NBER WORKING PAPER SERIES

THE CAUSAL IMPACT OF REMOVING CHILDREN FROM ABUSIVE AND NEGLECTFUL HOMES

Anthony Bald
Eric Chyn
Justine S. Hastings
Margarita Machelett

Working Paper 25419
http://www.nber.org/papers/w25419

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2019, Revised September 2019

The opinions and conclusions expressed are solely those of the authors. We thank the team at Research Improving People's Lives (RIPL), the state of Rhode Island, and the Rhode Island Department of Children, Youth, and Families for making this project possible. We are also thankful for helpful comments and feedback from Joseph Doyle, Sara Heller, Brian Jacob, Jonah Rockoff, Maya Rossin-Slater, Bruce Sacerdote, Emilia Simeonova, Sarah Turner, Christina Weiland, and seminar participants at the University of Michigan, the 2018 Annual Meeting of the Southern Economic Association, and the NBER Children's Program Meeting. Finally, we are grateful for financial support from the Smith Richardson Foundation and the Laura and John Arnold Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Causal Impact of Removing Children from Abusive and Neglectful Homes
Anthony Bald, Eric Chyn, Justine S. Hastings, and Margarita Machelett
NBER Working Paper No. 25419
January 2019, Revised September 2019
JEL No. H75,I21,I24,I28,I38,J12,J13,J24

ABSTRACT

This paper measures impacts of removing children from families investigated for abuse or neglect. We use removal tendencies of child protection investigators as an instrument. We focus on young children investigated before age 6 and find that removal significantly increases test scores and reduces grade repetition for girls. There are no detectable impacts for boys. This pattern of results does not appear to be driven by heterogeneity in pre-removal characteristics, foster placements, or the type of schools attended after removal. The results are consistent with the hypothesis that development of abused and neglected girls is more responsive to home removal.

Anthony Bald
Harvard University
79 John F. Kennedy St
Cambridge, MA 02138
abald@g.harvard.edu

Eric Chyn
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
eric.t.chyn@dartmouth.edu

Justine S. Hastings
Brown University
Department of Economics
64 Waterman Street
Providence, RI 02912
and NBER
justine.s.hastings@gmail.com

Margarita Machelett
Department of Economics
Brown University
margarita_machelett@brown.edu

1 Introduction

Each year, child protective service agencies in the U.S. investigate more than 4 million allegations of abuse or neglect (U.S. Department of Health and Human Services, 2016). As a result of these investigations, authorities annually remove nearly 200,000 children from their homes and place them into foster care (U.S. Department of Health and Human Services, 2016). The goal of removal is to protect children by reducing exposure to abuse and neglect.

There is relatively little evidence on the causal impact of child protective service removal on children. Abused children have lower academic performance and are more likely to have social or emotional conditions such as aggressive behavior or depression (Fantuzzo and Mohr, 1999; Wolfe et al., 2003; Holt et al., 2008; Doyle and Aizer, 2018).[1] Because removal is more likely in acute cases, the relationship between removal and outcomes may not be causal. Doyle (2007; 2008) addressed the endogeneity of removal from home by using the removal tendencies of quasi-randomly assigned child protective service investigators as an instrument for removal. He studied later life outcomes of older children who were subject to investigation between the ages of five and fifteen using data from Illinois and found that removal increased delinquency and arrests while decreasing labor market activity.

This paper focuses on young children and provides new evidence on the impact of removal based on comprehensive administrative data from Rhode Island. The data contain approximately two decades of child protective services case records joined to administrative records on academic outcomes in public schools. We study impacts of removal in early childhood (prior to age six) for two reasons. First, nearly half of removed children are under the age of six (U.S. Department of Health and Human Services, 2016). Second, the literature on child development suggests that early life events and interventions can have particularly strong influences on outcomes (Cunha et al., 2006; Heckman, 2006; Cunha and Heckman, 2007; Almond and Currie, 2011; Heckman et al., 2013; Heckman and Mosso, 2014; Elango

[1] Currie and Tekin (2012) study long-term outcomes of children, finding that maltreatment is associated with increases in the likelihood of committing crime.

et al., 2015; Almond et al., 2017). Our analysis is the first to estimate causal impacts of home removal for this important group of children.[2]

We use the removal tendency of child abuse investigators as an instrument for removal. Our main specification uses a standard leave-out mean removal rate as the measure of the tendency for each investigator. Prior literature has used this type of measure for judges and other authorities (Kling, 2006; Doyle, 2007, 2008; Aizer and Doyle, 2015; Bhuller et al., 2016; Eren and Mocan, 2017; Sampat and Williams, 2015; Dobbie, Goldin and Yang, 2018; Dobbie, Grönqvist, Niknami, Palme and Priks, 2018; Bhuller et al., 2018). We calculate the removal rate for all other cases assigned to an investigator using data from the Rhode Island Department of Children, Youth and Families (DCYF).[3] In our robustness checks, we allow the removal rate to vary by child and case characteristics and use this in an instrumental variable (IV) approach that relaxes the monotonicity assumption inherent in an IV approach using the standard measure (Mueller-Smith, 2015). To do this parsimoniously and avoid overfitting, we use the Least Absolute Shrinkage and Selection Operator (LASSO), a machine learning (ML) regularized regression technique, to select the child and case characteristics that define removal tendency by subgroup (Belloni et al., 2014). We present all regression results separately for girls and boys. Our analysis of effects by gender is motivated by prior research, which shows that girls and boys may respond differently to social programs and family conditions (Heckman et al., 2010; Bertrand and Pan, 2013; Heckman et al., 2013; Elango et al., 2015; Conti et al., 2016; Heckman et al., 2017; Garcia et al., 2018; Autor et al., 2019).

Our main finding is that there are significant and positive effects of removal on achievement outcomes for young girls and no corresponding significant effects for young boys. For

[2]Note that age six is the compulsory school starting age in Rhode Island during our sample period (Rhode Island, 2016). Benson and Fitzpatrick (2018) provide evidence that reports of child maltreatment increase when children enroll in school. Their findings suggest that the composition of investigated children may change at age 6 because educators may be an important source of information for instances of neglect and abuse.

[3]In our sample, the leave-out removal rate is a statistically significant predictor of removal and is uncorrelated with child and case characteristics.

young girls, the point estimate for the impact of removal indicates a 1.33 student-level standard deviation increase in average standardized test scores (math and reading) in the years after removal. These large effects are similar to findings from the Perry Preschool program, where girls randomly assigned to receive high-quality early education had 0.806 higher standardized test scores (Heckman et al., 2013). There are no statistically significant impacts of removal on standardized test scores for young boys. In our main specification, the point estimate is imprecise and suggests a negative impact of 0.06 student-level standard deviations.

In line with the results for test scores, we find that removal has beneficial impacts on additional measures of schooling achievement. For young girls, we find that removal reduces the likelihood of repeating a grade by 22.8 percentage points. Removed young girls are also significantly less likely to participate in special education.[4] As with the test score impacts, we find no detectable impacts of removal on grade repetition or special education participation for young boys.

We examine whether these results are due to multiple hypothesis testing or attrition in the form of changes in public school enrollment. Following Anderson (2008), we calculate adjusted "q-values" that control for the false discovery rate (FDR). Using the set of results for gender subgroups, we find that the estimates for young girls are significant at the 10 percent level using the FDR-adjusted q-values. We also study attrition and find no statistically significant impacts of removal on enrollment for young girls or young boys. The point estimates for girls and boys are not statistically different.

Next, we investigate potential explanations for the differences the impacts of removal on test scores by gender. Our analysis provides suggestive evidence that the pattern of results stems from differences in how girls and boys respond to removal. The strongest evidence to support this interpretation comes from analysis of siblings. In a within-family analysis, the pattern of point estimates is in line with our main analysis and suggests sisters and brothers

[4] We measure participation based on whether the child has a written Individualized Education Program (IEP). An IEP can be given as early as pre-school, and children are assessed each year until they are deemed to no longer be in need. Note that having an IEP does not generally exempt a student from testing in Rhode Island. In academic year 2013, 89 percent of Rhode Island students with an IEP took standardized exams.

respond differently to removal. Further, we find that young girls and boys have similar placement outcomes after removal (i.e., type of foster care and days spent in the foster care system) and attend schools with similar types of characteristics (in terms of school value-added and student body composition).[5] We also find little evidence that suggests the heterogeneous impacts on achievement are due to differences between girls and boys in terms of complier characteristics or parental responses to removal.[6]

As a final analysis, we study removal for older children (investigated at age six or later). We study schooling outcomes and later-life outcomes such as juvenile delinquency, high school graduation, the likelihood of having a teen birth, and post secondary school enrollment. This analysis of later-life outcomes focuses on older children because a child removed at a young age will not be old enough for us to observe outcomes by the end of our sample period. For older children of either gender, we find no statistically significant effects on any outcome. The point estimates tend to suggest that removed boys have worse outcomes in terms of juvenile convictions and high school graduation. For older girls, the non-significant point estimates have no consistent pattern. That is, the signs of the point estimates do not consistently indicate positive impacts for beneficial outcomes or negative estimates for disadvantageous outcomes. The estimates are sufficiently imprecise that we cannot reject that the effects of removal are equal for older girls and older boys.

Overall, these findings contribute to a broad literature on the impact of interventions for children from disadvantaged backgrounds that shows early-life interventions can have large causal impacts on children's outcomes (Garces et al., 2002; Ludwig and Miller, 2007; Almond et al., 2010; Heckman et al., 2013; Campbell et al., 2014; Elango et al., 2015; Aizer et al., 2016; Chetty et al., 2016; Hoynes et al., 2016; Isen et al., 2017; Chyn, 2018; Currie et al., 2018; Garcia et al., 2018).[7] Our results extend this literature by focusing on interventions

[5]We exclude sex abuse reports from all analyses because they make up only 5 percent of all investigations.

[6]For example, the share of compliers that have a married parent is similar in the young girl and young boy samples. To analyze parental behavior, we study parent perpetrators of abuse and neglect. Approximately 95 percent of the perpetrators in our sample are parents. Using samples of parent perpetrators for young girls and young boys, we find no statistically significant impacts of removal on criminal charges and incarceration.

[7]See Almond and Currie (2011) and Heckman and Mosso (2014) for a review of the literature on child

for young children at risk of abuse and neglect, and suggest that the impacts of removal found among young girls may be particular to age. Our findings complement the results from a growing set of studies showing heterogeneous program impacts by gender. As in our results, a number of studies find that schooling and social program interventions can have larger positive impacts for girls (Hastings et al., 2006; Kling et al., 2007; Angrist and Lavy, 2009; Heckman et al., 2013; Deming et al., 2014; Hoynes et al., 2016; Garcia et al., 2018).

2 Background: Child Protective Services and Case Assignment in Rhode Island

Figure 1 illustrates the process for child abuse and neglect investigations and home removal decisions in Rhode Island. An investigation of child abuse or neglect begins when an allegation is reported to the DCYF Child Protective Services (CPS) hotline.[8] The CPS hotline workers record details of the allegation, identify previous or pending investigations, and determine whether the report meets the criteria to initiate an investigation. If the criteria are not met, records of the allegation are expunged from CPS records after a specified period. If the allegations meet the criteria for an investigation, a CPS report is created and forwarded to the central Investigative Unit (IU) where a supervisor assign the case to a field Child Protective Investigator (CPI).[9]

The supervisor assigns the authorized reports using an internal "rotation list," which effectively randomizes the assignment of cases to available field CPIs. This rotation list is an ordered spreadsheet of CPIs, which does not depend on investigator characteristics such as age, ethnicity or any geographic consideration. Each day, the supervisor assigns cases as they arrive based on this ordered list, and CPIs with non-assigned cases are moved to the

development and the impact of interventions for children.

[8]Details on DCYF policies and procedures come from conversations with DCYF staff and documentation from the 2018 DCYF Policy Manual (Rhode Island Department of Children, Youth, and Families, 2018).

[9]In Rhode Island, there is one central Investigative Unit, which assigns cases regardless of geography.

top of the list for the next day's rotation.[10,11] In interviews, supervisors who assign cases state that the goal of the list is to provide "fairness" so that each field CPI will receive similar cases. The only exception for assigning cases through the list is when there is an allegation of sex abuse, in which case the supervisor may assign the case to a CPI of the same gender as the victim.[12] Every case assigned outside of the rotation list is flagged in the case management system, and we use this flag to exclude cases from our analysis (we discuss the sample criteria further in Section 3).

The CPI investigating the case decides whether there is enough evidence of abuse or neglect to warrant out-of-home placement.[13] If there is sufficient evidence, the CPI petitions the Rhode Island Family Court (RIFC) for removal of the child and placement into DCYF custody. According to conversations with DCYF staff, the RIFC typically follows the recommendations made by investigators. The average investigation (including those that do not end in removal) lasts less than one month.[14]

CPIs have limited ability to impact investigated children and their families other than through the removal decision.[15] The circumstances of the case largely determine the type of placement and the duration of time in the foster care system. DCYF places children in a family setting (relatives or a licensed foster family) or in a supervised environment such as a group home or shelter. The field CPI is not involved in a case once the investigation is closed following the removal decision. After removal, case management is handled by a

[10] Cases left unassigned on a day can be voluntarily picked by CPIs outside of this rotation list. These cases are flagged and excluded from the analysis.

[11] The supervisor uses the rotation list to assign cases even when the child has had previous investigations.

[12] Note that sex abuse cases comprise 5 percent of all investigations, and we exclude these from our analysis.

[13] The assigned CPI also makes decisions about whether an allegation of abuse or neglect is indicated or unfounded (see Figure 1). DCYF dismisses unfounded allegations, and children are not removed in those cases. The reports associated with unfounded cases are kept in the DCYF system and removed after a specified period. We obtained a limited sample of unfounded cases and found no statistically significant relationship between the CPI's substantiation rate (the rate of determining that an allegation in a case is founded) and the removal rate for the CPI's founded cases.

[14] In the sample of first investigations (described in Section 3), the average investigation lasts about 22 days. In cases where the CPI recommends removal, the average duration is 11 days.

[15] In Section 4.4, we provide a detailed discussion of the exclusion restriction necessary for our empirical analysis.

social worker.

When a child is in DCYF custody, parents can work with case workers to arrange visits, although the frequency of visitation varies depending on case-specific factors. DCYF releases children from custody due to reunification with parents, adoption, or aging out of the child welfare system by reaching the legal adult age. Reunification with parents occurs only after a parent has completed conditions stipulated by DCYF (e.g., parents may be required to follow a visitation plan or complete mental health counseling with a DCYF service provider). DCYF case management (intake) workers monitor whether a parent complies with conditions for reunification.

3 Data

We use data from anonymized administrative records housed in a secure enclave. All personally identifiable information has been removed from the data and replaced with anonymous identifiers. These identifiers allow researchers with approved access to join records associated with an individual across a range of social programs and government services (Hastings, 2019; Hastings et al., 2019). This section describes the samples and key measures that we construct. Appendix B provides further details and statistics on the process for joining records.

3.1 Sample of Children Investigated at Young Ages

There are 32,845 DCYF investigations that occurred between January 1, 2000, and December 31, 2015.[16] From these data, we create a sample of investigated children (with substantiated founded reports of abuse or neglect) based on three main restrictions. First, we exclude sex abuse investigations and investigations where the Investigative Unit supervisor assigned the case without using the rotation list ($N = 7,533$). Second, we drop investigations that occur after the first investigation associated with a child (ages 0-18) ($N = 5,474$). Third, we drop investigations assigned to CPIs with outlier removal tendencies and exclude

[16] See Appendix B for further details on the process for data cleaning and sample construction.

investigations assigned to CPIs who received less than 10 cases ($N = 508$).[17] This leaves us with 19,330 investigations. Of these investigations, 13,834 involve children *under* the age of six, referred to as "young" children from here forward.[18] In this young child investigations sample, there are 6,449 girls and 7,385 boys.

