in immediate or imminent danger, and the three most frequent allegations in this sample. The first column in Panel B provides details on all referrals, while the second referral provides details on the referrals where at least one child was scored.

not seem meaningfully different from the children that could not be scored. The rest of this paper will thus focus on the sample of children that could be scored.³

As referenced in Section 3.2.3, scores are generated to split the sample that was used to train the model in bins of 5%. Interestingly, in our setting, as shown in Figure 3.1, scores do

³Note that the sample of scored children ($N = 1932$) is slightly higher than the sample of children used for the analyses in the rest of the paper ($N = 1739$). That is because in the initial roll-out of the decision aid tool there was a issue in the storage of the treatment status for close to two hundred children. These children are excluded from the analyses for the remainder of the paper.

not seem uniformly distributed contrary to what one could have expected. In particular children are much more likely to receive a score of 0 and 20 and less likely than expected to receive a score between 2 and 14. Several features could explain this. First the training sample included several counties including the county examined for this study: this county could have a child risk distribution different from that of other counties (the current county only represented approximately 11% of cases in the training sample). Second, November 2020 to March 2021 is a period entirely within the covid epidemic whereas the model was trained on data prior to the epidemic. Our local agency has indicated multiple times that the type of the referrals they were receiving has changed substantially. Notably, they have indicated that teacher referrals to child welfare which tend to be low risk children have dramatically reduced relative to pre-covid times. Third, only three-quarters of referrals coming in are being scored in the roll-out of the tool as explained in Section 3.2.3 whereas the training dataset included all referrals. Finally, the training sample was based on a sample of cases over the course of a period of four years of child welfare referrals. High risk children are obviously much more likely to be re-referred often whereas low-risk children are more likely to be referred only once. Hence, one would expect an over-representation of high risk scores in a sample of referrals over 5 months.

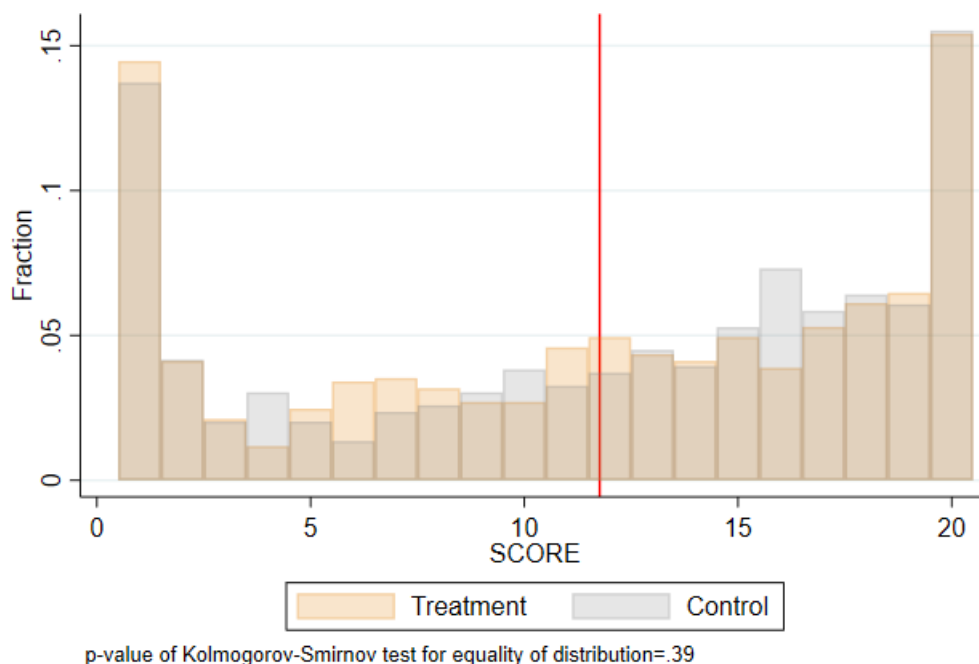
Given this observed non-uniform distribution, the top third of children are those with a score of 16 or above (Figure 3.1), which as mentioned in the previous section will be of particular interest in this study.

3.3.4 Balance Tests

In Figure 3.1, we present the distributions of scores for both treatment and control to assess balance in the score distributions by treatment status. Randomization seems to have been effective as children score similarly in the treatment and control group. The p-value of a Kolmogorov-Smirnov test that the two distributions are identical is reported at the bottom of the Figure and is quite high (0.39) confirming that the distributions seem the same.

In Table 3.2, we present a larger more complete balance test including all the demographic

Figure 3.1: Algorithm Score Distribution by Treatment Status



Notes: This graph presents the score distributions for children in treatment in orange (score available), and for children in control in grey (no score shown). The red bar is the mean score in the full sample. On the y-axis is the fraction of cases in each score bin. At the bottom of the figure we report the p-value from a Kolmogorov-Smirnov test for equality of the score distributions for the treated and control children.

characteristics and referral characteristics included in Table 3.1. We also include in this test indicators for what team the referral was sent to, how we tracked that this was a household (randomization step), and a few dummy variables indicating that the current referral may relate to an ongoing episode either already being investigated by child welfare or recently investigated by child welfare. The F-statistic and p-value at the bottom of this table clearly indicate that randomization seems to have been particularly effective (p-values= 0.436). In a test like this with 22 variables, one would expect that a few variables might show up individually as statistically significant just due to random chance (in this case the dummy variable for having an ongoing investigation). This thus provides comforting evidence that the experimental design was effective and that changes in the outcomes for treated children is entirely attributable precisely to showing a score.

Table 3.2: *Test of Randomized Assignment of Treatment Status*

	(1) Treatment
White	-0.05 [0.13]
Black	-0.15 [0.09]
Missing Race	-0.10 [0.14]
Hispanic	0.03 [0.05]
Missing Ethnicity	0.04 [0.06]
Female	0.02 [0.02]
Algorithm Score (out of 20)	-0.00 [0.00]
Number of Persons Listed on Referral	0.01 [0.01]
Number of Children	-0.02 [0.02]
Mandated Reporter	0.04 [0.04]
Child in Danger	-0.08 [0.12]
Alleged Injurious Environment	-0.01 [0.04]
Alleged Physical Abuse	0.08 [0.05]
Alleged Parental Substance Abuse	0.06 [0.05]
Team 1	-0.04 [0.04]
Team 2	-0.01 [0.05]
Randomization Step 2	-0.29 [0.19]
Randomization Step 3	-0.13 [0.09]
Randomization Step 4	-0.08 [0.10]
Prior Investigation Closed within 60 days	-0.03 [0.09]
Ongoing investigation	0.14* [0.08]
Case was Already Opened at Referral	-0.06 [0.05]
Number of Randomization Clusters	997

Continued on next page

Table 3.2 – Test of Randomized Assignment of Treatment Status (Continued from previous page)

F-statistic	1.020
p-value of F-test	0.436
Observations	1739

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. Given that households are only randomized once into treatment or control – the first time they are referred in this time period – this table restricts the sample to the first time that a given child is listed on a referral. This table reports the results of regressing the treatment status on individual demographic characteristics, reporter level characteristics, which team this referral was sent to, how we determined this was a household (randomization step), as well as whether the child had an investigation that was recently closed within 60 days, whether there was an ongoing investigation, and whether the child or someone else in the household had a case already opened with child welfare. At the bottom of the table, we also report the F-statistic from testing that all these coefficients are jointly equal to zero and the corresponding p-value.

3.4 Results

3.4.1 Process Outcomes

The first step in assessing whether the design was effective is to check whether caseworkers actually consulted the score when the score was made available. Results of this are presented for referrals from November 2020 to February 2021 in Table 3.3.

Table 3.3: Process Outcomes (First Stage)

	(1) Control Means	(2) Treatment (T)
Reported seeing score	0.003 (0.052)	0.653*** (0.028)
Reported not seeing a score	0.771* (0.420)	-0.651*** (0.031)
Score reported seen in this referral or a future referral	0.029 (0.168)	0.674*** (0.028)
N	726	1392
Number of clusters		795

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. A control for whether this first referral was part of an ongoing episode is included throughout. This table reports whether teams wrote down the score in the framework when children were in treatment. It also reports whether the red teams indicated not seeing a score. The last row in the panel also reports whether the team ever reported seeing the score for that child (either on the first referral or on a later referral).

When a score is made available to workers, they report seeing the score in approximately two-thirds of cases, as indicated by the first row coefficient of 0.653 in column 2 of Table 3.3 which is highly statistically significant. Conversely, when the score was not made available to workers they reported not seeing a score in 77.1% of cases (row 2 column 1). When the score was made available to workers, they report not seeing a score in $0.771 - 0.651 = 12\%$ of cases. That means that in $100\% - 65.3\% - 12\% = 22.7\%$ of cases we are not sure whether the score was actually seen or not based on the teams' comments from their discussions.

Interestingly, in Appendix Table C.1, we show that caseworkers vary in how often they report seeing the score depending on the type of allegations and the type of family structure (randomization step), and teams vary in their propensity to write and check scores. Future work will want to explore what can explain these differences.

In the last row of Table 3.3, we also present whether the score was seen on this referral or any other later referral. Remember that once a child has a treatment status, they will keep that treatment status. And indeed, the coefficient in column 2 does go up by a bit in row 3 relative to row 1, suggesting that sometimes the workers may not write it down or notice it on the first referral but they do on a later referral.⁴

These results are encouraging: they suggest that workers are being diligent in checking the score. Results in the next section will present the reduced form estimates for the score being *available* for a child, but this section indicates if we were to examine the effects of actually *seeing* the score the point estimates would be slightly higher than reported below.

3.4.2 Preliminary Results

The First Scored Referral

The first set of outcomes to examine are those relating to the child's first scored referral. Several changes can come out of the meeting because the team saw a score: their discussion

⁴The coefficient in column 1 of row 3 does go up a bit too because household structure is not fixed over time: children who were in separate households with different treatment statuses may at a later time point live within the same household. In those cases we re-randomize treatment to all children in the newly formed household.

length for the referral could change, and the score could affect the screen-in decision for marginal referrals. Indeed, there are clear legal guidelines as to when a family should be screened in and out, so it is inconceivable that teams would decide to screen in all high scoring families. It is only in cases where the teams are on the fence with regard to whether the family meets the guidelines for a screen in that seeing the score can make a difference. Those outcomes are reported in Panel A of Table 3.4.

Column 1 of Table 3.4 presents the means of the outcomes listed in the rows for the group of children who were in the control group and scored a score of 15 or lower (the bottom two thirds of the score distribution, T1 and T2). Columns 2-4 then report the estimates of equation (3.2). Columns 2-4 should respectively be interpreted as the general effects of seeing the score (column 2), how different are outcomes for the top third of scores relative to the bottom two thirds (column 3), how the effects of seeing the score are different for the top third of scores relative to the bottom two-thirds of scores. Finally, column 5 reports the frequency of each outcome to emphasize that certain outcomes are relatively rare and results are preliminary.

In the first row, we report the length of the text which summarizes the discussion. In this specification, the length of the text which can be viewed as a proxy for the length of the discussions does not seem particularly affected (note that for higher score thresholds of 18 and above it does seem to increase length of text however). Of course the standard errors are still quite large and cannot rule out meaningful changes in the length of the discussion transcription.

Seeing the score does not yet change whether a family is screened in, but the confidence intervals cannot yet rule out meaningfully sized effects. Results will need to be re-examined when more data is available.

In their discussion, teams also decide how fast they will go to visit a family: immediately, within 3 days or 5 days. However, only 3 referrals had a response time of less than 5 days among those scoring a 16 or above and in the control group, so results on response time are not presented at this stage given that sample size is still so small.

Table 3.4: Impacts of Score Availability for this Referral

	(1) Control Means T1-T2	(2) Treatment T	(3) Top Third T3	(4) Interaction T · T3	(5) Frequency
<i>Panel A: Decision Outcomes</i>					
Length of Text about Discussion	242.3 (257.6)	13.5 (27.2)	20.3 (31.5)	32.1 (41.8)	1340
Screened in	0.30 (0.46)	-0.033 (0.038)	0.053 (0.046)	-0.014 (0.062)	495
<i>Panel B: Investigation Results</i>					
Founded Investigation	0.019 (0.14)	-0.0026 (0.011)	0.0058 (0.014)	0.035 (0.022)	44
Inconclusive Investigation	0.0038 (0.062)	0.015 (0.010)	0.018 (0.011)	-0.028* (0.016)	25
Unfounded Investigation	0.023 (0.15)	-0.0027 (0.013)	0.051* (0.028)	-0.017 (0.033)	65
N	523	1739	1739	1739	1739
Number of clusters		997	997	997	

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. This table shows the effects of treatment on outcomes linked to the first scored referral for each child. A control for whether this first referral was part of an ongoing episode is included throughout. Panel A reports the effect on the length of text summarizing the discussion as measured in numbers of characters (note this outcome is currently only available for the subsample until late February like in Table 3.3), and whether this referral was investigated. Panel B reports the investigation results. Importantly, these investigation results are only available for the subset of investigations with allegations meeting the legal requirements for a *traditional* child welfare case. Column 1 reports means and the standard errors for those in control and scoring a 15 or less (the bottom two thirds of the score distribution, T1 and T2). Column 2 reports the effects of being treated. Column 3 reports the coefficient on the dummy indicating whether this individual scored a 16 or above (the top third of the score distribution, T3). Column 4 reports the effect of showing a score of 16 or above to caseworkers (T·T3). Column 5 reports the number of times that the outcome is not equal to 0, which is to say the frequency of each outcome.

In Panel B of Table 3.4, we present the results of the investigation if there was one. Remember that investigation results only need to be reported for the small set of screened in referrals that are on the high risk track: that explains why the percent of investigation results is so low relative to the screen-in rate in column 1. Furthermore, referrals that were investigated in February and March would likely not have investigation results yet. Given the frequency of each of these outcomes of the investigation, much caution needs to be kept with regards to the direction and the magnitude of these results. However, the point estimates seem to suggest that showing a high score increases the child's likelihood that the allegations were founded (insignificant point estimate of 0.035) and decreases the likelihood that the investigation was inconclusive. This may suggest either better targeting of investigations or additional diligence in the investigation when the child scored high. This result will need to be reassessed once more data are available. Indeed the percent of founded investigations for children in the top third of the score distribution but where the score was not shown is surprisingly low (column 3).

Of course, these are only the immediate short run outcomes of this referral. They raise the question of what happens then to child welfare's response as a result of the referral.

Outcomes after the First Scored Referral

Table 3.5 we thus report outcomes for what happens after this referral. Columns in this Table are to be interpreted similarly to what was explained in the previous section regarding Table 3.4.

In the first two rows of Table 3.5, we report outcomes that were included in the previous tables: screen in and founded investigation, except that this time instead of looking at just the first scored referral, we are also examining outcomes for all later referrals to see if the child was ever screened in or ever had a founded investigation. Two results here stand out as statistically significant. The first is that children with a high score are generally much more likely to be screened in: teams of workers are thus indeed deciding to visit more often families where the risk is higher! Indeed, for children where the score was not shown, the

Table 3.5: Longer-run Outcomes

	(1) Control Means T1-T2	(2) Treatment T	(3) Top Third T3	(4) Interaction T · T3	(5) Frequency
Ever screened in	0.37 (0.48)	-0.036 (0.041)	0.14*** (0.047)	-0.050 (0.066)	656
Ever founded investigation	0.023 (0.15)	-0.0025 (0.012)	0.017 (0.016)	0.047* (0.027)	63
Ever case opened after	0.31 (0.46)	-0.041 (0.038)	0.068 (0.045)	-0.025 (0.059)	498
Differential response case opened	0.30 (0.46)	-0.055 (0.038)	0.061 (0.044)	-0.028 (0.057)	469
Traditional case opened	0.10 (0.30)	-0.034 (0.024)	0.087** (0.034)	0.038 (0.045)	212
Prevention case opened	0.046 (0.21)	-0.0080 (0.016)	0.046* (0.026)	0.022 (0.031)	103
Any services received	0.040 (0.20)	0.010 (0.013)	0.085*** (0.023)	0.030 (0.037)	177
N	523	1739	1739	1739	1739
Number of clusters		997	997	997	

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. This table shows the effects of treatment for each child considering the outcomes not only of the first scored referral, but also of all later referrals. A control for whether the first referral was part of an ongoing episode is included throughout. Outcomes included are: whether the child was ever investigated, ever had a founded investigation, ever had a case opened after the referral, ever had a differential response case opened, ever had a traditional child welfare case opened, ever had a prevention case opened, and whether the family received any child welfare services (including assessments and child placements). Column 1 reports means and the standard errors for those in control and scoring an 15 or less (approximately the bottom two thirds of the score distribution, T1 and T2). Column 2 reports the effects of being treated. Column 3 reports the coefficient on the dummy variable indicating whether this individual scored a 16 or above (the top third of the score distribution, T3). Column 4 reports the effect of showing a score of 16 or above to caseworkers. Column 5 reports the number of times that the outcome is not equal to 0, which is to say the frequency of each outcome.

screen in rate goes up by an additional 14percentage points for children with a score of 16 or above relative to children scoring 15 or below for whom the screen in rate is 37% (row 1, columns 1 and 3). The standard error for the effect of seeing the score for high scoring children in row 1 of column 4 however is too large (0.066) to be able to conclude whether showing the score has any effect on screen ins at this stage. However, showing a high risk score increases the likelihood that a child has a founded investigation (row 2 column 4). This suggests that when the score is high seeing the risk score is helping child welfare to visit families where there were indeed concerns about the child's wellbeing and safety: the investigation was founded. Of course, the frequency of this outcome is still rather rare so results will need to be rerun once more data is available.

The following rows of Table 3.5 then report additional child welfare outcomes: whether a child welfare case was opened, whether a differential response case was opened (a low risk track case), whether a traditional case was opened (the high risk track), whether a prevention case was opened and whether the family received any services. Although all the point estimates are not statistically significant, they all indicate that children scoring in the top third of the distribution are more likely to have all types of cases opened, and more likely to receive services (column 3). This is what one would expect if social workers were picking up on which cases are high risk even without seeing the score. The point estimates for the effects of the treatment in columns (2) and (4) are not yet statistically significant: the standard errors in all specifications are still too large to conclude that there is an effect or no effect. However, we nonetheless want to flag the effect in last row given the magnitude of the effect size. Seeing a score for children scoring a 16 or above, leads them to be 3 percentage points more likely to receive any services. Relative to the mean of those scoring a 16 or above ($0.040 + 0.085 = 12.5\%$) that represents a 25% increase in likelihood of receiving services when the score is shown for high scoring children. This would be a substantial change in practice, if this point estimate were true. Future data will be able to tell us whether this persists. More work will also need to be done to assess what types of services are increasing.

One frequent concern however about child welfare outcomes is concerns about surveillance. This term of surveillance is often used in reference to families interacting with many professionals before a CPS interaction: mandated professionals would indeed be required to report anything that could be suspect. Yet, when child welfare is interacting more with a family themselves, or is giving access to services to a family, a similar phenomena might occur: child welfare would be more likely to know if anything went wrong. Thus, if seeing a score led child welfare workers or service providers to be present more often in the families' lives, then someone would be more likely to notice if anything was a concern. From child welfare's perspective, this is not an issue since they precisely would like to know any time there is something of concern. Yet, in the context of this evaluation, one would like to know how seeing the score is affecting outcomes: is there any improved targeting of interventions or are the effects just through better awareness of incidents of concern? One outcome should precisely help us to clarify this: child welfare should have no improved awareness of incidents in screened out families' lives. If the tool led to better targeting however, we would expect that the cases being screened out would be of lower risk and so less likely to be re-referred. Results are presented in Table 3.6 and indeed seem to suggest that seeing the score reduces the likelihood of a screened out referral being re-referred to child welfare: -7.2pp decrease in the re-referral rate of 33.1%. These effects are substantial by their size, though it is slightly surprising that they occur throughout the score distribution as indicated in column 2: even children scoring 15 or below are less likely to be re-referred (-5.3pp, row 1 column 2). Effects are especially large for those with a score of 16 or above: -5.3pp-4.1pp=-9.4pp, cutting by almost half the increased rate of re-referrals for high scoring children ($-0.094/0.222 \approx -42\%$). Of course, the coefficients for rows 1 and 3 in column 2 are currently not statistically significant, so whether these effects by score level will become statistically significant will depend on future data. Finally, column 2 also indicates that children with a score in the top third of the distribution are indeed much more likely to be re-referred, indicating that re-referrals is a good proxy for risk and that the tool is indeed flagging children with high likelihood of child welfare involvement (row 2 column 2).