3.2 School-Age Academic Outcomes

We join the sample of investigated young children to records from the Rhode Island Department of Education (RIDE), which are available for the academic years 2003-2016. The RIDE data include records for public school enrollment, identifiers for the school and grade enrolled, receipt of special education services as indicated by receipt of a written Individualized Education Program (IEP), and school attendance. Standardized test scores in reading and math are available in a subset of academic years (2005-2016).

Standardized test scores for exams taken during grades 3-8 are the main post-investigation outcome that we study. We construct a panel at the academic year level. Investigated children who were born before 1995 or after 2008 will *not* have observations in the panel because they are either too old or young to be enrolled in the testing grades (3-8) during the period 2005-2016. We focus on the average of the scores in math and reading (standardized by grade and academic year). There are 2,721 girls and 3,148 boys that have test scores for both exams in at least one year of the panel. Note that in a given academic year, there are no data for children who enrolled in a private school, although we do have test scores for children enrolled in charter schools.

We study additional post-investigation school outcomes such as grade repetition, participation in special education, and average attendance. We use enrollment records from RIDE to measure whether a child repeats a grade during grades 3-8 (which correspond to the grades that we study test score outcomes). This grade repetition outcome is only defined

[17] We define outliers as values of CPI removal tendency that fall below the first and above the ninety-ninth percentiles.

[18] In Section 5.5, we provide results using alternative age ranges to define a sample of young children. In Section 7, we also report results studying children who were aged 6 to 18 at the time of an investigation.

when individuals are enrolled in two consecutive years. There are 2,778 young girls and 3,225 young boys for whom we can measure grade repetition. For special education participation, we measure whether a child ever has a written IEP during grades 3-8.[19] A child who has an IEP has at least one of the thirteen disability categories as defined by the Individuals with Disabilities Education Act (IDEA).[20] The determination of an IEP can start as early as pre-school, when the child is three to four years old. Over half of students with an IEP in Rhode Island are identified with special needs prior to entering first grade.[21] For absences, we measure the number of absences within a school year and compute the mean across grades 3-8. For consistency, we examine IEP status and absences only for children who have the retention outcome defined.

Finally, RIDE enrollment records allow us to construct two additional types of outcomes. First, we use the end-of-year enrollment files from RIDE to measure whether a child is enrolled during ages 8-13. This age range corresponds to the grades that we use to analyze achievement and other schooling outcomes. We construct this measure for the investigated children with a resulting sample of 4,101 young girls and 4,750 young boys. Second, we study school mobility and characteristics of schools attended. For school mobility, we construct a measure of switching public schools. For characteristics, we construct school-level measures of test score value-added, mean test scores, the fraction of enrolled students who are minorities, and the fraction of students who receive a free or reduced-price lunch (FRL). Value-added for each school is estimated using all years available for the school and excluding the students in our DCYF investigation sample. We regress average standardized test scores (the average

[19] Not all children will have a complete set of years for which we can measure IEP enrollment. For example, if a child transfers (permanently) from public to private school in fifth grade, we would only observe IEP enrollment from third to fourth grade. We retain these children in our analysis and compute IEP participation for the grades available. Similarly, children born before 1995 or after 2008 can only be observed in a partial set of academic years due to the limited coverage of the RIDE data (from 2003-2016).

[20] The categories are: autism, deaf-blindness, deafness, developmental delay, emotional disturbance, hearing impairment, intellectual disability, learning disability, orthopedic impairment, speech or language impairment, traumatic brain injury, visual impairment including blindness, and other health impairment (Individuals with Disabilities Education Act, 2004).

[21] About 28 percent of children receive their IEP for the first time in kindergarten. An additional 25 percent of children receive an IEP before starting kindergarten and enroll in an Early Childhood Special Education program for young children with development delays and disabilities, as mandated by IDEA.

of math and reading scores) on lagged test scores (including their square and cube), as well as indicators for a student's race, gender, special education status, English learner status, and free or reduced price lunch status. We use the mean residuals within a school as a single measure of value-added.[22,23] The fraction of minority and fraction of FRL students at a school are calculated in each year. We join these school characteristics to a child-level panel (covering grades 3-8) to measure the impact of removal on the characteristics of the schools attended post-investigation. There are approximately 2,800 young girls and 3,300 young boys for whom we can measure mobility and school characteristic outcomes.[24]

3.3 Parent Perpetrators and Crime Outcomes

We also study impacts of removal on outcomes of parents of investigated children. We obtain information on parents from DCYF records on perpetrators associated with an investigation.[25] For young children in our sample, 95 percent of children have at least one parent listed as a perpetrator. We use this information to create a sample of parent perpetrators. We join this sample to criminal charge and incarceration records (1995-2017) from the Rhode Island Department of Corrections (RIDOC). The unit of analysis is a parent perpetrator, and the outcome is whether a parent perpetrator is charged or incarcerated at any point in the two or four years following the conclusion of an investigation. Because the criminal justice data source ends in 2017, these measures will be partially censored depending on the end date of the associated investigation.[26]

[22] Our approach follows prior studies such as Kane et al. (2008) and Chetty et al. (2014).

[23] See Appendix 3 for details on the estimation of school value-added and how we join this measure to the student-level data.

[24] Sample sizes slightly vary across these outcomes due to missing data. See Appendix B for further details.

[25] The DCYF investigation records have information on household characteristics, but there is no information on parent identity aside from the information contained in perpetrator records.

[26] There are 6,252 parent perpetrators associated with young girls and 7,141 parent perpetrators associated with young boys.

3.4 Descriptive Statistics

Table 1 presents summary statistics for the sample of investigated young children in Rhode Island. Column 1 shows that 58 percent and 16 percent of investigated children are white and Hispanic, respectively.[27] Race in the sample differs notably from Doyle (2007; 2008), which studied the impact of removal for a sample from Illinois where 76 percent of investigated children were African American. This contrast partly reflects differences in the demographics. That is, the African American shares of children in Rhode Island and Illinois are 9 and 15.8 percent, respectively (U.S. Census, 2018). In terms of family background, only 21 percent of the investigated children in Rhode Island are from married households.

The DCYF data report all allegations associated with an investigation. An allegation of neglect occurs in about 80 percent of investigations. Allegations of physical abuse or physical neglect (i.e., neglect that results in a physical injury) occur much less frequently in about 14 and 7 percent of investigations, respectively. These statistics are broadly in line with national statistics, where allegations of neglect and physical abuse occur in 75 and 18 percent of investigations, respectively (U.S. Department of Health and Human Services, 2016).

We also observe the reporter associated with child abuse investigations. For 82 percent of children, the reporter in the case was a professional such as a teacher, physician, social worker, or police officer. The remaining fraction of reports are provided by family, friends, or other individuals such as neighbors or anonymous reporters.

Column 1 shows that removal from home occurs in 20 percent of the sample of first investigations. This is lower than the removal rate observed in Doyle (2007; 2008), which studied older children from Illinois when the state's placement rate (27 percent) was one of the nation's highest in the 1990s. Columns 2 and 3 provide separate summary statistics for children subject to investigations that do not and do result in home removal, respectively.

[27]Nationally, 45 percent of child abuse victims were white, and 22 percent were Hispanic (U.S. Department of Health and Human Services, 2016).

Column 4 reports the p-values from tests of differences in means for each summary statistic. Investigations that do not end in removal have significantly different child and case characteristics from investigations where removal does occur. Children who are not removed are slightly older than those who are removed (1.9 years old versus 1.1), live in households with 7 percentage point higher marriage rates (p-value < 0.01), and are about 4 percentage points (p-value < 0.01) less likely to be African-American. The final row of Table 1 shows that Rhode Island children who are removed spend roughly 450 days in foster care, which is less than the average four-year stay in Doyle's (2007; 2008) study of Illinois.

4 Empirical Strategy

Consider the following regression model of the relationship between child outcomes and removal:

$$Y_i = \beta_0 + \beta_1 R_i + \beta_2 X_i + \epsilon_i, \tag{1}$$

where Y_i is a post-investigation outcome for child i, R_i is an indicator for whether the child was removed during the first investigation, X_i is a vector of child and case characteristics (including fixed effects for the investigation year), and ϵ_i is an error term.[28] Standard OLS estimates of Equation 1 will be biased if home removal (R_i) is correlated with unobserved determinants of child outcomes (ϵ_i). The descriptive statistics in Table 1, as well as prior research, suggest that observed and unobserved family and home conditions affect both the likelihood of removal and child outcomes (Berger et al., 2009, 2014; Wildeman and Waldfogel, 2014).

To address the endogeneity concern in Equation 1, we rely on an instrumental variable (IV) strategy that is based on a measure of the removal tendency of the investigator j who handles case c associated with child i. We denote the removal tendency as Z_{ijc}, and the

[28]Specifically, X_i contains the child and case characteristics listed in Table 1 and investigation year fixed effects.

first-stage equation is:

$$R_i = \alpha_0 + \alpha_1 Z_{ijc} + \alpha_2 X_i + \nu_i, \tag{2}$$

where Z_{ijc} is a leave-out removal tendency measure that is similar to measures calculated in the literature using judge decision tendencies as instruments for individual case decisions (Kling, 2006; Doyle, 2007, 2008, 2013; Dahl et al., 2014; Aizer and Doyle, 2015; Bhuller et al., 2016; Mueller-Smith, 2015; Eren and Mocan, 2017; Sampat and Williams, 2015; Dobbie, Grönqvist, Niknami, Palme and Priks, 2018; Dobbie, Goldin and Yang, 2018; Bhuller et al., 2018). In our context, we construct this measure to account for the fact that 30 percent of the cases in the DCYF sample include siblings.[29] We exclude the focal child and siblings on case c by defining the leave-out removal tendency for each case as:

$$Z_{ijc} = \frac{1}{N_j - n_c}\left(\sum_{k \neq c}^{N_c} R_k^{'}\right), \tag{3}$$

where N_j is the total number of children assigned to the investigator j, n_c is the number of children on case c, and N_c is the number of cases for the investigator. We define k to index the cases handled by investigator j, and $R_k^{'}$ is the number of children removed on case k. The leave out measure is an average that excludes children on the same case. We calculate the measure of removal tendency using all cases for investigator j within an eight-year window.[30] When we estimate Equation 1 using this leave-out measure, we report two-way clustered standard errors at the investigator (CPI) and case (i.e., family) level.

If there are heterogeneous impacts of removal, we must make two assumptions to interpret IV estimates of the parameter β_1 from Equation 1 as a local average treatment effect (LATE)

[29] See Appendix Section B.5 for further details on siblings.

[30] Ideally, we would calculate the leave-out instrument within each year to allow CPI tendency to evolve over time. In practice, we find that a version of our leave-out instrument defined at the yearly level has a relatively weak first stage. Instead, we allow CPI removal tendency to vary by calculating the measure separately for the 2000-2007 and 2008-2015 periods, respectively. As discussed in Section 5.5, our analysis is robust to using alternative definitions for defining the leave-out measure of CPI removal tendency.

of removal for marginal investigations (Imbens and Angrist, 1994). First, the measure of CPI removal tendency defined in Equation 3 must affect child outcomes only by changing the probability of removal. In Sections 4.3 and 4.4, we provide evidence that suggests this assumption is plausible in our setting by examining random assignment of investigators and analyzing whether CPI removal tendency is correlated with other post-removal decisions such as the type of placement or whether police were notified during an investigation.

Second, we assume that there is a monotonic impact of CPI assignment on removal across children. A violation of this assumption may occur if CPI removal tendencies vary with case characteristics. For example, a given CPI may be relatively strict when it comes to removing African American children, but lenient when it comes to removing all other children. If there is a non-monotone impact of removal tendency, the IV estimate would not identify a well-defined LATE.[31] As a test of monotonicity, Section 4.5 follows prior work and shows that the first-stage coefficient for the tendency measure defined in Equation 3 is positive in various subsamples (Bhuller et al., 2016; Dobbie, Goldin and Yang, 2018).[32]

To further address concerns regarding monotonicity, we carry out robustness tests in Section 5.5, which allow the CPI tendency and its impact on removal to vary with observed case characteristics. We do this by creating a set of potential instruments based on leave-out measures of removal tendency calculated for different categorizations of child and case characteristics (i.e., gender, minority status, marital status, reporter type, allegation type, and investigation level). This creates a set of potential instruments. Following Belloni et al. (2012, 2014), we use the machine-learning (ML) algorithm, Least Absolute Shrinkage and Selection Operator (LASSO), to select the instruments with greatest predictive power for

[31]Under non-monotonicity, the IV estimate would be a weighted average of marginal treatment effects where the weights do not sum to one (Angrist et al., 1996; Heckman and Vytlacil, 2005).

[32]In an additional monotonicity test, Section 4.5 also shows that the first-stage coefficient for a reverse-sample tendency measure is positive in various subgroups. The reverse sample tendency is calculated by dividing the sample into subgroups (e.g., by race) and constructing instruments using the complement for each subgroup. For example, we recalculate the removal tendency for white children using all observations outside this subgroup (all non-white children).

removal in the first stage equation.[33,34]

4.1 Variation in Child Protective Investigator (CPI) Removal Tendency

Figure 2 plots the distribution of the leave-out CPI removal tendency from Equation 3 in our sample of investigations of young children. We observe 102 CPIs during 2000-2015. The average number of children seen by CPIs across all years is 387. The mean of the removal tendency measure in Figure 2 is 0.178, while the 25th and 75th percentiles of the distribution are 0.138 and 0.217, respectively. The standard deviation is 0.056.[35] Further statistics and information on the CPIs in our analysis sample are provided in Appendix B.4.

4.2 First Stage Impact

Panel A of Table 2 reports results from Equation 2, measuring the impact of our instruments on whether an investigation resulted in removal of the child from the home. Column 1 shows that the leave-out measure of mean CPI removal tendency is highly predictive of removal. The estimate in Column 1 implies that moving from a CPI in the lowest quartile of removal tendency to one in the highest quartile would increase the likelihood of removal by 4.61 percentage points ($= 0.584 \times 0.079$), relative to a mean removal rate of 20.23 percentage points.[36] Columns 2 and 3 show that the effects by subgroups for gender are similar in magnitude. The point estimates suggest that removal tendency has a larger impact for boys, but we cannot reject the hypothesis of equal first stage impacts between girls and boys.

4.3 Instrument Validity: Testing Random Assignment

According to the assignment process described in Section 2, investigations in our sample should be quasi-randomly assigned to CPIs. To test this implication, we regress the removal

[33] The use of LASSO for regularization is necessary since there are several potential case characteristics by which CPI tendencies may vary. An unrestricted model would likely result in too many instruments and potentially weak instruments, creating challenges for causal inference (Bound et al., 1995).

[34] Further details on our ML approach are provided in Appendix D.

[35] In Doyle (2007), the standard deviation is 9 percent in the delinquency sample, 10 percent in the teen motherhood sample, and 7 percent in the labor market outcomes sample.

[36] Doyle (2007; 2008) discusses the possibility that the coefficient on the impact of CPI removal tendency may be less than one due to measurement error.

tendency measure on baseline child and investigation characteristics. Panel B of Table 2 reports results from a test for joint significance of baseline characteristics in determining investigator removal tendency. We report this for all investigations and subgroups based on gender and age of children at the time of an investigation. Baseline characteristics include the child demographics and case characteristics listed in Table 1. We consistently fail to reject the null hypothesis that the coefficients for investigation characteristics are jointly zero. For example, Column 1 shows that the chi-squared test statistic is 16.50 with a p-value of 0.284 in the sample of all investigations. Appendix Table A1 reports all point estimates associated with the regression we used to conduct tests of joint significance.