Table 3.6: Re-referrals of Screened Out Cases

	(1)	(2)
Score shown (T)	-0.072* (0.037)	-0.053 (0.039)
Top third of scores (T3)		0.222*** (0.054)
T · T3		-0.041 (0.074)
Observations	1244	1244
Number of Clusters	716	716
Reference Group Mean	.331	.23

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. A control for whether the first referral was part of an ongoing episode is included throughout. This table shows the effects of treatment on re-referrals for screened out children. The reference group mean in column (1) is the re-referral rate for all children in the control group, whereas in column (2) it is the re-referral rate for children in the bottom two terciles of the score distribution.

3.5 Conclusion

The emerging suite of algorithmic decision aids presents exciting possibilities for efficiency gains across numerous policy domains, but the intricacies of human-algorithm interaction may improve or hinder the efficacy and equity of such tools. Preliminary results from this study seem to suggest promising changes to child welfare practice as a result of seeing a child risk score. Caseworkers seem to be adhering to using the tool in practice. Interestingly, even without seeing a score, caseworkers are able to allocate services and interventions toward high risk families at a better-than-random rate. We find evidence that seeing the risk score increases the likelihood that high scoring cases have a founded investigation, and improves targeting as evidenced by decreased re-referrals for referrals that were screened out when the score was shown. Given a limited sample size for this preliminary evaluation, these latter results should be interpreted as suggestive until further evidence is collected.

Future work in this evaluation will examine effects on less frequent outcomes such as child removals, referrals of non-minor severity, and referrals where the child was injured. Our sample size is currently too limited to provide compelling evidence on those outcomes.

More generally, we hope to someday assess the effects of the intervention on child wellbeing by matching these data to hospital emergency room discharge data on health incidents for children.

Provided our preliminary results remain after the collection of more data, our results may have substantial consequences for policy and for the future design of machine learning tools. Current results suggest that thoughtfully designed and validated algorithm-based tools can be effective in practice, which would have consequences for the use of such tools both in child welfare and beyond.

References

- Abaluck, J., Agha, L., Chan, D., & Zhu, Y. (2020, July). *Mastering the Art of Cookbook Medicine: Machine Learning, Randomized Trials, and Misallocation* (SSRN Scholarly Paper No. ID 3644068). Rochester, NY: Social Science Research Network.
- Aizer, A., & Currie, J. (2014). The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science*, 344(6186), 856–861.
- Aizer, A., Devereux, P. J., & Salvanes, K. G. (2018). Grandparents, Moms, or Dads? Why Children of Teen Mothers Do Worse in Life. *NBER Working Paper #25165*.
- Akerlof, G. A. (2020). Sins of Omission and the Practice of Economics. *Journal of Economic Literature*, 58(2), 405–418.
- Albright, A. (2019). *If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions*. Working Paper.
- Alpert, A., Evans, W., Lieber, E. M., & Powell, D. (2019, November). Origins of the Opioid Crisis and Its Enduring Impacts. *NBER Working Paper #26500*.
- Alpert, A., Powell, D., & Pacula, R. L. (2018). Supply-Side Drug Policy in the Presence of Substitutes: Evidence from the Introduction of Abuse-Deterrent Opioids. *American Economic Journal: Economic Policy*, 10(4), 1–35.
- Alsan, M., & Wanamaker, M. (2018). Tuskegee and the Health of Black Men. *The Quarterly Journal of Economics*, 133(1), 407–455.
- Altindag, O., Greve, J., & Tekin, E. (2021, March). *Public Health Policy At Scale: Impact of a Government-sponsored Information Campaign on Infant Mortality in Denmark* (Tech. Rep. No. w28621). National Bureau of Economic Research. doi: 10.3386/w28621
- Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife Instrumental Variables Estimation. *Journal of Applied Econometrics*, 14(1), 57–67.
- Ash, M., & Robinson, D. E. (2009, June). Inequality, race, and mortality in U.S. cities: A political and econometric review of Deaton and Lubotsky (56:6, 1139–1153, 2003). *Social Science & Medicine*, 68(11), 1909–1913.
- Bailey, Z. D., Feldman, J. M., & Bassett, M. T. (2021). How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities. *The New England Journal of Medicine*, 384(8), 768–773.

- Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020). Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial. *American Economic Review*, 110(3), 824–859.
- Barth, R. P., Putnam-Hornstein, E., Shaw, T. V., & Dickinson, N. S. (2016, January). Safe Children: Reducing Severe and Fatal Maltreatment. *American Academy of Social Work and Social Welfare*(Working Paper No. 17), 21.
- Becker, G. S., & Murphy, K. M. (1988). A Theory of Rational Addiction. *Journal of Political Economy*, 96(4), 675–700.
- Bhuller, M., Dahl, G. B., Løken, K. V., & Mogstad, M. (2020). Incarceration, Recidivism, and Employment. *Journal of Political Economy*, 128(4), 1269–1324.
- Borusyak, K., & Jaravel, X. (2018, April). Revisiting Event Study Designs. *SSRN Electronic Journal*.
- Buckles, K., Evans, W. N., & Lieber, E. M. (2020). The Drug Crisis and the Living Arrangements of Children. *NBER Working Paper #27633*.
- Bundorf, K., Polyakova, M., & Tai-Seale, M. (2019). How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance. doi: 10.3386/w25976
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372.
- Capacity Building Center for States. (2019). *Strategies for Authentic Integration of Family and Youth Voice in Child Welfare*. Washington, DC: Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.
- Case, A., Lubotsky, D., & Paxson, C. (2002). Economic Status and Health in Childhood: The Origins of the Gradient. *American Economic Review*, 92(5), 1308–1334.
- CDC. (2020, September). *Infant Mortality*. <https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm>.
- Chaiyachati, B. H., Wood, J. N., Mitra, N., & Chaiyachati, K. H. (2020). All-Cause Mortality Among Children in the US Foster Care System, 2003-2016. *JAMA Pediatrics*, 174(9), 896–898.
- Chance, T., & Scannapieco, M. (2002). Ecological Correlates of Child Maltreatment: Similarities and Differences Between Child Fatality and Nonfatality Cases. *Child & Adolescent Social Work Journal*, 19(2), 139–161.
- Cherna Leadership Style*. (2019, April). <https://www.casey.org/leadership-style-cher-na/>.
- Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., ... Cutler, D. (2016, April). The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA*, 315(16), 1750.

- Chyn, E., Hastings, J. S., Machelett, M., & Bald, A. (2019). The Causal Impact of Removing Children from Abusive and Neglectful Homes. *NBER Working Paper #25419*.
- Conti, G., & Heckman, J. J. (2012, December). The Developmental Approach to Child and Adult Health. *NBER Working Paper #18664*. doi: 10.3386/w18664
- Covington, T. M. (2019). Child Death Review: An Effective Approach for the Surveillance of Sudden and Unexplained Infant Deaths in the US. In F. R. Hauck, I. B. Scheimberg, J. B. Beckwith, & M. C. Cohen (Eds.), *Investigation of Sudden Infant Death Syndrome* (pp. 75–85). Cambridge: Cambridge University Press.
- Crume, T. L., DiGuseppi, C., Byers, T., Sirotiak, A. P., & Garrett, C. J. (2002). Underascertainment of Child Maltreatment Fatalities by Death Certificates, 1990-1998. *Pediatrics*, 110(2), e18.
- Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, 79, 291–298.
- Currie, J., & Neidell, M. (2005). Air Pollution and Infant Health: What Can We Learn From California's Recent Experience? *The Quarterly Journal of Economics*, 120(3), 1003–1030.
- Currie, J., & Schwandt, H. (2016a). Inequality in mortality decreased among the young while increasing for older adults, 1990-2010. *Science*, 352(6286), 708–712.
- Currie, J., & Schwandt, H. (2016b). Mortality Inequality: The Good News from a County-Level Approach. *Journal of Economic Perspectives*, 30(2), 29–52.
- Cutler, D., Deaton, A., & Lleras-Muney, A. (2006). The Determinants of Mortality. *Journal of Economic Perspectives*, 20(3), 97–120.
- Dahl, G. B., Kostol, A. R., & Mogstad, M. (2014). Family Welfare Cultures. *The Quarterly Journal of Economics*, 129(4), 1711–1752.
- De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In (pp. 1–12). ACM.
- de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9), 2964–2996.
- Deaton, A. (2013). *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.
- Deaton, A., & Lubotsky, D. (2003, March). Mortality, inequality and race in American cities and states. *Social Science & Medicine*, 56(6), 1139–1153.
- DHS. (2019, May). *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions*.
- Dobbie, W., Yang, C. S., & Goldin, J. (2018). The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review*, 108(2), 201–240.

- Douglas, E. M. (2016, August). Testing if Social Services Prevent Fatal Child Maltreatment Among a Sample of Children Previously Known to Child Protective Services. *Child Maltreatment*, 21(3), 239–249.
- Douglas, E. M., & Mohn, B. L. (2013). Fatal and non-fatal child maltreatment in the US: An analysis of child, caregiver, and service utilization with the National Child Abuse and Neglect Data Set. *Child Abuse & Neglect*, 38(1), 42–51.
- Doyle, J. J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5), 1583–1610.
- Doyle, J. J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy*, 116(4), 746–770.
- Doyle, J. J. (2013). Causal effects of foster care: An instrumental-variables approach. *Children and Youth Services Review*, 35(7), 1143–1151.
- Eckerd Rapid Safety Feedback. (2021). <https://eckerd.org/family-children-services/ersf/>.
- Fitzpatrick, M., & Wildeman, C. (2020). *Final Report on Douglas County Decision Aid (DCDA). Predictive Risk Modeling Randomized Control Trial Experiment*. Working Paper.
- Flaherty, C. W., & Patterson, D. A. (2003). Predicting Child Physical Abuse Recurrence: Comparison of a Neural Network to Logistic Regression. *Journal of Technology in Human Services*, 21(4), 93–111.
- Ghertner, R., Baldwin, M., Crouse, G., Radcliff, L., & Waters, A. (2018, March). The Relationship between Substance Use Indicators and Child Welfare Caseloads. *ASPE Research Brief*.
- Ghertner, R., Waters, A., Radcliff, L., & Crouse, G. (2018). The role of substance use in child welfare caseloads. *Children and Youth Services Review*, 90, 83–93.
- Goodman-Bacon, A. (2018). Difference-in-Differences with Variation in Treatment Timing. *NBER Working Paper #25018*.
- Green, B., & Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on human-computer interaction*, 3(CSCW), 1–24. doi: 10.1145/3359152
- Grimon, M.-P. (2020). Effects of the Child Protection System on Parents. *Working Paper*.
- Gross, M. (2019). Temporary Stays and Persistent Gains: The Causal Effects of Foster Care.
- Gruber, J., & Köszegi, B. (2001). Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, 116(4), 1261–1303.
- Herman-Giddens, M. E., Brown, G., Verbiest, S., Carlson, P. J., Hooten, E. G., Howell, E., & Butts, J. D. (1999, August). Underascertainment of child abuse mortality in the United States. *JAMA*, 282(5), 463–467.
- Hotz, V. J., McElroy, S. W., & Todd, S. E. (2005). Teenage childbearing and its life cycle consequences: Exploiting a natural experiment. *Journal of Human Resources*, 40(3), 683–715.

- Hoynes, H., Schanzenbach, D. W., & Almond, D. (2016). Long-Run Impacts of Childhood Access to the Safety Net. *American Economic Review*, 106(4), 903–934. doi: 10.1257/aer.20130375
- Hull, P. (2017, March). Examiner Designs and First-Stage F Statistics: A Caution.
- Hurley, D. (2018, January). Can an Algorithm Tell When Kids Are in Danger? *The New York Times Magazine*, 30.
- Imai, K., Jiang, Z., Greiner, J., Halen, R., & Shin, S. (2020). Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment.
- Jackson, D., & Marx, G. (2017, December). *Data mining program designed to predict child abuse proves unreliable, DCFS says*. <https://www.chicagotribune.com/investigations/ct-dcfs-eckerd-met-20171206-story.html>.
- Johnson, M., Brown, C. H., & Wells, S. (2002). Using classification and regression trees (CART) to support worker decision making. *Social Work Research*, 26(1), 19–29. doi: 10.1093/swr/26.1.19
- Jolley, J. M. (2012). *Applying Neural Network Models to Predict Recurrent Maltreatment in Child Welfare Cases with Static and Dynamic Risk Factors* (Unpublished doctoral dissertation).
- Jolly, S., Griffith, K. A., DeCastro, R., Stewart, A., Ubel, P., & Jagsi, R. (2014). Gender Differences in Time Spent on Parenting and Domestic Responsibilities by High-Achieving Young Physician-Researchers. *Annals of Internal Medicine*, 160(5), 344–353.
- Karma, R. (2019, May). The Gross Inequality of Death in America. *The New Republic*.
- Kelly, C., Street, C., & Building, M. E. S. (2019). Child Maltreatment 2019. *Child Maltreatment*, 306.
- Kim, H., Wildeman, C., Jonson-Reid, M., & Drake, B. (2017). Lifetime Prevalence of Investigating Child Maltreatment Among US Children. *American Journal of Public Health*, 107(2), 274–280.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1), 237–293.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, 105(5), 491–495. doi: 10.1257/aer.p20151023
- Kleven, H., Landais, C., & Sogaard, J. E. (2019). Children and Gender Inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, 11(4), 181–209.
- Kling, J. R. (2006). Incarceration Length, Employment, and Earnings. *American Economic Review*, 96(3), 863–876.
- Lindquist, M. J., & Santavirta, T. (2014). Does placing children in foster care increase their adult criminality? *Labour economics*, 31, 72–83. doi: 10.1016/j.labeco.2014.10.001

- Mackenbach, J. P., Stirbu, I., Roskam, A.-J. R., Schaap, M. M., Menvielle, G., Leinsalu, M., . . . European Union Working Group on Socioeconomic Inequalities in Health (2008, June). Socioeconomic inequalities in health in 22 European countries. *The New England Journal of Medicine*, 358(23), 2468–2481.
- Maestas, N., Mullen, K. J., & Strand, A. (2013). Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt. *American Economic Review*, 103(5), 1797–1829.
- Marshall, D. B., & English, D. J. (2000). Neural Network Modeling of Risk Assessment in Child Protective Services. *Psychological Methods*, 5(1), 102–124. doi: 10.1037/1082-989X.5.1.102
- Medical Examiner in Allegheny County*. (n.d.). <https://www.alleghenycounty.us/medical-examiner/index.aspx>.
- Mueller-Smith, M. (2015, August). The Criminal and Labor Market Impacts of Incarceration.
- Nash, M. (2016, September). *Examination of Using Structured Decision Making and Predictive Analytics in Assessing Safety and Risk in Child Welfare*. County of Los Angeles, Office of Child Protection.
- National Center on Substance Abuse and Child Welfare. (2019). *Understanding Co-Occurring Substance Use Disorders, Mental Health/Trauma, and Domestic Violence*. U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration.
- Newton, A. W., & Vandeven, A. M. (2006). Unexplained infant and child death: A review of sudden infant death syndrome, sudden unexplained infant death, and child maltreatment fatalities including shaken baby syndrome. *Current opinion in pediatrics*, 18(2), 196–200.
- Olds, D. L., Kitzman, H., Knudtson, M. D., Anson, E., Smith, J. A., & Cole, R. (2014). Effect of Home Visiting by Nurses on Maternal and Child Mortality: Results of a 2-Decade Follow-up of a Randomized Clinical Trial. *JAMA Pediatrics*, 168(9), 800–806.
- Oliviere, D., Monroe, B., & Payne, S. (2011). *Death, dying, and social differences* (2nd ed. / edited by David Oliviere, Barbara Monroe, Sheila Payne. ed.). Oxford ; New York: Oxford University Press.
- Packard, T. (2016, February). *Literature Review: Predictive Analytics in Human Services*. Southern Area Consortium of Human Services.
- Parrish, G. (1995). Assessing and improving the quality of data from medical examiners and coroners. *Proceedings of the international collaborative effort on injury statistics*, 1(25), 1–10.
- Parrish, J. W., Schnitzer, P. G., Lanier, P., Shanahan, M. E., Daniels, J. L., & Marshall, S. W. (2017, May). Classification of maltreatment-related mortality by Child Death Review teams: How reliable are they? *Child Abuse & Neglect*, 67, 362–370.
- Peltzman, S. (2009). Mortality Inequality. *Journal of Economic Perspectives*, 23(4), 175–190.

- Posey, B. M., & Neuilly, M.-A. (2017, October). A fatal review: Exploring how children's deaths are reported in the United States. *Child Abuse & Neglect*, 72, 433–445.
- Prindle, J. J., Hammond, I., & Putnam-Hornstein, E. (2018). Prenatal substance exposure diagnosed at birth and infant involvement with child protective services. *Child Abuse & Neglect*, 76, 75–83.
- Putnam-Hornstein, E., Cleves, M. A., Licht, R., & Needell, B. (2013, October). Risk of Fatal Injury in Young Children Following Abuse Allegations: Evidence From a Prospective, Population-Based Study. *American Journal of Public Health*, 103(10), e39-e44.
- Quast, T., Storch, E. A., & Yampolskaya, S. (2018). Opioid Prescription Rates And Child Removals: Evidence From Florida. *Health Affairs*, 37(1), 134–139. doi: 10.1377/hlthaff.2017.1023
- Radel, L., Baldwin, M., Crouse, G., Ghertner, R., & Waters, A. (2018, March). Substance Use, the Opioid Epidemic, and the Child Welfare System: Key Findings from a Mixed Methods Study. *ASPE Research Brief*.
- Roberts, K. (2019). *Foster Care and Child Welfare* (Unpublished doctoral dissertation). Clemson University.
- Russell, J. (2015). Predictive analytics and child protection: Constraints and opportunities. *Child Abuse & Neglect*, 46, 182–189.
- Sankaran, V., Rideout, P. L., & Raimon, M. L. (2015). Strange Bedfellows: How Child Welfare Agencies Can Benefit from Investing in Multidisciplinary Parent Representation. *Center for the Study of Social Policy*.
- Schnell, M., & Currie, J. (2018). Addressing the Opioid Epidemic: Is There a Role for Physician Education? *American Journal of Health Economics*, 4(3), 383–410.
- Schnitzer, P. G., Covington, T. M., Wirtz, S. J., Verhoek-Oftedahl, W., & Palusci, V. J. (2008). Public Health Surveillance of Fatal Child Maltreatment: Analysis of 3 State Programs. *American Journal of Public Health* (1971), 98(2), 296–303.
- Schwartz, D. R., Kaufman, A. B., & Schwartz, I. M. (2004). Computational intelligence techniques for risk assessment and decision support. *Children and Youth Services Review*, 26(11), 1081–1095.
- Sepulveda, K., & Williams, S. C. (2019, February). *One in three children entered foster care in 2017 because of parental drug abuse*. ChildTrends.
- Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Law and Human Behavior*, 44(1), 51–59. doi: 10.1037/1hb0000360
- Sledjeski, E. M., Sledjeski, E. M., Dierker, L. C., Dierker, L. C., Brigham, R., Brigham, R., ... Breslin, E. (2008). The Use of Risk Assessment to Predict Recurrent Maltreatment: A Classification and Regression Tree Analysis (CART). *Prevention Science*, 9(1), 28–37.