4.4 Instrument Validity: Exclusion Restriction

The random assignment of cases to investigators is sufficient for a causal interpretation of the reduced form impact of being assigned to a stricter investigator. However, interpreting IV estimates as measuring the impact of removal in Equation 1 further requires that the removal tendency of an investigator should affect children only through the decision to remove a child from home and not through any other channel. For example, this exclusion restriction would be violated if CPIs also determined the duration of foster care or the type of foster placement. In Appendix Table A2, we test whether CPI removal tendency is correlated with other foster care outcomes for the subgroup of children who have been removed. We also test if there is a correlation between whether police are notified during an investigation. For girls and boys who have been removed, we find there are no statistically significant relationships between removal tendency and the time spent in foster care, the type of placement, or whether police authorities were notified during an investigation. This pattern of results is consistent with the idea that CPIs have limited ability to influence a child's outcomes once a child is placed into DCYF custody (as discussed in Section 2).

4.5 Monotonicity

To interpret IV estimates from Equations 1 as a LATE of removal for marginal investigations, we must assume monotonicity in the impact of the CPI removal tendency on the likelihood of removal across children in our sample. As noted in Bhuller et al. (2016) and Dobbie, Goldin and Yang (2018), one testable implication of monotonicity is that the first stage estimates should be non-negative for any subgroup of the investigations sample. Columns 2 and 3 in Panel A of Table 2 provide an initial indication that there is no evidence of a violation of monotonicity across all cases by showing that the first stage is non-negative for the subgroups defined by gender. Appendix Table A3 expands on these results by providing additional results for narrower subgroups based on various case characteristics. The first stage impacts of removal tendency are consistently positive.[37] An additional implication of monotonicity is that CPIs should be stricter for a specific type of investigation if they are stricter in other investigation types. To test this assumption, we estimate first stage models where we recalculate the leave-out instrument for each subgroup using all investigations outside of the subgroup. For example, we estimate a first stage model for Hispanic children using using the CPI's removal tendency calculated for all non-Hispanic investigations. Results are presented in Appendix Table A4 and are consistently positive and almost always statistically different from zero.

4.6 Interpreting the LATE in Our Analysis

Assuming the exclusion restriction and monotonicity assumptions hold, the IV estimates of the parameter β_1 from Equation 1 are a local average treatment effect (LATE) of removal for children who would have received a different removal decision had their case been assigned to a different investigator. To better understand this treatment effect parameter, we examine characteristics of compliers in our sample of first investigations for girls and boys, separately. To conduct this analysis, we calculate these characteristics following the approach from

[37] The magnitudes of the first stage estimates for subgroups defined by each case characteristic (shown in the rows of Appendix Table A3) are generally similar to the impact in the sample of all investigations.

Abadie (2003), Dahl et al. (2014), and Dobbie, Goldin and Yang (2018).[38]

Each row of Appendix Table A5 provides information on the overall sample mean for a case characteristic and the complier-specific mean. We provide these statistics separately for girls and boys investigated at young ages. For each gender, we see that compliers are generally similar to the average child in our investigation sample. The main exception is that compliers in the sample of young girls are less likely to be white. Comparing Columns 2 and 4, we also see that complier girls and boys have similar characteristics except in terms of race. For example, complier young girls are 9 percentage points (81 percent) more likely to be black than complier young boys. This difference is statistically significant at the one percent level.

5 Main Results

5.1 Standardized Test Scores for Young Children

Table 3 presents estimates of the impact of removal on standardized test scores for young girls (Panel A) and young boys (Panel B). Columns 1 and 2 provide IV estimates for effects on the average of reading and math scores with and without controls for case characteristics. Similarly, Columns 3-6 provide estimates separately for reading and math scores. Robust standard errors that are two-way clustered at the case (family) and investigator level are reported throughout. Note that the sample for test score analysis contains 2,721 and 3,148 children, which differs from the numbers for investigated children in Table 1. As noted in Section 3, investigated children will not be included in the test score panel if they are not old enough by the final year of the available school records (2016) or they do not attend public school in Rhode Island.

The results in Panel A show that the marginal removal has a significant and positive im-

[38]Similar to Dahl et al. (2014) and Dobbie, Goldin and Yang (2018), we define compliers in our setting as children whose removal decision would have been different had their case been assigned to the most lenient versus the strictest investigator. We consider investigators in the top percentile of removal tendency as "strict" and investigators in the bottom percentile of removal tendency as "lenient." See Appendix Section C and the notes to Appendix Table A5 for further details on our calculation of complier characteristics.

pact on the average standardized test scores for young girls. Column 1 shows that the point estimate for removal is 1.334 standard deviations. We obtain nearly identical results when we include controls for case characteristics in Column 2. Results for standardized math and reading scores are similarly large in magnitude and statistically significant. Evaluations of high-quality early education programs targeting disadvantaged children serve as an important point of comparison for these impacts. Heckman et al. (2013), for example, found that the Perry Preschool program increased female standardized test scores by 0.806 standard deviations. As another benchmark, Bharadwaj et al. (2013) and Chyn et al. (2019) find that neo-natal investments for babies born at very low birth weight increase standardized test scores by 0.15-0.34 standard deviations in elementary and middle school.

Our estimated impact is large in magnitude, but note that complier young girls in our sample would have had very low standardized test scores if they had not been removed. Following the approach from Dahl et al. (2014) and Bhuller et al. (2016), we calculate outcomes for compliers if they had not been removed, finding that the mean complier among young girls would have had an average standardized test score of -1.741.[39] This implies that young girls at the margin benefit from removal, but they are still likely to have below average test scores.

In contrast to the results for young girls, Panel B shows that there are no detectable impacts on any measure of test scores of young boys. The point estimates for boys are generally an order of magnitude smaller than what we obtain for girls, although the standard errors in our estimates are large and the confidence interval contains effect sizes that are substantively large. We can consistently reject the hypothesis of equal impacts of removal on test scores by gender (p-value < 0.10).

In Appendix Figure A1, we report separate estimates and confidence intervals for impacts on average standardized test scores in each grade (3-8) in the panel of test scores. For girls, we find positive point estimates that are similar in magnitude across grades. This pattern

[39]For detailed discussion of our calculation of the complier average outcome when not removed, see Appendix C.

suggests that the benefits of removal are persistent and may be due to permanent changes in child ability prior to third grade. We also find that the contrast between the impacts on test scores for girls and boys is constant across grades. For young boys, the estimates are never significant and are generally smaller in magnitude for the grades that we examine.

5.2 Grade Retention, Special Education, and Attendance for Young Children

Table 4 tests for impacts on additional schooling outcomes for children. As discussed in Section 3, we measure impacts on grade retention, special education participation (i.e., having a written IEP) and average absences during grades 3-8. The sample for this analysis differs slightly from the analysis of test scores since we include all enrolled children (regardless of whether they have a valid standardized test score).[40,41] Due to the number of outcomes, Table 4 only reports estimates from a specification with case controls. The estimates are robust to a specification without controls, as reported in Appendix Table A6.

The results in Panel A show that young girls are significantly less likely to be retained during grades 3-8. The point estimate shows that removal decreases the likelihood of any grade repetition by 22.8 percentage points. As with test scores, this impact is large, but the mean rate of repeating a grade when not removed is 29.1 percent for compliers. Panel A also shows that removal has a significant and large (44.1 percentage points) reduction in special education needs as measured by having a written IEP during grades 3-8.[42,43] There is suggestive evidence of a decrease in the mean number of absences for young girls, although the point estimate is not statistically significant.

For young boys, the results in Panel B shows that there are no statistically significant effects of removal on any of these non-testing outcomes. In Columns 1 and 2, the point

[40] Approximately 93 percent of investigated children who are enrolled in public school have a standardized test score.

[41] Section 3 provides details on all measures that we construct.

[42] All measures are constructed using records from grades 3-8. Because enrollment in special education can begin as early as pre-school, we also conducted analysis defining the IEP outcome for grades K-8. We do this for all the additional schooling outcome as well. We find similar results for the measures defined over this grade range. These results are reported in Appendix Table A7.

[43] Note that an IEP does not imply that a student is exempt from testing. In academic year 2013, 89 percent of students with an IEP took a standardized exam.

estimates suggest that removal decreases the likelihood that a young boy is retained or enrolls in special education by 5.3 and 20.1 percentage points, respectively. For absences, the estimate suggests that removal has relatively small benefits in terms of reducing absences.

Overall, the results for retention, special education participation, and absences match the pattern of heterogeneous impacts by gender observed for test scores. To summarize these schooling results, we construct a school index measure, which is the equally weighted average of the standardized (z-score) measures for the three outcomes.[44] One interpretation of this index is that higher values indicate that children have less schooling ability or more difficult experiences in school. Column 4 of Table 4 shows that removal leads to a large and significant 0.92 standard deviation improvement in the school index for young girls. The corresponding estimate for boys is the much smaller in magnitude, and we can reject the hypothesis that the effects on this index outcome are equal for boys and girls at the ten percent level.

5.3 Attrition due to Changes in Public School Enrollment or Test-Taking

A concern for interpreting the test score and schooling results is that removal may affect whether a child attends a Rhode Island public school or sits for a standardized exam. This would generate selection into the panel of test scores and other schooling outcomes that we use for our analysis. To address this concern, we construct a balanced panel with indicators for enrollment and exam taking during ages 8-13. This is the age range that corresponds to grades 3-8, which are the focus of the test score analysis. The sample for this analysis is larger than what appears in Tables 3 or 4 because we include investigated children that never appear in the enrollment records.

Table 5 shows that we do not find any significant impacts of removal on enrollment or test-taking for young girls. The insignificant point estimates suggest that, if anything, removed young girls are more likely to be observed in the test score panel. For young boys,

[44] To standardize each component, we calculate the mean and standard deviation of each outcome using investigated children by gender. Next, we compute the standard score by taking each outcome and subtracting the mean for all investigated children of the same gender and dividing by the standard deviation.

we also find no detectable effects, and the point estimates are negative for both outcomes. We cannot reject the hypothesis of equal impacts on enrollment or test-taking for girls and boys. Overall, this lack of significant impacts suggests that attrition from public school or selective test-taking are unlikely to explain our results for young girls.

5.4 Multiple Hypothesis Testing

Given that our analysis tests for impacts for multiple outcomes, one concern is that the findings for young girls are an artifact of multiple hypothesis testing. To manage the risk of false positives, we follow the recommended practice of adjusting per comparison p-values (Anderson, 2008). We use the two-step procedure from Benjamini et al. (2006) to calculate "q-values" that control for the false discovery rate (FDR), which is the proportion of rejections that are false positives (Type I errors). Appendix Table A8 shows that the IV estimates for test scores and retention of young girls are significant at the 5 percent level after adjusting for the fact that we analyzed multiple outcomes (i.e., impacts for average test scores, retention, participation in special education, and average absences).

5.5 Robustness Tests

Appendix Tables A9–A11 provide robustness tests for the main analysis. We begin with checks related to the construction of the sample. For comparison, Column 1 of Appendix Table A9 provides the estimate for the impact on the average of standardized tests from our preferred specification. Recall that this specification includes only those children investigated before age 6 whose assigned CPI handled at least 10 cases. Columns 2-4 provide results for the samples of children with CPIs who handle at least 100, 200 or 300 cases. The point estimates largely do not change across these alternative samples. The main change for young investigated girls is that we lose statistical significance when we impose the 300-case restriction and exclude approximately half of the original sample. Column 5 shows that our results do not change when we include children associated with investigations involving sex abuse.[45] Columns 6 and 7 test whether the results change when we change the age range used to define the sample of young investigated children. The estimates do not change when we define the sample of young children as those investigated during ages 0-4 or during ages 0-6.

Next, we check whether the results are robust to using an approach that allows CPI removal tendency to vary with case characteristics. As discussed in Section 4, this allows us to relax the assumption of monotonicity necessary to interpret our main results as the LATE of removal for marginal investigations (Imbens and Angrist, 1994). The case characteristics that we consider are sex, race (non-minority and minority), marital status, reporter type, allegation type, and investigation level. For each of these characteristics, we define mutually exclusive groups of children and calculate CPI removal tendency for the group. For example, each CPI will have a leave-out removal tendency calculated separately for non-minority (white) and minority (non-white) children. Since we consider multiple characteristics, we have a set of instruments. We use LASSO to select the instruments with greatest predictive power for removal in the first stage equation (Belloni et al., 2014).[46]

Appendix Table A10 presents the machine-learning (ML) IV estimates for impacts of removal on test scores, grade retention, special education (IEP), and the school index measure from Section 5.[47] The ML IV estimates are similar to the main results that we report in Tables 3 and 4. Panel A shows that we consistently find significant and positive impacts of removal on test scores for young girls. Similarly, the ML IV estimates indicate that removal significantly reduces retention and the likelihood of IEP participation. In Panel B, the ML IV estimates for boys are never statistically significant, and the point estimates suggest that removal has negative impacts on all achievement and schooling outcomes.

[45] Recall that sex allegations are excluded from the main analysis since case assignment to a CPI may take into consideration the gender of the CPI.

[46] See Appendix Section D for further details on the measures and implementation of the LASSO-based approach.

[47] We estimate LASSO separately for separately for each gender and outcome that we consider. The instruments selected change across specifications. Appendix D provides details on our implementation and the instruments selected.

Finally, we test whether our results are robust to the types of investigations or time period used to calculate CPI removal tendencies. Column 1 of Appendix Table A11 reproduces our main preferred estimates for test scores. The main estimate is based on calculating removal tendency with an eight-year period for each CPI and using all children (i.e., those with first and repeat investigations).[48] By calculating the removal rate within an 8-year period, we allow the removal tendency of a given CPI to change over time.[49] Columns 2-4 shows that we obtain similar results when we use a measure of removal tendency that is based only on first investigations or based on pooling investigation decisions for all years (2000-2015).

5.6 Marginal Treatment Effects

To further examine impacts of removal on test scores of young children, we explore heterogeneity by examining marginal treatment effects (MTEs). MTEs are treatment effects for individuals with a particular "resistance" to treatment (Cornelissen et al., 2016). These effects are defined under a generalized Roy model. In our context, let Y_1 and Y_0 denote the potential outcomes if a child is removed or not removed, respectively. We assume that each of these are linear functions of both observable (X) and unobservable factors. The choice to remove a child by a CPI is given by the indicator function $I = 1(v(X, Z)–U)$, where v is any function, Z is the leave-out removal tendency instrument, and U is an unobserved continuous random variable. Since U enters the removal equation with a negative sign, it is interpreted as resistance to treatment (removal). We can re-write the CPI choice equation as $P(X, Z) > U_d$, where $P(X, Z)$ is the propensity score and U_d represents quantiles of the unobserved resistance to removal (U).

The MTE is defined as $\mathbb{E}(Y_{1i} - Y_{0i} | X = x, U_d = u)$, and the dependence of the MTE on U_d reflects unobserved heterogeneity in treatment effects (Heckman and Vytlacil, 1999; Heckman

[48] For the analysis of impacts on outcomes, we only use the first investigation associated with a child. In the construction of the instrument, we use first and repeat investigations. The latter provide more observations for calculating removal tendencies, allowing us to have more statistical power.

[49] As discussed in Section 4, we allow the CPI removal tendency to vary over time by calculating the measure separately for the 2000-2007 and 2008-2015 periods, respectively. Ideally, we would calculate the leave-out instrument for each year to allow CPI tendency to evolve over time. A concern is that some CPIs may see relatively few children within a year, thereby making it difficult to infer their tendency.

et al., 2001; Heckman and Vytlacil, 2005, 2007). As in prior studies, we assume separability between observed and unobserved heterogeneity in treatment effects (Carneiro et al., 2011; Bhuller et al., 2016; Brinch et al., 2017). Given this assumption and the exogenous instrument condition from Section 4, the MTE is identified over the common support of the propensity score $P(X, Z)$ (Carneiro et al., 2011; Bhuller et al., 2016; Brinch et al., 2017). Panels A and B of Appendix Figure A2 shows the propensity score distribution for the removed and non-removed children in the young girl and young boy samples, respectively. The dashed red lines indicate the upper and lower points of the propensity score with common support (after trimming 5 percent of the sample with overlap in the distributions).