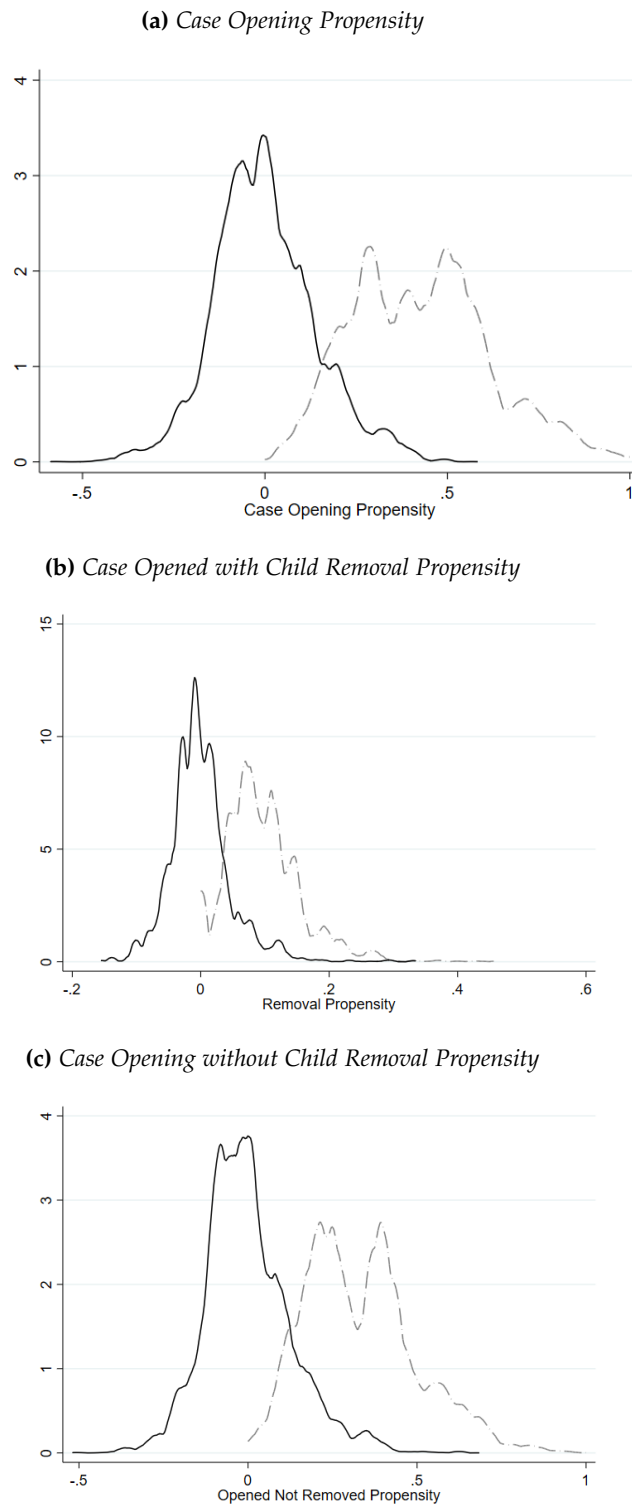
- Smedley, B. D., Stith, A. Y., Nelson, A. R., & Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, D.C.: National Academies Press.
- Stevenson, M., & Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. *SSRN Electronic Journal*.
- Thurston, H., Freisthler, B., Bell, J., Tancredi, D., Romano, P. S., Miyamoto, S., & Joseph, J. G. (2017). Environmental and individual attributes associated with child maltreatment resulting in hospitalization or death. *Child Abuse & Neglect*, 67, 119–136.
- Tuljapurkar, S. (2011). *The Final Inequality*. University of Chicago Press.
- U.S. Census Bureau QuickFacts: Allegheny County, Pennsylvania. (2019). <https://www.census.gov/quickfacts/alleghenycountypennsylvania>.
- Vaithianathan, R., Maloney, T., Jiang, N., De Haan, I., Dale, C., Putnam-Hornstein, E., & Dare, T. (2012, September). *Vulnerable Children - Can administrative data be used to identify children at risk of adverse outcomes?* Centre for Applied Research in Economics (CARE), University of Auckland.
- Vaithianathan, R., Putnam-Hornstein, E., Chouldechova, A., Benavides-Prado, D., & Berger, R. (2020). Hospital Injury Encounters of Children Identified by a Predictive Risk Model for Screening Child Maltreatment Referrals: Evidence From the Allegheny Family Screening Tool. *JAMA Pediatrics*, 174(11), e202770–e202770.
- Vaithianathan, R., Rouland, B., & Putnam-Hornstein, E. (2018, February). Injury and Mortality Among Children Identified as at High Risk of Maltreatment. *Pediatrics*, 141(2).
- Vogl, T. S. (2018). Dynamics of Child Mortality Inequality. *AEA Papers and Proceedings*, 108, 348–352.
- Warburton, W. P., Warburton, R. N., Sweetman, A., & Hertzman, C. (2014). The Impact of Placing Adolescent Males into Foster Care on Education, Income Assistance, and Convictions. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 47(1), 35–69.
- Williams, Z. (2018, July). *Dramatic spike in emergency removals of children by ACS* [Text]. <https://nynmedia.com/new-york-city-administration-childrens-services-emergency-removals-children-increase>.
- Zandi, I. (2001). Use of artificial neural network as a risk assessment tool in preventing child abuse. In (Vol. 2, pp. 1438–1442 vol.2). IEEE.

Appendix A

Appendix to Chapter 1

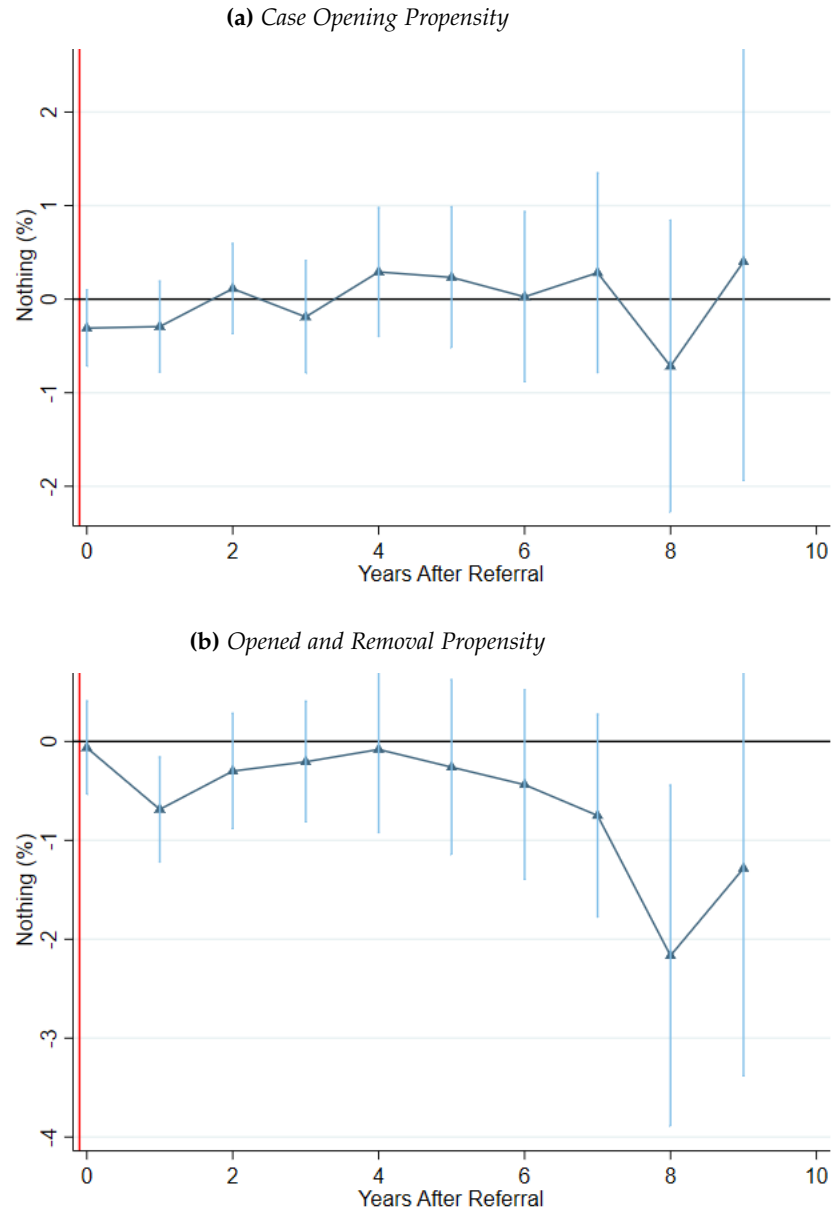
A.1 Figures

Figure A.1: *Density of Propensity Instruments*



Notes: Dashed gray lines are the raw propensity distributions. Solid black line are the propensity distributions residualized for office-year fixed effects. Restricted to referrals assigned to an investigator overseeing at least 10 referrals.

Figure A.2: *Selective Attrition - Mothers Not Involved in Any Program*



Notes: For the sample of mothers assigned to an investigator that investigates at least 10 referrals, this graph plots the probability of not being involved in any program in the years subsequent to the referral as a proxy for moving out of county. Figure A.2 (a) plots the effects when assigned to an investigator one standard deviation more likely to open a case. Figure A.2 (b) plots those when assigned to an investigator one standard deviation more likely to open and remove a child, holding constant your investigator's propensity to open and not remove a child. Coefficients and 90% confidence intervals are reported, with two-way clustering for investigators and parents.

A.2 Tables

Table A.1: *Number of Observations, May 2009 - Dec 2018*

Restrictions	Adults	Unique Adults	Referrals
On Referrals	279250	113300	107934
Parents	232606	92998	106348
Investigated	121962	59975	53489
Not Reassigned	100214	58499	44331
Outcomes	96877	55877	43906
Found Investigator	88296	52467	39860
One Investigator	78125	48652	35386
General Protective Services	62199	40532	28167

Notes: This table reports sample size restrictions. The first row reports the total number of adults, unique adults and referrals listed between May 2009 and December 2018. Then each subsequent row adds an additional restriction. The second row restricts to parents, and referrals with at least one parent listed. The third row lists parents that were investigated. The fourth row excludes those that were reassigned to the previous investigator because the last investigation was not long before. The fifth row restricts to parents for whom I have outcomes data, the sixth to those where the agency's records list at least one investigator, the seventh to where they do not list more than one investigator, and finally the last row restricts to the sample of cases that were not initially reported as a case of abuse.

Table A.2: *Randomization Test*

	Open Case	Open & No Removal	Open & Removal
	(1)	(2)	(3)
White	-0.10** [0.04]	-0.10** [0.05]	-0.01 [0.01]
Race Missing	-0.05** [0.02]	-0.05* [0.02]	-0.01 [0.01]
Hispanic	0.01 [0.05]	-0.03 [0.05]	0.11* [0.06]
Ethnicity Missing	0.01 [0.01]	0.01 [0.01]	-0.01 [0.01]
Women	-0.00 [0.01]	0.00 [0.01]	-0.01 [0.01]
Age	-0.00 [0.00]	0.00 [0.00]	-0.00* [0.00]
Mandated Reporter	-0.10 [0.07]	-0.09 [0.06]	-0.03 [0.07]
Anonymous Reporter	-0.01 [0.06]	0.01 [0.06]	-0.06 [0.07]
Courts Reported	0.05 [0.06]	0.07 [0.06]	-0.04 [0.03]
Family Reported	-0.00 [0.06]	0.01 [0.05]	-0.03 [0.06]
Health Professional Reported	0.04 [0.04]	0.05 [0.04]	-0.02 [0.02]
Neighbor or Friend Reported	-0.02 [0.07]	0.00 [0.07]	-0.06 [0.07]
Police Reported	0.09** [0.04]	0.09** [0.04]	0.02 [0.02]
School Reported	0.00 [0.04]	0.02 [0.04]	-0.04 [0.03]
Social worker Reported	0.09* [0.04]	0.11** [0.04]	-0.04 [0.03]
TANF 3 months prior	0.02 [0.02]	0.01 [0.02]	0.03 [0.02]
SNAP 3 months prior	0.01 [0.01]	0.01 [0.01]	-0.00 [0.01]
MH 3 months prior	0.01 [0.02]	0.01 [0.02]	0.00 [0.02]
Medicaid 3 months prior	-0.02 [0.02]	-0.01 [0.02]	-0.01 [0.02]
Jail 3 months prior	0.02	0.01	0.01