Panels A and B of Appendix Figure A3 shows the MTEs for young girls and young boys, respectively. We use a local instrumental variable approach using a global quadratic polynomial specification.[50] We construct confidence intervals using 100 bootstrap replications. The results in Panel A show that the MTE estimates for test scores are most positive for young girls with low unobserved resistance to treatment. The estimates decrease as the unobserved resistance increases and become negative at the highest quantiles. The decline at the upper levels of resistance suggests that young girls on the margin of placement with the highest removal rate CPIs (who likely have less severe unobserved abuse levels) benefit less from removal. For young boys, the results in Panel B show that the MTE estimates are always negative, and the estimates decline with increases in the resistance to treatment.

6 Understanding Gender Differences in the Impact of Removal for Young Children

What explains the pattern of gender differences in the impacts of removal for young children? This section considers three categories of explanations. First, it is possible there are gender differences in the pre-investigation characteristics of compliers that could help determine the effects of removal. Second, removal may have heterogeneous effects on medi-

[50] We conducted robustness checks on the MTE estimates and found similar results when we use linear, quadratic, or cubic specifications.

ating factors such as the type of foster care placement, school mobility or characteristics, or parental behavior. Third, girls and boys may be responding differently to the same treatment of removal in early life.

6.1 Complier Characteristics

One possibility is that the compliers among young girls are different in terms of pre-investigation background characteristics relative to their male counterparts. If effects vary by these complier characteristics, this could explain why we observe gendered effects on test scores and the other schooling outcomes such as grade repetition. As noted in Section 4.6, the average characteristics for compliers are generally similar for young girls and young boys except in terms of racial composition. Specifically, the average young girl complier is much more likely to be a minority relative to their male counterparts.[51]

To understand the importance of race in our analysis, Appendix Table A12 reports impacts of removal by gender and minority status subgroups. These results do not provide strong evidence that the difference in the minority share among girl and boy compliers explains the pattern of test score effects. Although the results are not always statistically significant, Panel A generally suggests that removal has large and beneficial estimated impacts on test scores and the school index measure for both minority and non-minority young girls. In contrast, the results in Panel B show that there are generally no significant impacts of removal for either the group of minority or non-minority young boys.

6.2 Mediating Factors

As detailed in Section 3, we have extensive measures of mediating factors that could help determine the impact of removal. Specifically, we focus on factors such as types of foster care placement associated with the first investigation, school mobility and characteristics, and the behavior of parents. Our focus is on testing whether there are gender differences in

[51]For example, Appendix Table A5 shows that the fraction of compliers who are white is only 42.6 percent for young girls compared with 62.7 percent for complier young boys.

any of these potential mediators.

Table 6 reports impacts of removal on foster care outcomes associated with the first investigation such as the number of days spent in each type of foster care and the likelihood of adoption.[52] The results show little evidence of differences in these post-removal outcomes for young girls (Panel A) and young boys (Panel B). For example, removal has statistically significant and large positive impacts on the days spent in any type of foster care for both genders, and there are no detectable impacts of removal on the likelihood of adoption for boys or girls.[53,54]

Across placement outcomes, the main gender distinction is the effects of removal on the number of days spent in the residual category of "other" DCYF placements.[55] Column 5 shows that removal has no detectable impact on days spent in the other category of foster care for young girls. In contrast, the impact for young boys is significant, but it is worth noting that the point estimate is just 38 days in these residual categories of foster care placement. This impact is approximately 10 percent of the amount of time that boys typically spend in foster care.

Next, we test for gender differences in the impact of removal on school mobility or the types of schools that children attend. Table 7 provides estimates for impacts on schooling mobility and several characteristics of schools attended for grades 3-8. The results provide no strong evidence of gendered treatment effects. For young girls and young boys, there are no statistically significant impacts, and we cannot reject the hypothesis that the point estimates are equal for young girls and boys, though the standard errors are generally large.[56]

[52]The foster care outcomes in Table 6 are based on placement records for the first investigation. Alternatively, we can measure foster care outcomes associated with any subsequent investigation. When we analyze the total days spent in foster care including days from the first and subsequent investigations, we also find statistically significant and large impacts of removal on the first investigation. While the point estimates are larger for young girls, we fail to reject the hypothesis that these estimated effects on total days are equal for young girls and young boys.

[53]By definition, the number of days spent in each type of foster care after the first investigation is zero for children who are not removed from home.

[54]These results for adoption reflect the fact that this is a relatively rare outcome in our sample. Fewer than 3 percent of children in our DCYF sample of first investigations are later adopted.

[55]These are placements that are not with a relative, licensed foster family or group home.

[56]The confidence intervals for mobility and school characteristic results are sufficiently large that we

As a last test of mediating factors, we examine post-investigation outcomes for adult household members. Specifically, we study parent perpetrators of child abuse or neglect and estimate the impact of child removal on their criminal charges and incarceration within the two- or four-year period after an investigation concludes.[57,58]

Appendix Table A13 reports impacts separately for the parent perpetrators of young girls (Panel A) and young boys (Panel B).[59] We find no statistically significant impacts of removal on the likelihood that a parent perpetrator is charged or incarcerated in the time after an investigation concludes. The small and imprecise results for young girls provide no strong evidence for the hypothesis that child removal leads to changes in household environment that mediate the impacts of removal that we detect.

6.3 Analysis of Siblings

A final explanation that we consider is that girls and boys respond differently to the same treatment of home removal during early life. This hypothesis is motivated by prior research that finds that biology and social processes drive development advantages for young girls in terms of language, temperament, and socioemotional development (Else-Quest et al., 2006; Zahn-Waxler et al., 2008; Schore, 2017; Magnuson et al., 2016). To test for gender differences in the impact of removal per se, we compare impacts of removal between brothers and sisters from the same household.

Table 8 reports estimated impacts of removal, where the sample is limited to young

cannot rule out large positive or large negative impacts. The exception is for school-level value added where we can rule out positive or negative impacts that are larger than one tenth of a standard deviation in test scores.

[57]Note that all perpetrators in the sample are associated with an investigation where DCYF has substantiated the report of abuse or neglect. The data contain no information on the residence of a perpetrator.

[58]There are at least three reasons why charges and incarceration of parent perpetrators might increase following removal. First, during the hearing and removal decision process, evidence may be uncovered which would trigger an adult criminal charge that results in post-investigation incarceration. Second, the DCYF system could affect reporting behavior because parents must regularly check-in with case management staff (who are not CPIs) as part of a child reunification plan. Third, removal may adversely affect the mental health of perpetrators resulting in changes in criminal behavior.

[59]The unit of analysis is parent who is listed as a perpetrator. We split the analysis by gender of the investigated child. If a parent is associated with siblings of the opposite sex, they are included in the results for both young girls (Panel A) and young boys (Panel B).

(investigated before age 6) siblings, and the specification is a modified version of Equation 1 that interacts removal with indicators for gender. In this approach, the IV model has two endogenous variables, which are interaction terms for removal and an indicator for being a girl and for removal and an indicator for being a boy.[60] While the results are imprecisely estimated, Columns 1 and 2 and show the point estimates for young girls who have siblings are nearly identical to the effects in Table 3 for the main sample. In contrast to these large and positive estimates, the effects for boy siblings are negative. We obtain a similar pattern of results in Columns 3 and 4 when we restrict the sample to the oldest sibling of the opposite gender per family. Due to the large standard errors in our estimates, we cannot statistically reject the hypothesis of equal effects for young siblings. Overall, we interpret these results as suggestive evidence that young girls are more positively and significantly impacted by home removal than their male counterparts.

7 Impacts of Removal on Outcomes for Older Children

Finally, we study effects of removal on the older children who are investigated at ages 6-18. We study post-investigation schooling outcomes and the following (post-investigation) later-life outcomes: having any juvenile court conviction by age 18, graduation from high school by age 19, teen birth, and enrollment in any post secondary institution by age 22.[61] In contrast to the analysis in Section 5, we study these later-life outcomes only for older children since a child investigated before age six will generally not be old enough to be at risk for a given later-life outcome by the end of the period covered by the data sources.

Appendix Table A14 reports tests of randomization for the sample of older children. These results provide an important caveat for the analysis of impacts of removal for older

[60]The first stage has two instruments, which are the leave-out measures interacted with gender. The second and first stages both control for the main effects for gender.

[61]Details on the sample construction and outcomes are provided in Appendix B. Note that we construct schooling outcomes of older children (i.e., the measures of grade retention, special education participation (IEP), and average absences) using only school year observations that occur after the year that an investigation concludes. Most test score results for older children are based on a sample of children investigated at ages 6-12 because children investigated at later ages will not be enrolled in the testing grades (3-8).

children. Column 2 shows that we reject the null hypothesis at the one percent level in a joint test of the statistical significance of case characteristics in the sample of older investigated girls.[62] To help assess whether this imbalance threatens the validity of IV estimates for older children, we conduct two tests, which we discuss in detail in Appendix E. First, we find that estimates of the impact of removal are not sensitive to the inclusion of case characteristic controls. This provides some reassurance to the extent that observed case characteristics are correlated with unobserved explanatory variables (Altonji et al., 2005). Second, unlike the analysis for young children, we can analyze test scores in the periods *before* an investigation begins for older children. This placebo analysis finds that there are no statistically significant impacts on pre-investigation test scores.

Appendix Table A15 reports estimates for the impact of removal for older girls (Panel A) and older boys (Panel B). Across outcomes, we find no statistically significant impacts of removal for either gender. The estimates are relatively imprecise, and we cannot rule out substantively large positive or negative impacts. For older girls, the point estimates do not consistently point to beneficial impacts. For example, the results suggest removal increases the likelihood of having a teenage birth and improves enrollment in a post secondary institution. The results for older boys provide some weak, but suggestive evidence that removal has detrimental effects in terms of reduced test scores, increases in adverse school experience, and lower likelihoods of both high-school graduation or post secondary attendance.

Comparing the estimates for older and younger girls allows us to examine whether the effects of removal are specific to age. For girls investigated at older ages, the estimated impact on average test scores is -0.230 standard deviations. Despite the large standard error associated with this estimate, we can reject that the hypothesis that the effects for older and younger girls are equal at the five percent significance level. This pattern is consistent with the literature on the importance of early-life interventions (Cunha et al., 2006; Heckman, 2006; Cunha and Heckman, 2007; Almond and Currie, 2011; Heckman et al., 2013; Bharadwaj

[62]The regression estimates show that older girls who have physical neglect or emergency cases see CPIs who have 1.5 and 1.6 percentage points higher removal tendencies.

et al., 2013; Heckman and Mosso, 2014; Elango et al., 2015; Almond et al., 2017; Chyn et al., 2019).

As a final discussion point, we benchmark our results relative to prior studies of home removal. Using a similar IV approach, Doyle (2007) studied older children investigated at ages 5 to 15 in Illinois. He found statistically significant and large positive impacts on teenage pregnancy (29 percentage points) and juvenile delinquency (47 percentage points) for older girls.[63] In our sample, the positive point estimate for teenage pregnancy for older girls is much smaller in magnitude, but the standard error is sufficiently large that we cannot rule out the effect size observed by Doyle (2007).

8 Conclusion

Child protection authorities in the U.S. remove more than 200,000 children from their homes annually (U.S. Department of Health and Human Services, 2016). Despite this fact, there is relatively little research on the causal impacts of this policy. This paper provides new evidence on the effects of home removal by using comprehensive administrative data on educational outcomes. We focus on children removed before the age of six and examine heterogeneous effects by gender. Our analysis is motivated by the growing literature showing the importance of early-life interventions (Cunha et al., 2006; Heckman, 2006; Cunha and Heckman, 2007; Almond and Currie, 2011; Heckman et al., 2013; Heckman and Mosso, 2014; Elango et al., 2015; Almond et al., 2017) and differential responses by gender (Heckman et al., 2010; Bertrand and Pan, 2013; Heckman et al., 2013; Elango et al., 2015; Conti et al., 2016; Heckman et al., 2017; Garcia et al., 2018; Autor et al., 2019).

We use the removal tendency of quasi-experimentally assigned child protective investigators as an instrument for removal. We find that removal causes statistically significant and substantial improvements in performance on standardized exams for young girls, as well as

[63] Warburton et al. (2014) also study crime for older investigated children and use an IV strategy based on caseworkers. They find imprecise IV estimates of the impact of foster care placement. Lindquist and Santavirta (2014) conduct a descriptive study and find that, among children placed at ages 13-18, foster care is associated with higher crime.

decreases in grade retention and special education needs. Estimates show similar impacts on test scores starting from the first testing grade and onward, which suggests a permanent change in ability prior to when a young removed girl begins taking exams. We do not find significant positive impacts of removal for young boys. In contrast, we find insignificant results with point estimates implying, if anything, negative impacts on test scores. We show that our results are robust to several checks, including allowing for heterogeneity in investigator removal tendency by case and child characteristics.

We investigate several potential explanations for the gendered pattern of treatment effects. An analysis of siblings suggests that young girls benefit from removal while their brothers do not. We find no evidence of notable differences in the complier characteristics of girls and boys, and we find that young children of both genders have similar foster care and school experiences subsequent to removal. Overall, this suggests that the impact of home removal per se varies based on the gender of young children.

Our findings echo prior studies of schooling and social program interventions that find girls respond positively and significantly to interventions aimed at improving educational opportunity or community environment (Hastings et al., 2006; Kling et al., 2007; Angrist and Lavy, 2009; Heckman et al., 2013; Deming et al., 2014; Hoynes et al., 2016). In addition, our finding that increases in academic performance accrue to girls removed before age six contributes to the literature on the importance of early-life conditions (Cunha et al., 2006; Heckman, 2006; Cunha and Heckman, 2007; Almond and Currie, 2011; Heckman et al., 2013; Heckman and Mosso, 2014; Elango et al., 2015; Almond et al., 2017).

Given the prevalence of home removal, we conclude by emphasizing the need for additional research on the impacts of home removal. To the best of our knowledge, we provide the first estimates of the causal impacts of home removal at early ages. Prior work by Doyle (2007; 2008) provides compelling evidence on the causal effects for children removed at older ages. Future research that uses administrative data from other states can help facilitate a more complete understanding of the effects of removal on neglected and abused children.

References

Abadie, Alberto. (2003). 'Semiparametric instrumental variable estimation of treatment response models', *Journal of Econometrics* 113(2), 231–263.

Aizer, Anna and Doyle, Joseph J. (2015). 'Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges*', *The Quarterly Journal of Economics* .

Aizer, Anna, Eli, Shari, Ferrie, Joseph and Lleras-Muney, Adriana. (2016). 'The Long-Run Impact of Cash Transfers to Poor Families', *American Economic Review* 106(4), 935–971.

Almond, Douglas and Currie, Janet. (2011), Chapter 15 - Human capital development before age five, *in* **David Card and Orley Ashenfelter and David Card**., eds, 'Handbook of Labor Economics', Vol. 4, Part B, Elsevier, pp. 1315–1486.

Almond, Douglas, Currie, Janet and Duque, Valentina. (2017), Childhood Circumstances and Adult Outcomes: Act II, NBER Working Paper 23017, National Bureau of Economic Research.

Almond, Douglas, Doyle, Joseph J., Kowalski, Amanda E. and Williams, Heidi. (2010). 'Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns', *The Quarterly Journal of Economics* 125(2), 591–634.

Altonji, Joseph G., Elder, Todd E. and Taber, Christopher R. (2005). 'Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools', *Journal of Political Economy* 113(1), 151–184.

Anderson, Michael L. (2008). 'Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects', *Journal of the American Statistical Association* 103(484), 1481–1495.

Angrist, Joshua D., Imbens, Guido W. and Rubin, Donald B. (1996). 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association* 91(434), 444–455.

Angrist, Joshua and Lavy, Victor. (2009). 'The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial', *American Economic Review* 99(4), 1384–1414.