Continued on next page

Table A.2 – Randomization Test (Continued from previous page)

	Open Case	Open & No Removal	Open & Removal
	(1)	(2)	(3)
Substance Abuse 3 months prior	[0.02] 0.00	[0.02] -0.00	[0.02] 0.02
Courts 3 months prior	[0.03] -0.00	[0.03] 0.01	[0.03] -0.03*
Screen-in 3 months prior	[0.02] -0.07	[0.02] -0.06	[0.01] -0.02
Referral 3 months prior	[0.09] 0.05	[0.08] 0.06*	[0.08] -0.02
Open Case 3 months prior	[0.03] -0.07	[0.03] -0.06	[0.03] -0.02
Open Case prior year 1	[0.12] -0.05*	[0.11] -0.04	[0.12] -0.03
Referral prior year 1	[0.03] 0.00	[0.03] -0.00	[0.02] 0.02
Screen-in prior year 1	[0.02] -0.01	[0.02] -0.01	[0.01] -0.00
Courts prior year 1	[0.02] 0.03	[0.02] 0.02	[0.02] 0.02
Substance Abuse prior year 1	[0.02] 0.01	[0.02] 0.00	[0.01] 0.02
Jail prior year 1	[0.02] -0.03**	[0.02] -0.03**	[0.02] 0.00
Medicaid prior year 1	[0.01] 0.02	[0.01] 0.02	[0.01] 0.01
MH prior year 1	[0.02] -0.00	[0.02] 0.00	[0.01] -0.00
SNAP prior year 1	[0.01] -0.01	[0.01] -0.01	[0.01] 0.01
TANF prior year 1	[0.01] -0.01	[0.01] -0.00	[0.01] -0.02
Number of investigators	460	460	460
Chi-squared	53.81	55.61	43.08
P-value	0.0220	0.0148	0.164
Observations	58548	58548	58548

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors are cluster bootstrapped by investigator, with 800 bootstraps. Regional office year fixed effects are included throughout. The sample is restricted to investigators overseeing at least ten referrals. This table reports the details from the tests assessing randomness of assignments of investigators with varying levels of propensity to open a case (column 1), propensity to open a case without a removal (column 2), and propensity to remove a child (column 3). The propensities are standardized so that the

size of the coefficients can be interpreted in a meaningful way. The first six characteristics are demographic characteristics of parents (race, ethnicity, gender, age). The next nine variables indicate what type of individual called child welfare to report abuse or neglect: a reporter mandated by law to call child welfare, courts, family, health professional, police, school, social worker, family, etc. The missing category is "other reporter". Finally, the remaining variables included in the test are indicator variables for whether parents received any of the services listed in the third month prior to the referral ("3 months prior"), or at any point in time in the year prior to the referral ("prior year 1").

Table A.3: Tests for Monotonicity

	(1) First Stage	(2) Reverse Sample First Stage
Whites	0.667*** [0.029]	0.608*** [0.034]
Non-whites	0.700*** [0.032]	0.685*** [0.034]
Mothers	0.716*** [0.026]	0.865*** [0.018]
Fathers	0.697*** [0.030]	0.891*** [0.021]
Mandated Reporter	0.709*** [0.031]	0.500*** [0.041]
Not a Mandated Reporter	0.692*** [0.041]	0.618*** [0.048]
Used Substance Abuse Services in year prior	0.738*** [0.051]	0.839*** [0.050]
No Substance Abuse Services in year prior	0.703*** [0.028]	0.578*** [0.048]
Used Mental Health Services in year prior	0.728*** [0.035]	0.832*** [0.034]
No Mental Health Services in year prior	0.700*** [0.029]	0.637*** [0.035]

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors are clustered at the investigator level. Year and regional office fixed effects are included throughout. The sample is restricted to investigators overseeing at least ten referrals. This table reports a subset of the tests run to assess whether there were any monotonicity violations. Column (1) reports the first stage coefficient on the propensity to open a case for the sample indicated in the left-most column. Column (2) reports the first stage coefficient on the propensity to open a case for the sample in the left-most column, where the propensity is constructed for the opposite sample. So for the first row, the first column reports the first stage coefficients for whites using the same propensity as in the rest of the paper. The second column reports the first stage coefficients for whites where the propensity is constructed using the sample of non-whites that the investigator investigated.

Table A.4: *Robustness Tests for Month Results*

	Health			Justice		Social Services		
	(1) Medicaid	(2) MH	(3) SA	(4) Courts	(5) Jail	(6) TANF	(7) SNAP	(8) Housing
1. Original Specification								
Opened no Removal Propensity	0.23 [0.18]	0.30** [0.14]	0.38*** [0.13]	0.01 [0.17]	0.07 [0.06]	-0.21 [0.18]	0.30 [0.21]	0.04 [0.12]
Opened and Removal Propensity	0.20 [0.18]	0.58*** [0.16]	0.29** [0.14]	0.46** [0.22]	0.05 [0.07]	-0.47** [0.18]	-0.37 [0.23]	-0.01 [0.14]
After referral	2.51*** [0.17]	2.71*** [0.15]	4.24*** [0.18]	2.08*** [0.19]	0.95*** [0.08]	1.70*** [0.18]	1.08*** [0.23]	1.20*** [0.15]
2. Controlling for Allegation Propensities								
Opened no Removal Propensity	0.25 [0.20]	0.21 [0.15]	0.32** [0.14]	-0.08 [0.19]	0.08 [0.06]	-0.19 [0.20]	0.21 [0.26]	-0.04 [0.16]
Opened and Removal Propensity	0.11 [0.19]	0.50*** [0.18]	0.16 [0.15]	0.53** [0.23]	0.04 [0.08]	-0.59*** [0.20]	-0.47* [0.25]	-0.02 [0.15]
After referral	2.51*** [0.17]	2.70*** [0.15]	4.23*** [0.18]	2.09*** [0.19]	0.95*** [0.08]	1.71*** [0.18]	1.08*** [0.23]	1.19*** [0.14]
3. Keeping only First Referral								
Opened no Removal Propensity	0.26 [0.23]	0.29* [0.15]	0.54*** [0.15]	0.16 [0.20]	0.05 [0.06]	-0.34 [0.24]	0.55** [0.26]	0.04 [0.16]
Opened and Removal Propensity	0.38 [0.24]	0.67*** [0.18]	0.32** [0.15]	0.67** [0.28]	0.06 [0.08]	-0.30 [0.26]	-0.35 [0.29]	-0.03 [0.17]
After referral	3.09*** [0.20]	2.73*** [0.17]	4.41*** [0.20]	2.06*** [0.23]	0.85*** [0.09]	2.18*** [0.21]	1.60*** [0.29]	1.32*** [0.17]

Continued on next page

Table A.4 – Robustness Tests for Month Results (Continued from previous page)

	Health			Justice		Social Services		
	(1) Medicaid	(2) MH	(3) SA	(4) Courts	(5) Jail	(6) TANF	(7) SNAP	(8) Housing
4. Before One Caseworker Model								
Opened no Removal Propensity	-0.11 [0.22]	0.13 [0.15]	0.39** [0.18]	0.10 [0.21]	0.08 [0.08]	-0.23 [0.21]	0.11 [0.27]	-0.31 [0.20]
Opened and Removal Propensity	0.09 [0.24]	0.47*** [0.17]	0.40** [0.18]	0.56* [0.30]	0.07 [0.09]	-0.62** [0.25]	-0.63** [0.29]	-0.37 [0.30]
After referral	2.34*** [0.22]	2.55*** [0.18]	3.71*** [0.21]	2.32*** [0.23]	0.95*** [0.10]	1.34*** [0.21]	1.24*** [0.28]	0.79*** [0.20]
5. Control for Imbalanced Characteristics After								
Opened no Removal Propensity	0.26 [0.18]	0.30** [0.14]	0.39*** [0.13]	-0.02 [0.16]	0.06 [0.06]	-0.21 [0.18]	0.33 [0.21]	0.03 [0.12]
Opened and Removal Propensity	0.22 [0.18]	0.58*** [0.16]	0.29** [0.14]	0.42** [0.21]	0.05 [0.07]	-0.44** [0.18]	-0.37 [0.23]	-0.01 [0.14]
After referral	1.60*** [0.28]	2.30*** [0.22]	3.01*** [0.24]	1.05*** [0.29]	0.57*** [0.10]	1.19*** [0.32]	0.57 [0.36]	2.11*** [0.26]
6. Control for Risk and Safety Levels After								
Opened no Removal Propensity	0.22 [0.19]	0.22 [0.14]	0.38*** [0.14]	-0.02 [0.17]	0.07 [0.06]	-0.17 [0.18]	0.32 [0.24]	0.05 [0.12]
Opened and Removal Propensity	0.10 [0.19]	0.52*** [0.17]	0.20 [0.15]	0.37 [0.23]	0.04 [0.08]	-0.37* [0.20]	-0.44* [0.24]	-0.05 [0.15]
After referral	1.50*** [0.34]	1.73*** [0.26]	2.59*** [0.22]	0.22 [0.36]	0.47*** [0.13]	0.92*** [0.34]	0.01 [0.41]	0.68*** [0.24]
7. Truncate Propensity Values								
Opened no Removal Propensity	0.24	0.33**	0.38***	-0.02	0.07	-0.21	0.25	0.06

Continued on next page

Table A.4 – Robustness Tests for Month Results (Continued from previous page)