Autor, David, Figlio, David, Karbownik, Krzysztof, Roth, Jeffrey and Wasserman, Melanie. (2019). 'Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes', *American Economic Journal: Applied Economics* 11(3), 338–381.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). 'Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain', *Econometrica* 80(6), 2369–2429.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (2014). 'High-Dimensional Methods and Inference on Structural and Treatment Effects', *The Journal of Economic Perspectives* 28(2), 29–50.

Benjamini, Yoav, Krieger, Abba M. and Yekutieli, Daniel. (2006). 'Adaptive Linear Step-up Procedures that Control the False Discovery Rate', *Biometrika* 93(3), 491–507.

Benson, Cassandra and Fitzpatrick, Maria D. (2018), Beyond Reading, Writing, and Arithmetic: The Role of Teachers and Schools in Reporting Child Maltreatment.

Berger, Lawrence M., Bruch, Sarah K., Johnson, Elizabeth I., James, Sigrid and Rubin, David. (2009). 'Estimating the "Impact" of Out-of-Home Placement on Child Well-Being: Approaching the Problem of Selection Bias', *Child Development* 80(6), 1856–1876.

Berger, Lawrence M., Cancian, Maria, Han, Eunhee, Noyes, Jennifer and Rios-Salas, Vanessa. (2014). 'Children's Academic Achievement and Foster Care', *Pediatrics* pp. 2014–2448.

Bertrand, Marianne and Pan, Jessica. (2013). 'The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior', *American Economic Journal: Applied Economics* 5(1), 32–64.

Bharadwaj, Prashant, Loken, Katrine and Neilson, Christopher. (2013). 'Early Life Health Interventions and Academic Achievement', *American Economic Review* 103(5), 1862–1891.

Bhuller, Manudeep, Dahl, Gordon B., Løken, Katrine V. and Mogstad, Magne. (2016), Incarceration, Recidivism and Employment, NBER Working Paper 22648, National Bureau of Economic Research.

Bhuller, Manudeep, Dahl, Gordon B., Loken, Katrine V. and Mogstad, Magne. (2018). 'Intergenerational Effects of Incarceration', *AEA Papers and Proceedings* 108, 234–240.

Bound, John, Jaeger, David A. and Baker, Regina M. (1995). 'Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak', *Journal of the American Statistical Association* 90(430), 443–450.

Brinch, Christian N., Mogstad, Magne and Wiswall, Matthew. (2017). 'Beyond LATE with a Discrete Instrument', *Journal of Political Economy* 125(4), 985–1039.

Campbell, Frances, Conti, Gabriella, Heckman, James J., Moon, Seong Hyeok, Pinto, Rodrigo, Pungello, Elizabeth and Pan, Yi. (2014). 'Early Childhood Investments Substantially Boost Adult Health', *Science* 343(6178), 1478–1485.

Carneiro, Pedro, Heckman, James J and Vytlacil, Edward J. (2011). 'Estimating Marginal Returns to Education', *American Economic Review* 101(6), 2754–2781.

Chetty, Raj, Friedman, John N. and Rockoff, Jonah E. (2014). 'Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates', *American Economic Review* 104(9), 2593–2632.

Chetty, Raj, Hendren, Nathaniel and Katz, Lawrence F. (2016). 'The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment', *American Economic Review* 106(4), 855–902.

Chyn, Eric. (2018). 'Moved to Opportunity: The Long-Run Effects of Public Housing Demolition on Children', *American Economic Review* 108(10), 3028–3056.

Chyn, Eric, Gold, Samantha and Hastings, Justine S. (2019), The Returns to Early-life Interventions for Very Low Birth Weight Children, NBER Working Paper 25753, National Bureau of Economic Research.

Conti, Gabriella, Heckman, James J. and Pinto, Rodrigo. (2016). 'The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour', *The Economic Journal* 126(596), F28–F65.

Cornelissen, Thomas, Dustmann, Christian, Raute, Anna and Schönberg, Uta. (2016). 'From LATE to MTE: Alternative methods for the evaluation of policy interventions', *Labour Economics* 41, 47–60.

Cunha, Flavio and Heckman, James. (2007). 'The Technology of Skill Formation', *American Economic Review* 97(2), 31–47.

Cunha, Flavio, Heckman, James J., Lochner, Lance and Masterov, Dimitriy V. (2006), Interpreting the Evidence on Life Cycle Skill Formation, *in* **E. Hanushek and F. Welch**., eds, 'Handbook of the Economics of Education', Vol. 1, Elsevier, pp. 697–812.

Currie, Janet, Mueller-Smith, Michael and Rossin-Slater, Maya. (2018), Violence while in Utero: The Impact of Assaults During Pregnancy on Birth Outcomes, NBER Working Paper 24802, National Bureau of Economic Research.

Currie, Janet and Tekin, Erdal. (2012). 'Understanding the Cycle: Childhood Maltreatment and Future Crime', *The Journal of Human Resources* 47(2), 509–549.

Dahl, Gordon B., Kostøl, Andreas Ravndal and Mogstad, Magne. (2014). 'Family Welfare Cultures', *The Quarterly Journal of Economics* 129(4), 1711–1752.

Deming, David J., Hastings, Justine S., Kane, Thomas J. and Staiger, Douglas O. (2014). 'School Choice, School Quality, and Postsecondary Attainment', *American Economic Review* 104(3), 991–1013.

Dobbie, Will, Goldin, Jacob and Yang, Crystal S. (2018). 'The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges', *American Economic Review* 108(2), 201–240.

Dobbie, Will, Grönqvist, Hans, Niknami, Susan, Palme, Mårten and Priks, Mikael. (2018), The Intergenerational Effects of Parental Incarceration, Technical Report w24186, National Bureau of Economic Research.

Doyle, Joseph J. (2007). 'Child protection and child outcomes: Measuring the effects of foster care', *The American Economic Review* 97(5), 1583–1610.

Doyle, Joseph J. (2008). 'Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care', *Journal of Political Economy* 116(4), 746–770.

Doyle, Joseph J. (2013). 'Causal effects of foster care: An instrumental-variables approach', *Children and Youth Services Review* 35(7), 1143–1151.

Doyle, Joseph J and Aizer, Anna. (2018). 'Economics of Child Protection: Maltreatment, Foster Care, and Intimate Partner Violence', *Annual Review of Economics* 10, 87–108.

Elango, Sneha, García, Jorge Luis, Heckman, James J and Hojman, Andrés. (2015), Early Childhood Education, Working Paper 21766, National Bureau of Economic Research.

Else-Quest, Nicole M., Hyde, Janet Shibley, Goldsmith, H. Hill and Hulle, Carol Ann Van. (2006). 'Gender differences in temperament: a meta-analysis.', *Psychological bulletin* 132(1), 33–72.

Eren, Ozkan and Mocan, Naci. (2017), Juvenile Punishment, High School Graduation and Adult Crime: Evidence from Idiosyncratic Judge Harshness, NBER Working Paper 23573, National Bureau of Economic Research, Cambridge, MA.

Fantuzzo, John W. and Mohr, Wanda K. (1999). 'Prevalence and Effects of Child Exposure to Domestic Violence', *The Future of Children* 9(3), 21–32.

Garces, Eliana, Thomas, Duncan and Currie, Janet. (2002). 'Longer-Term Effects of Head Start', *The American Economic Review* 92(4), 999–1012.

Garcia, Jorge Luis, Heckman, James J. and Ziff, Anna L. (2018). 'Gender differences in the benefits of an influential early childhood program', *European Economic Review* 109, 9–22.

Hastings, Justine S. (2019), Fact-Based Policy: How Do State and Local Governments Accomplish It?, Technical report, The Hamilton Project (Brookings Institution).

Hastings, Justine S., Howison, Mark, Lawless, Ted, Ucles, John and White, Preston. (2019), Unlocking Data to Improve Public Policy, preprint, Open Science Framework.

Hastings, Justine S, Kane, Thomas J and Staiger, Douglas O. (2006). 'Gender and Performance: Evidence from School Assignment by Randomized Lottery', *American Economic Review* 96(2), 232–236.

Heckman, James J, Holland, Margaret L, Makino, Kevin K, Pinto, Rodrigo and Rosales-Rueda, Maria. (2017), An Analysis of the Memphis Nurse-Family Partnership Program, NBER Working Paper 23610, National Bureau of Economic Research.

Heckman, James J. and Mosso, Stefano. (2014). 'The Economics of Human Development and Social Mobility', *Annual Review of Economics* 6(1), 689–733.

Heckman, James J. and Vytlacil, Edward. (2005). 'Structural Equations, Treatment Effects, and Econometric Policy Evaluation1', *Econometrica* 73(3), 669–738.

Heckman, James J. and Vytlacil, Edward J. (1999). 'Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects', *Proceedings of the national Academy of Sciences* 96(8), 4730–4734.

Heckman, James J. and Vytlacil, Edward J. (2007), Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation, *in* **James J. Heckman and Edward E. Leamer**., ed., 'Handbook of Econometrics', Vol. 6, Part B, Elsevier, pp. 4779–4874.

Heckman, James, Moon, Seong Hyeok, Pinto, Rodrigo, Savelyev, Peter and Yavitz, Adam. (2010). 'Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program', *Quantitative Economics* 1(1), 1–46.

Heckman, James, Pinto, Rodrigo and Savelyev, Peter. (2013). 'Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes', *American Economic Review* 103(6), 2052–2086.

Heckman, James, Tobias, Justin L. and Vytlacil, Edward. (2001). 'Four Parameters of Interest in the Evaluation of Social Programs', *Southern Economic Journal* 68(2), 211–223.

Heckman, J. J. (2006). 'Skill Formation and the Economics of Investing in Disadvantaged Children', *Science* 312(5782), 1898–1900.

Holt, Stephanie, Buckley, Helen and Whelan, Sadhbh. (2008). 'The impact of exposure to domestic violence on children and young people: A review of the literature', *Child Abuse & Neglect* 32(8), 797–810.

Hoynes, Hilary, Schanzenbach, Diane Whitmore and Almond, Douglas. (2016). 'Long-Run Impacts of Childhood Access to the Safety Net', *American Economic Review* 106(4), 903–934.

Imbens, Guido W. and Angrist, Joshua D. (1994). 'Identification and Estimation of Local Average Treatment Effects', *Econometrica* 62(2), 467.

Isen, Adam, Rossin-Slater, Maya and Walker, W. Reed. (2017). 'Every Breath You Take—Every Dollar You'll Make: The Long-Term Consequences of the Clean Air Act of 1970', *Journal of Political Economy* 125(3), 848–902.

Kane, Thomas J., Rockoff, Jonah E. and Staiger, Douglas O. (2008). 'What does certification tell us about teacher effectiveness? Evidence from New York City', *Economics of Education Review* 27(6), 615–631.

Kling, Jeffrey R. (2006). 'Incarceration Length, Employment, and Earnings', *The American Economic Review* 96(3), 863–876.

Kling, Jeffrey R., Liebman, Jeffrey B. and Katz, Lawrence F. (2007). 'Experimental Analysis of Neighborhood Effects', *Econometrica* 75(1), 83–119.

Lindquist, Matthew J. and Santavirta, Torsten. (2014). 'Does placing children in foster care increase their adult criminality?', *Labour Economics* 31, 72–83.

Ludwig, Jens and Miller, Douglas L. (2007). 'Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design', *Quarterly Journal of Economics* .

Magnuson, Katherine A., Kelchen, Robert, Duncan, Greg J., Schindler, Holly S., Shager, Hilary and Yoshikawa, Hirokazu. (2016). 'Do the effects of early childhood education programs differ by gender? A meta-analysis', *Early Childhood Research Quarterly* 36, 521–536.

Mueller-Smith, Mike. (2015), The Criminal and Labor Market Impacts of Incarceration.

Rhode Island. (2016), 'Rhode Island General Law. The Rhode Island Board of Education Act, Stat. 16-19-1'.

Rhode Island Department of Children, Youth, and Families. (2018), 'Policy Manual'.

Sampat, Bhaven and Williams, Heidi L. (2015), How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome, NBER Working Paper 21666, National Bureau of Economic Research.

Schore, Allan N. (2017). 'All Our Sons: The Developmental Neurobiology and Neuroendocrinology of Boys at Risk', *Infant Mental Health Journal* 38(1), 15–52.

U.S. Census. (2018), 'S0901: Children Characteristics, 2012-16 American Community Survey 5-year Estimates.'.

U.S. Department of Health and Human Services. (2016), 'Child Maltreatment'.

Warburton, William P., Warburton, Rebecca N., Sweetman, Arthur and Hertzman, Clyde. (2014). 'The Impact of Placing Adolescent Males into Foster Care on Education, Income Assistance, and Convictions: Impact of Placing Adolescent Males into Foster Care', *Canadian Journal of Economics/Revue canadienne d'économique* 47(1), 35–69.

Wildeman, Christopher and Waldfogel, Jane. (2014). 'Somebody's Children or Nobody's Children? How the Sociological Perspective Could Enliven Research on Foster Care', *Annual Review of Sociology* 40(1), 599–618.

Wolfe, David A., Crooks, Claire V., Lee, Vivien, McIntyre-Smith, Alexandra and Jaffe, Peter G. (2003). 'The Effects of Children's Exposure to Domestic Violence: A Meta-Analysis and Critique', *Clinical Child and Family Psychology Review* 6(3), 171–187.

Zahn-Waxler, Carolyn, Shirtcliff, Elizabeth A. and Marceau, Kristine. (2008). 'Disorders of Childhood and Adolescence: Gender and Psychopathology', *Annual Review of Clinical Psychology* 4(1), 275–303.

9 Tables and Figures

Table 1: Descriptive Statistics for in the DCYF Investigations Sample

		(1)	(2)	(3)	(4)
		Sample: Young Children (Age $<$ 6)			
		All	Non-removed	Removed	p-value
Demographics	Female	0.466	0.463	0.479	0.125
		(0.499)	(0.499)	(0.500)	
	White	0.588	0.590	0.581	0.364
		(0.492)	(0.492)	(0.494)	
	Black	0.168	0.160	0.198	0.000
		(0.374)	(0.367)	(0.398)	
	Hispanic	0.162	0.170	0.131	0.000
		(0.369)	(0.376)	(0.337)	
	Other race	0.082	0.079	0.091	0.050
		(0.274)	(0.270)	(0.288)	
	Age	1.81	1.984	1.125	0.000
		(1.762)	(1.760)	(1.593)	
Family	Married couple	0.123	0.137	0.069	0.000
		(0.328)	(0.343)	(0.253)	
	Unmarried couple	0.293	0.304	0.251	0.000
		(0.455)	(0.460)	(0.434)	
	Single/other	0.584	0.560	0.680	0.000
		(0.493)	(0.496)	(0.466)	
	English language	0.972	0.970	0.978	0.010
		(0.166)	(0.171)	(0.146)	
	Other language	0.028	0.030	0.022	0.010
		(0.166)	(0.171)	(0.146)	
Allegation	Neglect	0.795	0.812	0.729	0.000
		(0.404)	(0.391)	(0.444)	
	Physical neglect	0.065	0.059	0.088	0.000
		(0.246)	(0.236)	(0.283)	
	Physical abuse	0.140	0.129	0.183	0.000
		(0.347)	(0.336)	(0.387)	
Reporter	Professional	0.825	0.827	0.816	0.189
		(0.380)	(0.378)	(0.388)	
	Family/friend	0.127	0.125	0.137	0.093
		(0.333)	(0.330)	(0.344)	
	Other reporter	0.048	0.049	0.047	0.753
		0.214	(0.215)	(0.212)	
Invest. Type	Emergency	0.104	0.054	0.298	0.000
		(0.305)	(0.227)	(0.457)	
	Immediate	0.572	0.607	0.433	0.000
		(0.495)	(0.488)	(0.496)	
	Routine	0.325	0.339	0.269	0.000
		(0.468)	(0.473)	(0.444)	
Post Invest.	Removed	0.202	0.000	1.000	0.000
		(0.402)	(0.000)	(0.000)	
	Days, Foster Care	92.520	0.000	457.278	0.000
		(268.119)	(0.000)	(434.224)	
N		13,834	11,035	2,799	

Notes: This table reports descriptive statistics for young children (investigated before age 6) in the DCYF sample. Columns 2 and 3 report statistics for non-removed and removed children, respectively. Column 4 reports the p-value from a t-test of difference in means for Columns 2 and 3.

Table 2: Tests of Random Case Assignment and First-Stage Results

Panel A. First-stage Impact of CPI Removal Tendency			
	(1)	(2)	(3)
Dependent variable:	Removed (=1)		
CPI removal tendency	0.584***	0.539***	0.622***
	(0.056)	(0.078)	(0.075)
Sample	Young Children	Young Girls	Young Boys
Mean of dependent variable	0.202	0.208	0.197
Case controls	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes
N	13,834	6,449	7,385

Panel B. Tests of Randomization			
	(1)	(2)	(3)
Dependent variable:	CPI removal tendency		
Chi-squared statistic	16.500	12.950	16.640
p-value of joint significance	0.284	0.451	0.216
Sample	Young Children	Young Girls	Young Boys
Case controls	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes
N	13,834	6,449	7,385

Notes: This table summarizes tests of random case assignment (Panel A) and the first-stage impact of CPI removal tendency (Panel B). Column 1 reports results for all young children (investigated before age six). Columns 2 and 3 report results for young female and male children, respectively. In Panel A, the joint test statistics are from an OLS regression of CPI removal tendency on the set of case characteristics listed in Table 1. All models include controls for case characteristics and investigation year fixed effects. The chi-square test-statistic and p-value reported are from a test for joint significance of all variables except investigation year fixed effects. Standard errors in parentheses are two-way clustered at the family and CPI level. In Panel B, the first-stage results are from an OLS regression of removal on CPI removal tendency, controls for case characteristics, and investigation year fixed effects (FE). Removed is an indicator for home removal at the child's first investigation. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table 3: Impact of Removal on Test Scores of Young Children

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Young Girls (Age < 6)						
Dependent variable:	Average z-score		Math z-score		Reading z-score	
Removed (= 1)	1.334**	1.366**	1.540***	1.544***	1.141*	1.201**
	(0.586)	(0.575)	(0.591)	(0.578)	(0.625)	(0.612)
Mean of dependent variable	-0.392	-0.392	-0.460	-0.460	-0.328	-0.328
Complier mean if not removed	-1.741	-1.741	-1.922	-1.922	-1.550	-1.550
Case controls	No	Yes	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	15.444	17.152	14.657	16.321	15.952	17.681
N	10,391	10,391	10,418	10,418	10,430	10,430
Individuals	2,721	2,721	2,722	2,722	2,725	2,725
Panel B. Young Boys (Age < 6)						
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score		Math z-score		Reading z-score	
Removed (= 1)	0.051	-0.059	0.013	-0.123	0.109	0.024
	(0.596)	(0.561)	(0.584)	(0.570)	(0.652)	(0.606)
Mean of dependent variable	-0.571	-0.571	-0.518	-0.518	-0.630	-0.630
Complier mean if not removed	-0.931	-0.931	-0.867	-0.867	-1.022	-1.022
Case controls	No	Yes	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	10.492	14.238	10.543	14.154	10.638	14.474
N	12,345	12,345	12,387	12,387	12,406	12,406
Individuals	3,148	3,148	3,149	3,149	3,149	3,149

Notes: This table reports results for the impact of removal on test scores for young girls (Panel A) and young boys (Panel B). We standardize scores at the grade-year level and construct a yearly panel of tests taken in grades 3-8 during school years 2005-2016. All results are from two-stage least squares models with a leave-out measure of CPI removal tendency as an instrument for removal. Columns 1-2 report impacts for the average of standardized math and reading scores. Columns 3-4 and 5-6 report results for reading scores and math scores, respectively. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table 4: Impact of Removal on Additional Schooling Outcomes of Young Children

Panel A. Young Girls (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	-0.228**	-0.441*	-5.629	-0.918**
	(0.108)	(0.248)	(5.218)	(0.400)
Mean of dependent variable	0.043	0.258	11.982	0.000
Complier mean if not removed	0.291	0.726	10.039	0.694
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	21.007	21.007	21.007	21.007
N	2,778	2,778	2,778	2,778
Individuals	2,778	2,778	2,778	2,778

Panel B. Young Boys (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	-0.053	-0.201	-0.550	-0.228
	(0.114)	(0.241)	(4.859)	(0.327)
Mean of dependent variable	0.064	0.428	12.380	0.000
Complier mean if not removed	0.239	0.759	12.955	0.483
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	16.159	16.159	16.159	16.159
N	3,225	3,225	3,225	3,225
Individuals	3,225	3,225	3,225	3,225

Notes: This table reports results for the impact of removal on schooling outcomes for young girls (Panel A) and young boys (Panel B). Columns 1-3 report impacts on measures of whether an investigated child was ever retained, ever participated in special education (i.e., has an IEP), and the average number of days absent during grades 3-8. Column 4 reports results for an index that is constructed from standardized measures of the retention, IEP, and absence measures. All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table 5: Impact of Removal on School Enrollment and Test-taking

Panel A. Young Girls (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Enrolled, Ages 8-13		Tested, Ages 8-13	
Removed (= 1)	0.134	0.072	0.094	0.045
	(0.243)	(0.209)	(0.231)	(0.201)
Mean of dependent variable	0.633	0.633	0.570	0.570
Complier mean if not removed	0.595	0.595	0.537	0.537
Case controls	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	25.734	33.848	25.734	33.848
N	17,774	17,774	17,774	17,774
Individuals	4,101	4,101	4,101	4,101

Panel B. Young Boys (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Enrolled, Ages 8-13		Tested, Ages 8-13	
Removed (= 1)	-0.388	-0.369	-0.336	-0.317
	(0.248)	(0.242)	(0.255)	(0.247)
Mean of dependent variable	0.639	0.639	0.564	0.564
Complier mean if not removed	0.946	0.946	0.740	0.740
Case controls	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	25.647	31.068	25.647	31.068
N	21,293	21,293	21,293	21,293
Individuals	4,750	4,750	4,750	4,750

Notes: This table reports results for the impact of removal on public school enrollment and test-taking outcomes for young girls (Panel A) and young boys (Panel B). These measures are contained in a panel for investigated children that covers ages 8-13 (i.e., the ages associated with grades 3-8). All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table 6: Impact of Removal on Foster Care Outcomes

Panel A. Young Girls (Age < 6)						
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Days in any foster care	Days w/ relative	Days w/ foster family	Days in group home	Days in other care	Adopted (=1)
Removed (= 1)	334.804***	203.456***	124.461	10.094*	-3.207	0.026
	(116.263)	(60.046)	(109.533)	(5.616)	(13.520)	(0.068)
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	46.552	46.552	46.552	46.552	46.552	46.552
N	6,449	6,449	6,449	6,449	6,449	6,449
Individuals	6,449	6,449	6,449	6,449	6,449	6,449
Panel B. Young Boys (Age < 6)						
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Days in any foster care	Days w/ relative	Days w/ foster family	Days in group home	Days in other care	Adopted (=1)
Removed (= 1)	398.959***	151.555***	181.317**	27.485*	38.603***	-0.019
	(97.514)	(50.360)	(71.145)	(14.922)	(14.688)	(0.054)
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	67.242	67.242	67.242	67.242	67.242	67.242
N	7,385	7,385	7,385	7,385	7,385	7,385
Individuals	7,385	7,385	7,385	7,385	7,385	7,385

Notes: This table reports results for the impact of removal on foster care placement outcomes for young girls (Panel A) and young boys (Panel B). All foster care outcomes are associated with the child's first investigation, which implies the means of placement outcomes are zero for non-removed children. Days in foster care is a measure of total time spent in foster care as a result of the child's first investigation. We split days in foster care into four categories: days spent with relatives, days spent with a foster family (non-relatives), days spent in a group home, and other days spent in foster care. Adoption is an indicator for whether the child is adopted upon discharge from foster care. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table 7: Impact of Removal on School Mobility and School-level Characteristics

	Panel A. Young Girls (Age < 6)				
	(1)	(2)	(3)	(4)	(5)
Dependent variable:	Moved Schools (=1)	Value-Added	Avg. Test Scores	% Minority	% FRL
Removed (= 1)	-0.017	0.041	0.095	0.098	-0.033
	(0.114)	(0.034)	(0.179)	(0.189)	(0.158)
Mean of dependent variable	0.351	-0.046	-0.118	0.469	0.578
Complier mean if not removed	0.417	-0.086	-0.284	0.377	0.652
Case controls	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	21.239	21.285	21.285	21.371	21.371
N	11,314	11,380	11,380	11,418	11,418
Individuals	2,836	2,852	2,852	2,855	2,855
	Panel B. Young Boys (Age < 6)				
	(1)	(2)	(3)	(4)	(5)
Dependent variable:	Moved Schools (=1)	Value-Added	Avg. Test Scores	% Minority	% FRL
Removed (= 1)	-0.021	-0.022	-0.285	0.251	0.211
	(0.108)	(0.032)	(0.230)	(0.152)	(0.135)
Mean of dependent variable	0.362	-0.050	-0.158	0.468	0.576
Complier mean if not removed	0.410	-0.045	-0.247	0.447	0.511
Case controls	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	24.941	31.068	25.677	25.866	25.866
N	13,662	13,759	13,759	13,796	13,796
Individuals	3,299	3,321	3,321	3,325	3,325

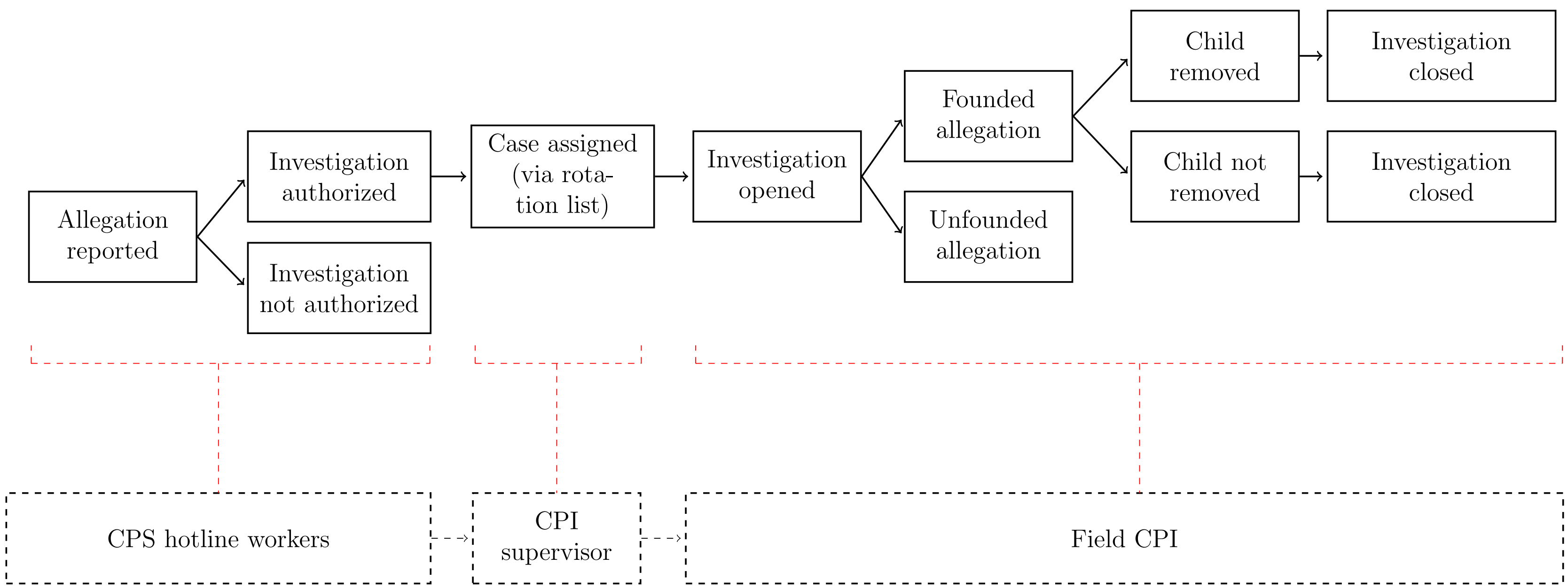
Notes: This table reports results for the impact of removal on school mobility and school-level characteristics for young girls (Panel A) and young boys (Panel B). All measures are based on a panel of observations covering grades 3-8. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table 8: Impact of Removal on Test Scores of Young Siblings

	(1)	(2)	(3)	(4)
Dependent variable:	Average z-score			
Removed (= 1) × Female	1.475	1.336	1.170	0.964
	(1.152)	(0.961)	(1.096)	(0.876)
Removed (= 1) × Male	-0.108	-0.262	-0.241	-0.420
	(0.892)	(0.827)	(0.973)	(0.917)
Sample	All	All	Oldest	Oldest
Mean of dependent variable				
Female	-0.496	-0.496	-0.506	-0.506
Male	-0.671	-0.671	-0.661	-0.661
Complier mean if not removed				
Female	-1.504	-1.504	-1.317	-1.317
Male	-0.230	-0.230	-0.201	-0.201
Case controls	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	3.333	4.168	3.615	4.531
N	5,562	5,562	4,776	4,776
Individuals	1,347	1,347	1,158	1,158

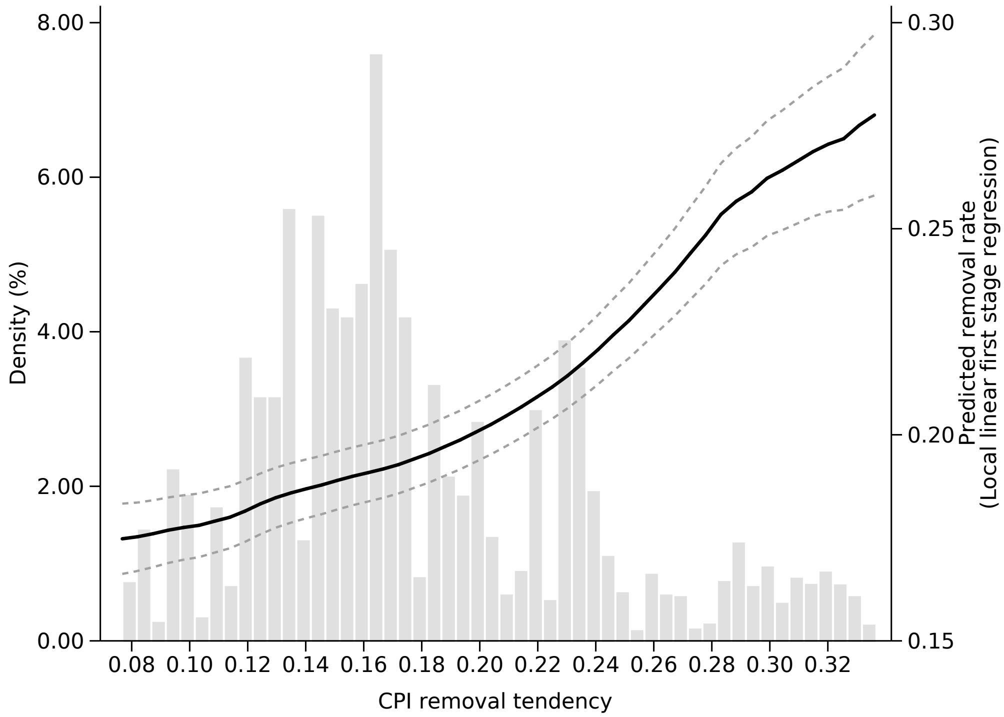
Notes: This table reports results for the impact of removal on test score outcomes for young girls and young boys who are siblings. Results are based on estimating IV models where there are two endogenous variables which are interactions between removal status gender dummy variables. The first-stage has two instruments which are the leave-out measures interacted with the same gender dummy variables. Columns 1 and 2 report impacts using all young siblings. Columns 3 and 4 report estimates using a sample that only includes the oldest (below age 6) opposite sex siblings in the young children sample. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Figure 1: DCYF Process for Abuse and Neglect Allegations



Notes: This figure illustrates the process by which an allegation of abuse or neglect is processed by DCYF in Rhode Island. See Section 2 for further details.