	Health			Justice		Social Services		
	(1) Medicaid	(2) MH	(3) SA	(4) Courts	(5) Jail	(6) TANF	(7) SNAP	(8) Housing
Opened and Removal Propensity	[0.19]	[0.14]	[0.14]	[0.18]	[0.06]	[0.18]	[0.22]	[0.12]
	0.19	0.57***	0.36**	0.19	0.02	-0.53**	-0.30	-0.17
	[0.22]	[0.18]	[0.16]	[0.26]	[0.08]	[0.22]	[0.28]	[0.18]
	2.47***	2.67***	4.18***	2.09***	0.96***	1.67***	1.17***	1.17***
After referral	[0.17]	[0.16]	[0.18]	[0.19]	[0.08]	[0.18]	[0.23]	[0.15]
8. Excluding you always in Propensity								
Opened no Removal Propensity	0.21	0.27*	0.39***	0.05	0.06	-0.23	0.25	0.03
	[0.18]	[0.14]	[0.13]	[0.16]	[0.06]	[0.18]	[0.21]	[0.12]
Opened and Removal Propensity	0.31*	0.57***	0.33**	0.44**	0.04	-0.34*	-0.16	-0.07
	[0.18]	[0.16]	[0.14]	[0.22]	[0.07]	[0.19]	[0.24]	[0.15]
After referral	2.51***	2.72***	4.24***	2.08***	0.95***	1.70***	1.08***	1.19***
	[0.17]	[0.16]	[0.18]	[0.19]	[0.08]	[0.18]	[0.23]	[0.15]
9. Excluding your year in Propensity								
Opened no Removal Propensity	0.20	0.30**	0.34***	0.09	0.05	-0.24	0.18	0.06
	[0.19]	[0.14]	[0.13]	[0.18]	[0.06]	[0.17]	[0.20]	[0.12]
Opened and Removal Propensity	0.11	0.44***	0.15	0.24	-0.01	-0.14	-0.23	-0.11
	[0.18]	[0.14]	[0.12]	[0.18]	[0.07]	[0.17]	[0.21]	[0.14]
After referral	2.52***	2.69***	4.14***	2.18***	1.02***	1.69***	1.25***	1.16***
	[0.19]	[0.17]	[0.19]	[0.21]	[0.09]	[0.19]	[0.24]	[0.17]
10. Propensity Residualized on Past Year Outcomes								
Opened no Removal Propensity	0.30	0.29**	0.36***	0.06	0.04	-0.21	0.35	0.01
	[0.18]	[0.14]	[0.13]	[0.16]	[0.06]	[0.18]	[0.21]	[0.12]
Opened and Removal Propensity	0.37**	0.59***	0.38***	0.43*	0.02	-0.27	-0.17	-0.08

Continued on next page

Table A.4 – Robustness Tests for Month Results (Continued from previous page)

	Health			Justice		Social Services		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Medicaid	MH	SA	Courts	Jail	TANF	SNAP	Housing
After referral	[0.18]	[0.16]	[0.14]	[0.22]	[0.07]	[0.19]	[0.24]	[0.15]
	2.52***	2.72***	4.25***	2.09***	0.95***	1.71***	1.09***	1.20***
	[0.17]	[0.16]	[0.18]	[0.19]	[0.08]	[0.18]	[0.23]	[0.15]

Notes: $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Table presents results for robustness checks for the month-level analyses for mothers presented in Table 1.4. Each set of three rows shows the estimated coefficients from separate regressions including the interaction of being after a referral with the propensity to open a case without removal, and the removal propensity. The propensities are standardized and outcomes are reported in percentage points (%). Columns list which type of services are received by the parent, from left to right: enrolled in Medicaid, receiving mental health services, receiving substance abuse services, case open in court, being in jail, TANF recipient, receiving food stamps, receiving housing assistance. Section 1 is the original specification from Table 1.4 for comparability. Section 2 controls for the propensity instruments for substantiating different types of allegations. Section 3 keeps only the first referral for each parent. Section 4 excludes all the later data when the investigator assigned becomes the caseworker for the family if the case is opened. Section 5 includes controls for outcomes being different after the referral for whites, those with missing race, referrals reported by the police and referrals reported by a social worker (characteristics that were imbalanced in balance tests). Section 6 truncates the propensity excluding the top and bottom 1% of propensities where there might be concern with regards to monotonicity. Section 7 constructs the propensity by excluding you not only in this current investigation but in all future investigations with that investigator. Section 8 constructs the propensity excluding all investigations done by your investigator that year. Section 9 residualizes the propensity on all characteristics included in the balance test (individual characteristics, referral characteristics, lagged 3m outcomes, lagged outcomes for the past year).

Table A.5: *Moms Receiving Any Services 3-6 years After Referral Relative to Prior (%)*

	Health			Justice		Social Services		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Medicaid	MH	SA	Courts	Jail	TANF	SNAP	Housing
Case Opening Propensity	-0.02 [0.35]	-0.10 [0.28]	0.08 [0.25]	0.02 [0.30]	-0.03 [0.16]	-0.67*** [0.33]	-0.19 [0.37]	0.47 [0.36]
Investigators	460	460	460	460	460	460	460	443
Prior Mean	65.6	23.9	11.5	32.8	5.62	24.4	58.6	7.32
Observations	180597	180597	180597	180597	177670	180366	180597	79494

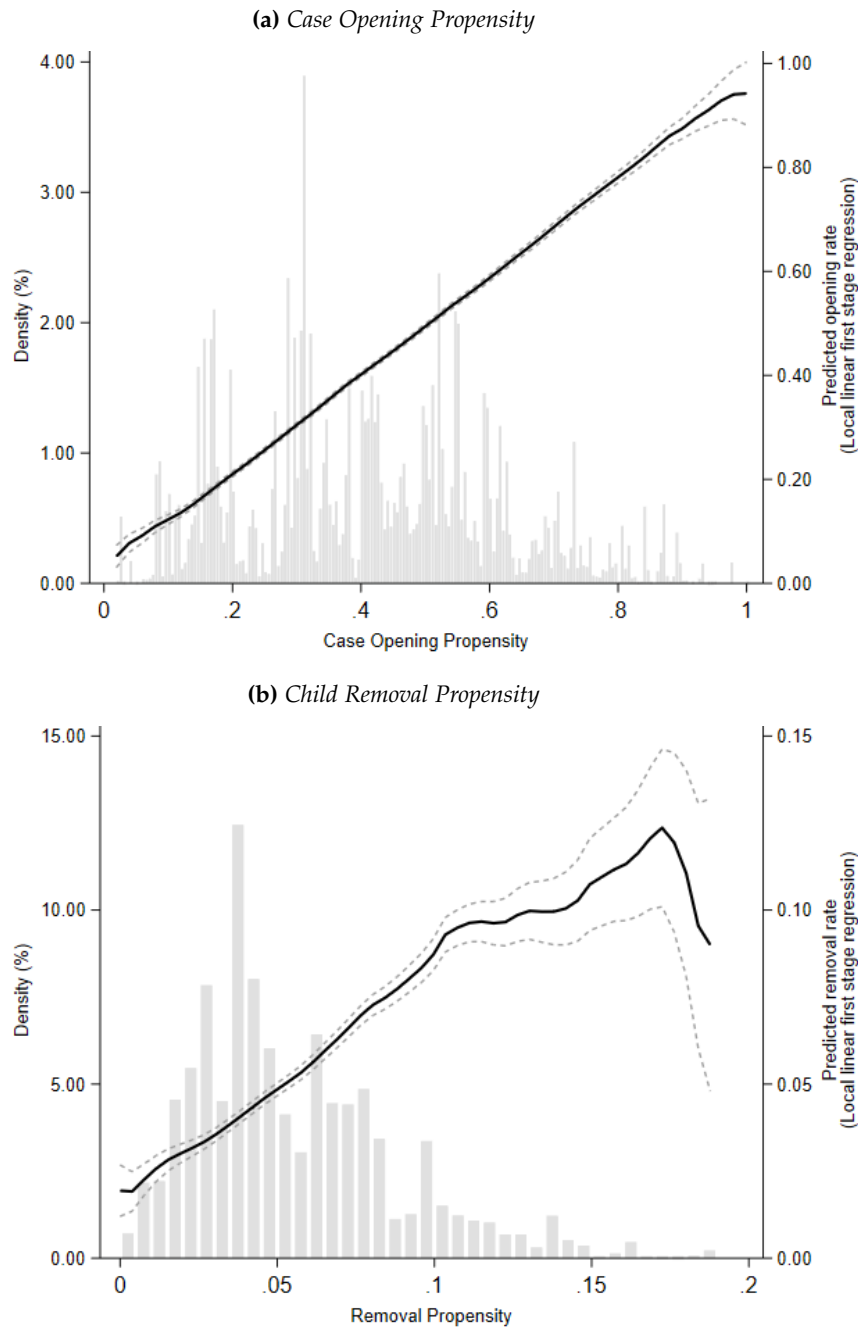
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors are clustered two-way by investigator and individual. Individual-referral, time, and office-year fixed effects included throughout. Propensity is standardized and outcomes are reported in percentage points (%). The variables labeled as the propensities are the interaction of the propensity and being after the referral. Columns list which type of services are received by the parent, from left to right: enrolled in Medicaid, receiving mental health services, receiving substance abuse services, case open in court, being in jail, TANF recipient, receiving food stamps, receiving housing assistance.

Appendix B

Appendix to Chapter 2

B.1 Figures

Figure B.1: Monotonicity Test



Notes: Figures present the distribution of propensities on the left axis for cases assigned to an investigator with at least 50 investigations, and on the right axis the corresponding local linear predictions with 95% confidence intervals. These local linear predictions only include the corresponding propensity instrument as independent variable. They also do not control for differences across offices and years of referral so as to show the raw propensity on the x-axis.

B.2 Tables

Table B.1: *Number of Child-Referrals and Rates of Medically Examined Deaths, May 2009 - Dec 2018*

Sample	(1) Nb Children	(2) Nb Examined Deaths	(3) Rate (‰)
Referred	215202 (100%)	822	3.8
Investigated	108790 (51%)	439	4
Opened a Case	46919 (22%)	239	5.1
Removed	5733 (3%)	29	5.1

Notes: Children may be referred to child welfare several times, and may also be investigated, have a case opened or removed several times from May 2009 to December 2018. This table reports the number of children in each of these groups, with repeat listings of the same child on different referrals counting as separate observations. The percent of referral-specific child observations in each category is in parentheses below the raw counts in column 1. The second column reports the number of children that died among each of these groups of children, and the third columns reports the rate of medically examined deaths per thousand children (ratio of columns 2 and 1).

Table B.2: *Probability of a Medically Examined Death Among Children Always Screened Out*

	(1) Medically Examined Death
Girls	-0.247*** [0.060]
Black American	0.143** [0.064]
Missing Race	-0.141 [0.094]
Hispanic	0.000 [.]
Missing Ethnicity	0.233*** [0.061]
Age	-0.295*** [0.048]
Squared Age	0.025*** [0.007]
Cubed Age	-0.001** [0.000]
TANF Recipient Before	-0.114 [0.117]
SNAP Recipient Before	-0.117 [0.086]
Medicaid Recipient Before	0.256*** [0.069]
Received Mental Health Services Before	0.143* [0.082]
Constant	-2.251*** [0.090]
Pseudo R-squared	0.0780
Observations	40828

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table reports the probit regression results predicting having a child death that was medically examined based on individual characteristics: gender, race, ethnicity, a cubic polynomial of age, and whether the child was a recipient of TANF, SNAP, Medicaid or mental health services at any point before dying. The model is estimated off of the sample of children that were *never* investigated in the decade worth of data that I have (not investigated in this referral nor any other referral where the child may be listed).