Figure 2: CPI Removal Tendency



Notes: This figure reports the distribution of CPI removal tendency for the sample of young children investigated by DCYF. Section 4 describes how the measure is constructed. The total number of children is 13,384, and the number of unique CPIs is 102.

A Appendix Tables and Figures

Table A1: Tests of Random Case Assignment (Full Regression Results), Young Children Sample

	(1)	(2)	(3)
Dependent variable:		CPI removal tendency	
Female	-0.001		
	(0.001)		
Black	-0.001	-0.002	0.000
	(0.002)	(0.002)	(0.003)
Hispanic	0.002	0.004*	0.001
	(0.001)	(0.002)	(0.002)
Other race	0.003	0.001	0.005*
	(0.002)	(0.003)	(0.003)
Age	-0.000	-0.000	0.000
	(0.000)	(0.001)	(0.000)
Married couple	0.000	-0.002	0.001
	(0.002)	(0.003)	(0.002)
Unmarried couple	-0.001	-0.000	-0.001
	(0.001)	(0.002)	(0.001)
English language	-0.000	-0.005	0.004
	(0.004)	(0.004)	(0.005)
Neglect	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)
Physical neglect	0.003	0.002	0.004
	(0.002)	(0.004)	(0.003)
Professional reporter	-0.003	-0.004	-0.002
	(0.003)	(0.004)	(0.003)
Family/friend reporter	-0.003	-0.007	0.000
	(0.003)	(0.005)	(0.003)
Emergency investigation	0.001	0.000	0.002
	(0.002)	(0.003)	(0.003)
Immediate investigation	0.002	0.001	0.002
	(0.001)	(0.002)	(0.002)
Chi-squared statistic	16.500	12.950	16.640
p-value of joint significance	0.284	0.451	0.216
Sample	Young Children	Young Girls	Young Boys
Mean of CPI removal tendency	0.178	0.177	0.179
Investigation year FE	Yes	Yes	Yes
N	13,834	6,449	7,385

Notes: This table reports regression results testing the random assignment of cases to CPIs. Results are from a regression of CPI removal tendency on the case characteristics listed and investigation year fixed effects. Column 1 reports estimates for all young children. Columns 2 and 3 report estimates for young female and male children, respectively. The chi-square statistic and p-value reported are from an test of joint significance of all variables except investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table A2: Exclusion Restriction Tests

	Panel A. Removed Young Girls (Age < 6)			
	(1)	(2)	(3)	(4)
Dependent variable:	Days in any foster care	Number of placements	Placed with relative (=1)	Police Notified (=1)
CPI removal tendency	-262.755	0.898	-0.250	0.059
	(261.601)	(1.002)	(0.233)	(0.094)
Mean of dependent variable	457.079	2.078	0.363	0.958
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
N	1,341	1,341	1,341	1,341
	Panel B. Removed Young Boys (Age < 6)			
	(1)	(2)	(3)	(4)
Dependent variable:	Days in any foster care	Number of placements	Placed with relative (=1)	Police Notified (=1)
CPI removal tendency	-78.021	1.248	-0.405	-0.024
	(240.844)	(1.113)	(0.264)	(0.090)
Mean of dependent variable	457.461	2.233	0.353	0.966
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
N	1,458	1,458	1,458	1,458

Notes: The sample for this analysis is restricted to removed children. The table reports regression results testing whether placement and other investigation outcomes of removed children are correlated with CPI removal tendency. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A3: First-Stage Impact of CPI Removal Tendency, by Subgroup

	(1)	(2)	(3)
Dependent variable:		Removed (= 1)	
White	0.570*** (0.076) [0.200] N=8,136	0.432*** (0.102) [0.206] N=3,768	0.686*** (0.098) [0.194] N=4,368
Black	0.411** (0.205) [0.238] N=2,321	0.493** (0.241) [0.242] N=1,103	0.340 (0.283) [0.235] N=1,218
Hispanic	0.566*** (0.186) [0.163] N=2,247	0.730*** (0.239) [0.171] N=1,032	0.408* (0.238) [0.156] N=1,215
Married couple	0.648*** (0.162) [0.113] N=1,699	0.556** (0.221) [0.120] N=743	0.715*** (0.204) [0.108] N=956
Unmarried couple	0.578*** (0.124) [0.173] N=4,054	0.561*** (0.189) [0.179] N=1,942	0.586*** (0.147) [0.168] N=2,112
Single/other	0.592*** (0.080) [0.236] N=8,081	0.555*** (0.099) [0.240] N=3,764	0.624*** (0.115) [0.232] N=4,317
Neglect	0.563*** (0.071) [0.186] N=10,997	0.496*** (0.096) [0.190] N=5,225	0.623*** (0.093) [0.181] N=5,772
Physical abuse	0.776*** (0.180) [0.264] N=1,940	0.827*** (0.282) [0.281] N=833	0.757*** (0.231) [0.251] N=1,107
Professional reporter	0.615*** (0.063) [0.200] N=11,407	0.575*** (0.088) [0.205] N=5,280	0.646*** (0.082) [0.196] N=6,127
Family/friend reporter	0.493*** (0.184) [0.218] N=1,759	0.535** (0.246) [0.218] N=849	0.482* (0.254) [0.218] N=910
Immediate	0.787*** (0.082) [0.153] N=7,911	0.744*** (0.107) [0.155] N=3,595	0.828*** (0.106) [0.152] N=4,316
Routine	0.372*** (0.116) [0.168] N=4,490	0.335** (0.148) [0.177] N=2,193	0.419*** (0.143) [0.158] N=2,297
Sample:	Young Children	Young Girls	Young Boys
Case controls	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes

Notes: This table summarizes the first-stage relationship between removal and CPI removal tendency for subgroups. Subgroups are based on the characteristics listed in Table 1. The subgroups for physical neglect, other reporter, and emergency cases are not reported because these have relatively few (< 1500) observations. We also omit reporting results based on language since 97 percent of cases are English language. Standard errors in parentheses are two-way clustered at the family and CPI level. Means for removal for each subgroup are reported in brackets. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A4: First-Stage Impact of CPI Removal Tendency, Reverse Sample Calculation for Subgroups

	(1)	(2)	(3)
Dependent variable:		Removed (=1)	
White	0.365***	0.296***	0.429***
	(0.061)	(0.077)	(0.083)
	[0.200]	[0.206]	[0.195]
	N=8,107	N=3,759	N=4,348
Black	0.392**	0.486**	0.293
	(0.160)	(0.233)	(0.215)
	[0.238]	[0.242]	[0.235]
	N=2,321	N=1,103	N=1,218
Hispanic	0.488***	0.731***	0.261
	(0.177)	(0.241)	(0.233)
	[0.163]	[0.171]	[0.156]
	N=2,247	N=1,032	N=1,215
Married couple	0.545***	0.415**	0.636***
	(0.143)	(0.198)	(0.183)
	[0.113]	[0.120]	[0.108]
	N=1,698	N=743	N=955
Unmarried couple	0.565***	0.540***	0.580***
	(0.114)	(0.169)	(0.136)
	[0.173]	[0.179]	[0.168]
	N=4,051	N=1,941	N=2,110
Single/other	0.546 ***	0.593***	0.510***
	(0.102)	(0.120)	(0.138)
	[0.236]	[0.240]	[0.232]
	N=8,029	N=3,740	N=4,289
Neglect	0.362***	0.340***	0.386***
	(0. 074)	(0.086)	(0.095)
	[0.185]	[0.190]	[0.181]
	N=10,916	N=5,192	N=5,724
Physical abuse	0.741***	0.753***	0.741***
	(0.180)	(0.266)	(0.225)
	[0.264]	[0.281]	[0.251]
	N=1,940	N=833	N=1,107
Professional reporter	0.304***	0.316***	0.289***
	(0.054)	(0.079)	(0.079)
	[0.199]	[0.204]	[0.195]
	N=11,267	N=5,218	N=6,049
Family/friend reporter	0.471**	0.491**	0.494*
	(0.191)	(0.234)	(0.252)
	[0.217]	[0.218]	[0.217]
	N=1,758	N=849	N=909
Immediate	0.464***	0.470***	0.459***
	(0.092)	(0.105)	(0.113)
	[0.153]	[0.155]	[0.152]
	N=7,873	N=3,577	N=4,296
Routine	0.311***	0.272**	0.361***
	(0.093)	(0.121)	(0.119)
	[0.167]	[0.177]	[0.158]
	N=4,485	N=2,191	N=2,294
Sample	Young Children	Young Girls	Young Boys
Case controls	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes

Notes: This table summarizes the first-stage relationship between removal and CPI removal tendency for different subgroups. The instrument is recalculated for each subgroup using all observations outside the subgroup ("reverse" sample definition). Subgroups are based on the characteristics listed in Table 1. The subgroups for physical neglect, other reporter, and emergency cases are not reported because these have relatively few (< 1500) observations. We also omit reporting results based on language since 97 percent of cases are English language. Standard errors in parentheses are two-way clustered at the family and CPI level. Means for removal for each subgroup are reported in brackets. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table A5: Characteristics of Compliers for Young Girls and Young Boys

		Young Girls (Age < 6)		Young Boys (Age < 6)	
		(1)	(2)	(3)	(4)
		$P(X=x)$	$P(X=x\|complier)$	$P(X=x)$	$P(X=x\|complier)$
Demographics	White	0.584	0.426	0.591	0.627
		(0.007)	(0.080)	(0.006)	(0.060)
	Black	0.171	0.201	0.165	0.111
		(0.005)	(0.070)	(0.004)	(0.049)
	Hispanic	0.160	0.231	0.165	0.119
		(0.005)	(0.056)	(0.004)	(0.050)
	Other race	0.085	0.128	0.079	0.142
		(0.004)	(0.048)	(0.003)	(0.036)
Family	Married couple	0.115	0.113	0.129	0.137
		(0.004)	(0.048)	(0.004)	(0.037)
	Unmarried couple	0.301	0.286	0.286	0.268
		(0.006)	(0.071)	(0.005)	(0.055)
	Single/other	0.584	0.629	0.585	0.600
		(0.006)	(0.078)	(0.006)	(0.056)
	English language	0.970	0.992	0.973	0.946
		(0.002)	(0.023)	(0.002)	(0.025)
	Other language	0.030	0.013	0.027	0.052
		(0.002)	(0.022)	(0.002)	(0.020)
Allegation	Neglect	0.810	0.756	0.782	0.793
		(0.005)	(0.068)	(0.005)	(0.040)
	Physical neglect	0.061	0.053	0.069	0.022
		(0.003)	(0.042)	(0.003)	(0.004)
	Physical abuse	0.129	0.196	0.150	0.187
		(0.004)	(0.061)	(0.004)	(0.045)
Reporter	Professional	0.819	0.899	0.830	0.864
		(0.005)	(0.069)	(0.005)	(0.055)
	Family/friend	0.132	0.106	0.123	0.101
		(0.004)	(0.057)	(0.004)	(0.042)
	Other reporter	0.050	0.013	0.047	0.041
		(0.003)	(0.036)	(0.003)	(0.024)
Investigation	Emergency	0.102	0.071	0.105	0.041
		(0.004)	(0.056)	(0.003)	(0.050)
	Immediate	0.557	0.750	0.584	0.763
		(0.006)	(0.084)	(0.006)	(0.068)
	Routine	0.340	0.218	0.311	0.226
		(0.005)	(0.078)	(0.005)	(0.061)

Notes: This table reports characteristics of compliers in the DCYF sample. We define compliers as children whose removal decision would have been different had they been assigned the most strict versus the most lenient investigator. To identify compliers, we follow Abadie (2003), Dahl et al. (2014), and Dobbie, Goldin and Yang (2018). Let $\overline{z}$ denote the maximum value of the instrument (the most strict investigator) and $\underline{z}$ denote the minimum value of the instrument (the most lenient investigator). We can then express the share of compliers in our sample as: $p_c = Pr(Removed = 1|Z_i = \overline{z}) - Pr(Removed = 1|Z_i = \underline{z})$. In practice, we assign the top percentile of our instrument to $\overline{z}$ and the bottom percentile of our instrument to $\underline{z}$. As discussed in Dahl et al. (2014) and Dobbie, Goldin and Yang (2018), the share of compliers can be directly estimated as $p_c = \alpha$, where α is the coefficient on the instrument from the first stage regression (Equation 2). In this table, we estimate the share of compliers in each subgroup (row) and report the likelihood of being a complier in each subgroup. Standard errors in parentheses are obtained using 500 bootstrap replications. See Appendix C for further details.

Table A6: Impact of Removal on Additional Schooling Outcomes of Young Children (No Case Controls)

	(1)	(2)	(3)	(4)
Panel A. Young Girls (Age < 6)				
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	-0.234**	-0.442*	-4.645	-0.892**
	(0.111)	(0.246)	(5.416)	(0.304)
Mean of dependent variable	0.043	0.258	11.982	0.000
Complier mean if not removed	0.291	0.726	10.039	0.694
Case controls	No	No	No	No
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	19.743	19.743	19.743	19.743
N	2,778	2,778	2,778	2,778
Individuals	2,778	2,778	2,778	2,778
Panel B. Young Boys (Age < 6)				
	(1)	(2)	(3)	(4)
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	-0.047	-0.192	-0.976	-0.228
	(0.116)	(0.236)	(5.003)	(0.326)
Mean of dependent variable	0.064	0.428	12.380	0.000
Complier mean if not removed	0.239	0.759	12.955	0.483
Case controls	No	No	No	No
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	12.677	12.677	12.677	12.677
N	3,225	3,225	3,225	3,225
Individuals	3,225	3,225	3,225	3,225

Notes: This table reports results for the impact of removal on schooling outcomes for young girls (Panel A) and young boys (Panel B). This table differs from Table 4 because the specifications omit controls for case characteristics. Columns 1-3 report impacts on measures of whether an investigated child was ever retained, ever participated in special education (i.e., has an IEP), and the average number of days absent during grades 3-8. Column 4 reports results for an index that is constructed from standardized measures of the retention, IEP, and absence measures. All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A7: Impact of Removal on Additional Schooling Outcomes of Young Children (Grades K-8)

Panel A. Young Girls (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	-0.360***	-0.365*	-4.227	-0.803**
	(0.138)	(0.215)	(4.912)	(0.325)
Mean of dependent variable	0.119	0.273	12.789	0.000
Complier mean if not removed	0.442	0.718	11.298	0.609
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	27.367	27.367	27.367	27.367
N	3,655	3,655	3,655	3,655
Individuals	3,655	3,655	3,655	3,655

Panel B. Young Boys (Age < 6)

	(1)	(2)	(3)	(4)
Dependent variable:	Retention (=1)	IEP (=1)	Absences	School index
Removed (= 1)	0.041	0.038	1.516	0.120
	(0.164)	(0.205)	(4.476)	(0.315)
Mean of dependent variable	0.153	0.462	12.871	0.000
Complier mean if not removed	0.238	0.562	10.087	0.042
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	26.889	26.889	26.889	26.889
N	4,261	4,261	4,261	4,261
Individuals	4,261	4,261	4,261	4,261

Notes: This table reports results for the impact of removal on schooling outcomes for young girls (Panel A) and young boys (Panel B). This table differs from Table 4 because the measures are based on data from grades K-8 (rather than grades 3-8). Note that the sample size increases because we include children who have not yet reached grade 3 by the end of the schooling records. Columns 1-3 report impacts on measures of whether an investigated child was ever retained, ever participated in special education (i.e., has an IEP), and the average number of days absent during grades 3-8. Column 4 reports results for an index that is constructed from standardized measures of the retention, IEP, and absence measures. All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A8: Adjusted p-values for Impact of Removal on Outcomes of Young Children

Panel A. Young Girls (Age < 6)			
	(1)	(2)	(3)
Dependent variable (below):	2SLS Estimate	p-value	FRD q-value
Average z-score	1.366**	0.017	0.070
	(0.575)		
Retention (=1)	-0.288**	0.034	0.070
	(0.108)		
IEP (=1)	-0.441*	0.075	0.101
	(0.248)		
Absences	-5.629	0.281	0.281
	(5.218)		
Panel B. Young Boys (Age < 6)			
	(1)	(2)	(3)
Dependent variable (below):	2SLS Estimate	p-value	FRD q-value
Average z-score	-0.059	0.916	0.917
	(0.561)		
Retention (=1)	-0.053	0.642	0.917
	(0.114)		
IEP (=1)	-0.201	0.404	0.917
	(0.241)		
Absences	-0.550	0.910	0.917
	(4.869)		

Notes: This table reports adjusted p-values for the impact of removal on outcomes of young children. Column 1 of Panels A and B reproduce the results for young girls and young boys from Tables 3 and 4. Columns 2 and 3 report per-comparison (pairwise) and false discovery rate (FDR) adjusted p-values ("q-values"). The adjustment takes into account the fact that we tested the four listed outcomes for the gender subgroup. The FDR-adjusted p-values control for the number of false positives when multiple hypotheses are tested. These adjusted p-values are calculated using the two-step procedure in Benjamini et al. (2006).