Table B.3: *Randomization Test*

	(1) Open Case	(2) Remove
Black American	0.008* [0.005]	-0.000 [0.000]
Missing Race	0.001 [0.003]	0.000 [0.000]
Hispanic	-0.003 [0.006]	0.001 [0.001]
Missing Ethnicity	-0.001 [0.002]	0.000 [0.000]
Girls	-0.003 [0.002]	0.000 [0.000]
Age	-0.000*** [0.000]	-0.000 [0.000]
Referred for Abuse	0.001 [0.007]	0.001 [0.001]
Assigned several investigators	0.012 [0.010]	0.003 [0.002]
Risk score: Moderate Risk	0.001 [0.004]	0.001 [0.001]
Risk score: High Risk	-0.001 [0.004]	0.001 [0.001]
Safety score: Impending Danger	0.001 [0.003]	0.001 [0.000]
Safety score: Present Danger	0.005 [0.004]	0.001** [0.001]
Missing risk score	0.008 [0.008]	0.001 [0.002]
Missing safety score	0.004 [0.004]	0.001 [0.001]
Mandated reporter	-0.016* [0.009]	-0.002 [0.002]
Reporter group: anonymous	-0.001 [0.008]	-0.002 [0.002]
Reporter group: courts	0.014* [0.008]	0.002 [0.001]
Reporter group: family	-0.004 [0.008]	-0.000 [0.002]
Reporter group: health professional	0.007 [0.005]	0.001 [0.001]
Reporter group: neighbor or friend	-0.007	-0.002

Continued on next page

Table B.3 – Randomization Test (Continued from previous page)

	(1) Open Case	(2) Remove
Reporter group: police	[0.010] 0.014***	[0.002] 0.001
Reporter group: school	[0.005] 0.005	[0.001] 0.000
Reporter group: social worker	[0.005] 0.011**	[0.001] 0.001
Response time: <48h (newborns)	[0.005] -0.025	[0.001] -0.008
Response time: within 3-7 days	[0.023] -0.038*	[0.007] -0.010*
Reponse time: within 7-10 days	[0.021] -0.038*	[0.006] -0.009
Response time: immediate	[0.021] -0.027	[0.006] -0.007
Response time: <24h	[0.022] -0.032	[0.006] -0.010*
One father and one mother listed on referral	[0.021] -0.000	[0.006] 0.000
Nb of individuals listed on referral	[0.002] 0.000	[0.000] 0.000
Age of Youngest child on referral	[0.000] -0.000	[0.000] -0.000**
Any mother listed on referral	[0.000] -0.000	[0.000] 0.000
Any father listed on referral	[0.005] -0.004	[0.001] -0.001
Months on SNAP in Past Yr (%)	[0.004] 0.000	[0.001] 0.000
Months on TANF in Past Yr	[0.002] 0.004	[0.000] -0.001
Months Used Mental Health Services in Past Yr (%)	[0.003] -0.000	[0.001] 0.000
Months on Medicaid in Past Yr (%)	[0.003] 0.001	[0.000] 0.000
Months with Court Case in Past Yr (%)	[0.002] 0.012***	[0.000] 0.001
Trend in use of TANF, Past Yr	[0.004] 0.009	[0.001] 0.000
Trend in use of SNAP, Past Yr	[0.023] -0.014	[0.004] -0.001
	[0.018]	[0.003]

Continued on next page

Table B.3 – Randomization Test (Continued from previous page)

	(1) Open Case	(2) Remove
Trend in use of Mental Health Services, Past Yr	-0.010 [0.024]	0.005 [0.004]
Trend in use of Medicaid, Past Yr	-0.005 [0.018]	0.002 [0.004]
Trend in Court involvement, Past Yr	0.051 [0.032]	0.010* [0.005]
Number of investigators	226	226
F-test	1.689	1.395
P-value	0.00809	0.0647
Observations	75329	75329

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table reports the balance test for children assigned to investigators more likely to open a child welfare case (column 1) and more likely to remove children (column 2). Standard errors are clustered two-way for investigator and child. The reference groups for risk score is low risk, for safety score is no present or impending danger, for reporters is other, and for response time is missing. At the bottom of the table I report the statistic of an F-test with its corresponding p-value.

Table B.4: Tests for Monotonicity

	(1) First Stage	(2) Reverse Sample First Stage
Blacks	0.668*** [0.038]	0.508*** [0.048]
Non-Blacks	0.677*** [0.042]	0.549*** [0.045]
Girls	0.692*** [0.037]	0.744*** [0.036]
Boys	0.702*** [0.037]	0.752*** [0.033]
Abuse	0.348*** [0.119]	0.108* [0.060]
Neglect	0.729*** [0.032]	0.079 [0.059]
Mandated Reporter	0.715*** [0.041]	0.451*** [0.048]
Not a Mandated Reporter	0.648*** [0.054]	0.597*** [0.056]
Infant (<1y)	0.693*** [0.059]	0.750*** [0.059]
Non-Infant	0.697*** [0.037]	0.476*** [0.041]
Toddler (1-3y)	0.660*** [0.052]	0.742*** [0.047]
Non-Toddler	0.704*** [0.035]	0.606*** [0.035]

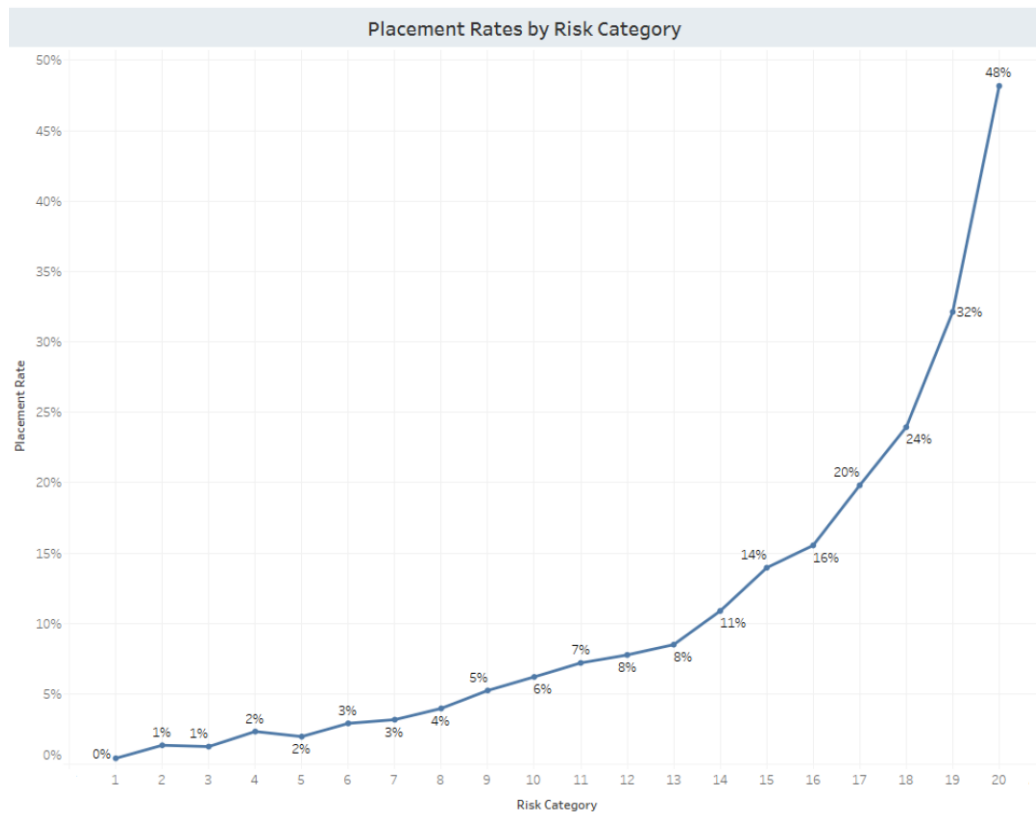
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Standard errors are clustered at the investigator level. Office-year fixed effects are included throughout. The sample is restricted to investigators overseeing at least fifty referrals. This table reports a subset of the tests run to assess whether there were any monotonicity violations. Column (1) reports the first stage coefficient on the propensity to open a case for the sample indicated in the left-most column. Column (2) reports the first stage coefficient on the propensity to open a case for the sample in the left-most column, where the propensity is constructed for the opposite sample. So for the first row, the first column reports the first stage coefficients for blacks using the same propensity as in the rest of the paper. The second column reports the first stage coefficients for blacks where the propensity is constructed using the sample of non-blacks that the investigator investigated.

Appendix C

Appendix to Chapter 3

C.1 Figures

Figure C.1: *Placement Risk by Risk Score on Validation Holdout Sample*



Notes: This graph shows the estimated risk for a child to have an out of home placement (a removal) in the two years following a child welfare referral by child risk score level as predicted by the LASSO model. Estimates come from a validation sample which was not used to train the model.

C.2 Tables

Table C.1: *Characteristics of Those with a Score Reported to be Seen*

Characteristic	(1) Estimate	(2) Relative Likelihood
White	.002 (.056)	1.02
Black	-.256 (.168)	.63
Missing Race	.013 (.056)	1.02
Hispanic	.017 (.064)	1.03
Female	-.027 (.037)	.99
Top third of scores (T3)	.009 (.056)	1.04
Child in Danger	-.331 (.207)	.51
Alleged Injurious Environment	.134** (.059)	1.09
Alleged Physical Abuse	.035** (.068)	1.05
Alleged Parental Substance Abuse	-.041** (.074)	.95
Team 1	-.097* (.057)	.91
Team 2	.006* (.058)	1.01
Randomization Step 2	.351*** (.029)	1.54
Randomization Step 3	.05*** (.056)	1.03
Randomization Step 4	-.099* (.058)	.88

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the household level since randomization is done at the household level. A control for whether this first referral was part of an ongoing episode is included throughout. This table presents how likely caseworkers wrote down that they saw the score among treated individuals with each characteristic. Column 1 reports the coefficient and standard error of a regression estimating the interaction of the characteristic listed in the left-most column with the treatment status (main effects not reported). Column 2 then reports relative to the mean how much more or less likely individuals with this characteristic were to have a score reported. For example, the relative likelihood of 0.63 for Black Americans indicates that the caseworkers were 37% less likely to have Blacks' score seen and/or reported relative to the full sample.