Table A9: Robustness to Changes in Sample Definition

	Panel A. Young Girls						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	Average z-score						
Removed (= 1)	1.366**	1.339**	1.395*	0.917	1.136*	1.183**	1.017**
	(0.575)	(0.587)	(0.760)	(1.123)	(0.587)	(0.530)	(0.563)
Sample	Main sample	CPI > 100 cases	CPI > 200 cases	CPI > 300 cases	With sex cases	Ages 0-4	Ages 0-6
Mean of dependent variable	-0.392	-0.391	-0.401	-0.392	-0.383	-0.385	-0.388
Complier mean if not removed	-1.741	-1.552	-1.630	-1.741	-1.545	-1.692	-1.526
Case controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	17.152	16.958	10.419	3.229	12.442	15.064	17.771
N	10,391	9,909	8,024	4,696	11,028	8,683	12,216
Individuals	2,721	2,597	2,117	1,379	2,873	2,290	3,189
	Panel B. Young Boys						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	Average z-score						
Removed (= 1)	-0.059	-0.061	-0.526	-1.325	-0.234	-0.877	0.003
	(0.561)	(0.619)	(0.794)	(1.283)	(0.570)	(0.833)	(0.428)
Sample	Main sample	CPI > 100 cases	CPI > 200 cases	CPI > 300 cases	With sex cases	Ages 0-4	Ages 0-6
Mean of dependent variable	-0.571	-0.571	-0.568	-0.583	-0.568	-0.593	-0.573
Complier mean if not removed	-0.931	-0.942	-0.871	-0.907	-0.755	-0.974	-0.886
Case controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	14.238	11.729	10.360	3.125	15.177	7.110	22.618
N	12,345	11,756	9,767	5,694	12,685	8,245	14,592
Individuals	3,148	2,997	2,488	1,588	3,222	2,156	3,721

Notes: This table reports results for the impact of removal on the average of standardized test scores for young girls (Panel A) and young boys (Panel B). For comparison, Column 1 reproduces estimates from our main sample and preferred specification (as reported in Table 3). Columns 2-7 report results using alternative samples. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A10: Robustness to Estimating Impacts using ML-IV Approach

	Test score outcomes			Schooling outcomes		
Panel A. Young Girls (Age < 6)						
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score	Math z-score	Reading z-score	School index	Retention (=1)	IEP (=1)
Removed (= 1)	1.071**	1.213**	0.940*	-0.837**	- 0.229**	-0.365*
	(0.482)	(0.489)	(0.508)	(0.335)	(0.102)	(0.214)
Mean of dependent variable	-0.392	-0.459	0.001	-0.059	0.043	0.258
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	27.042	26.322	27.448	14.560	14.560	14.560
N	10,196	10,221	10,234	2,768	2,768	2,768
Individuals	2,672	2,673	2,676	2,768	2,768	2,768
Panel B. Young Boys (Age < 6)						
	Test score outcomes			Schooling outcomes		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score	Math z-score	Reading z-score	School index	Retention (=1)	IEP (=1)
Removed (= 1)	-0.389	-0.564	-0.19	-0.234	-0.083	-0.060
	(0.559)	(0.593)	(0.563)	(0.291)	(0.303)	(0.239)
Mean of dependent variable	-0.572	-0.520	-0.630	0.001	0.064	0.428
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	33.427	33.502	34.435	18.211	18.211	18.211
N	12,220	12,262	12,281	3,209	3,209	3,209
Individuals	3,108	3,109	3,109	3,209	3,209	3,209

Notes: This table reports results based on an IV approach where the CPI removal rate to vary with case characteristics. See Sections 5.5 and Appendix D for details on the IV calculations. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A11: Robustness to Estimating Impacts using Alternative Instruments

	Panel A. Young Girls			
	(1)	(2)	(3)	(4)
Dependent variable:	Average z-score			
Removed (= 1)	1.366**	1.082**	1.480**	1.040**
	(0.575)	(0.451)	(0.620)	(0.429)
IV Version	All Cases 8-year periods	All Cases All (16) years	First Cases 8-year periods	First Cases All (16) years
Mean of dependent variable	-0.392	-0.392	-0.392	-0.392
Complier mean if not removed	-1.741	-1.741	-1.741	-1.741
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	17.152	15.029	17.809	18.303
N	10,391	10,391	10,374	10,391
Individuals	2,721	2,721	2,718	2,721
	Panel B. Young Boys			
	(1)	(2)	(3)	(4)
Dependent variable:	Average z-score			
Removed (= 1)	-0.059	-0.527	-0.249	-0.856
	(0.561)	(0.624)	(0.505)	(0.601)
IV Version	All Cases 8-year periods	All Cases All (16) years	First Cases 8-year periods	First Cases All (16) years
Mean of dependent variable	-0.571	-0.571	-0.571	-0.571
Complier mean if not removed	-0.931	-0.931	-0.931	-0.931
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	14.238	12.612	16.445	15.272
N	12,345	12,345	12,330	12,345
Individuals	3,148	3,148	3,143	3,148

Notes: This table reports results based on an IV approach where the CPI removal rate is calculated over different investigations and time periods. Column 1 reproduces the estimates from our preferred measure, which calculates removal across all of a CPI's cases (first-time and repeat) during an 8-year window. Column 2 reports estimates using a measure based on all cases (first and subsequent) and the entire sample period (2000-2015). Columns 3 reports estimates using a measure based on first cases and an 8-year window. Column 4 reports estimates using a measure based on first cases and the entire sample period (2000-2015). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table A12: Impact of Removal on Outcomes of Young Children, By Minority Status

	Panel A. Young Girls (Age < 6)			
	(1)	(2)	(3)	(4)
Dependent variable:	Average z-score	School index	Average z-score	School index
Removed (= 1)	1.091**	-0.639*	1.593	-1.323**
	(0.485)	(0.358)	(1.196)	(0.725)
Mean of dependent variable	0.015	-0.022	-0.261	0.020
Complier mean if not removed	-1.424	0.401	-1.698	1.030
Sample	Minority	Minority	Not Minority	Not Minority
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	9.972	11.331	6.751	7.631
N	4,652	1,272	5,739	1,506
Individuals	1,255	1,272	1,466	1,506
	Panel B. Young Boys (Age < 6)			
Dependent variable:	Average z-score	School index	Average z-score	School index
Removed (= 1)	0.305	0.238	-0.57	-0.967*
	(0.595)	(0.440)	(1.211)	(0.580)
Mean of dependent variable	-0.732	0.009	-0.442	-0.007
Complier mean if not removed	-1.206	0.103	-0.505	1.036
Sample	Minority	Minority	Not Minority	Not Minority
Case controls	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes
F-statistic (instrument)	8.600	8.723	4.567	7.449
N	5,468	1,467	6,877	1,758
Individuals	1,430	1,467	1,718	1,758

Notes: This table reports results for the impact of removal on test score outcomes for young girls and young boys by minority status (non-white versus white). Columns 2 and 4 reports results for a school index measure that is constructed from standardized measures of the retention, IEP, and absence measures. All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table A13: Impact of Removal on Criminal Justice Outcomes for Parent Perpetrators

Panel A. Parent Perpetrators of Young Girls (Age < 6)

	(1)	(2)
Dependent variable:	Charged/Incar., 2-year Post	Charged/Incar., 4-year Post
Removed (= 1)	-0.024	0.021
	(0.190)	(0.197)
Mean of dependent variable	0.212	0.266
Complier mean if not removed	0.254	0.270
Case controls	Yes	Yes
Investigation year FE	Yes	Yes
F-statistic (instrument)	29.331	29.331
N	6,252	6,252
Individuals	6,252	6,252

Panel B. Parent Perpetrators of Young Boys (Age < 6)

	(1)	(2)
Dependent variable:	Charged/Incar., 2-year Post	Charged/Incar., 4-year Post
Removed (= 1)	-0.004	-0.113
	(0.102)	(0.105)
Mean of dependent variable	0.202	0.255
Complier mean if not removed	0.251	0.360
Case controls	Yes	Yes
Investigation year FE	Yes	Yes
F-statistic (instrument)	55.868	55.868
N	7,141	7,141
Individuals	7,141	7,141

Notes: This table reports results for the impact of removal on criminal justice outcomes for the parents of young girls (Panel A) and young boys (Panel B). Information on parent perpetrators comes from DCYF records. In the sample of young investigated children, 95 percent of children have at least one perpetrator who is a parent. As described in Section 3, we construct samples of parent perpetrators of young girls and young boys and measure whether parents are charged or incarcerated within 2-year and 4-year windows after the conclusion of an investigation. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table A14: Tests of Random Case Assignment (Full Regression Results), Older Children Sample

	(1)	(2)	(3)
Dependent variable:		CPI removal tendency	
Female	-0.001		
	(0.001)		
Black	-0.001	0.000	0.001
	(0.002)	(0.002)	(0.003)
Hispanic	-0.000	-0.003	0.003
	(0.002)	(0.002)	(0.004)
Other race	0.001	0.001	-0.000
	(0.003)	(0.003)	(0.004)
Age	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
Married couple	0.000	0.001	-0.001
	(0.002)	(0.003)	(0.002)
Unmarried couple	-0.000	0.002	-0.002
	(0.002)	(0.002)	(0.002)
English language	-0.003	-0.005*	-0.001
	(0.003)	(0.003)	(0.004)
Neglect	-0.001	-0.004	0.001
	(0.002)	(0.002)	(0.001)
Physical neglect	-0.007*	0.015***	0.001
	(0.004)	(0.003)	(0.005)
Professional reporter	-0.003	-0.005*	-0.001
	(0.003)	(0.004)	(0.004)
Family/friend reporter	-0.002	-0.004	0.000
	(0.003)	(0.003)	(0.004)
Emergency investigation	0.014**	0.016**	0.013
	(0.006)	(0.007)	(0.008)
Immediate investigation	0.002	0.000	0.004
	(0.001)	(0.002)	(0.001)
Chi-squared statistic	23.640	31.590	12.740
p-value of joint significance	0.051	0.003	0.469
Sample	Older Children	Older Girls	Older Boys
Mean of CPI removal tendency	0.177	0.177	0.177
Investigation year FE	Yes	Yes	Yes
N	13,120	6,643	6,477

Notes: This table reports regression results testing the random assignment of cases to CPIs. Results are from a regression of CPI removal tendency on the case characteristics listed and investigation year fixed effects. Column 1 reports estimates for all older children (investigated at ages 6-18). Columns 2 and 3 report estimates for young female and male children, respectively. The chi-square statistic and p-value reported are from an test of joint significance of all variables except investigation year fixed effects (FE). Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table A15: Impact of Removal on Outcomes of Older Children

Panel A. Older Girls (Age $\geq$ 6)

	School-age outcomes		Later-life outcomes			
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score	School Index	Delinquent (=1)	HS Grad. (=1)	Teen Birth (=1)	College (=1)
Removed (= 1)	-0.230	-0.373	-0.030	-0.010	0.089	0.133
	(0.582)	(0.326)	(0.261)	(0.187)	(0.162)	(0.222)
Mean of dependent variable	0.068	-0.005	0.055	0.351	0.194	0.303
Complier mean if not removed	-0.337	0.138	0.101	0.263	0.210	0.032
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	22.129	35.092	9.296	49.213	33.333	38.718
N	7,517	3,029	1,829	4,136	2,956	3,326
Individuals	2,581	3,029	1,829	4,136	2,956	3,326

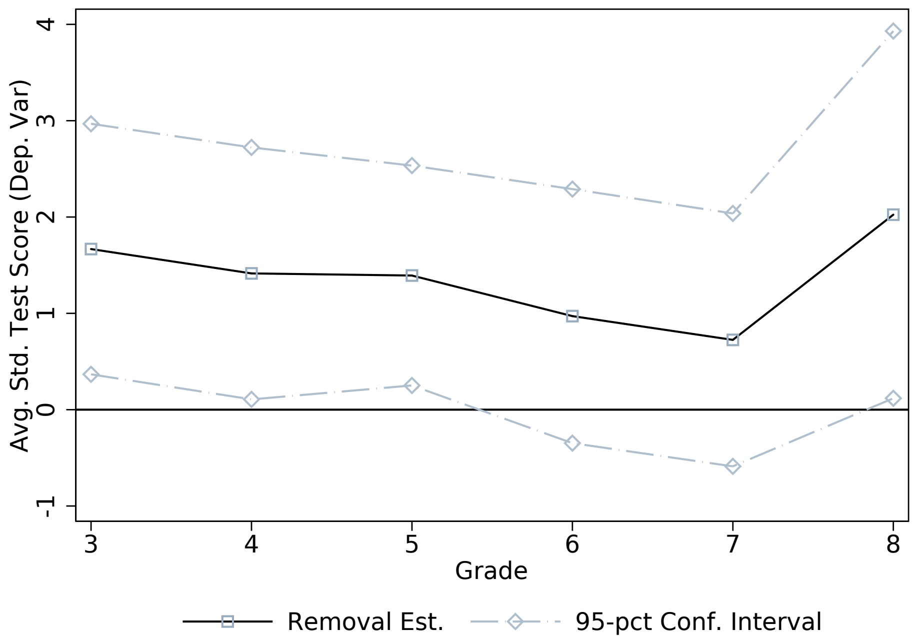
Panel B. Older Boys (Age $\geq$ 6)

	School-age outcomes		Later-life outcomes			
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score	School Index	Delinquent (=1)	HS Grad. (=1)	Teen Birth (=1)	College (=1)
Removed (= 1)	-0.237	0.323	-0.016	-0.144	0.119	-0.127
	(0.429)	(0.216)	(0.156)	(0.157)	(0.115)	(0.187)
Mean of dependent variable	0.053	-0.003	0.147	0.319	0.059	0.239
Complier mean if not removed	-0.414	-0.297	0.096	0.385	0.000	0.367
Case controls	Yes	Yes	Yes	Yes	Yes	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	30.911	34.273	24.610	44.810	26.860	41.145
N	8,838	3,440	2,185	3,770	3,025	2,953
Individuals	2,965	3,440	2,185	3,770	3,025	2,953

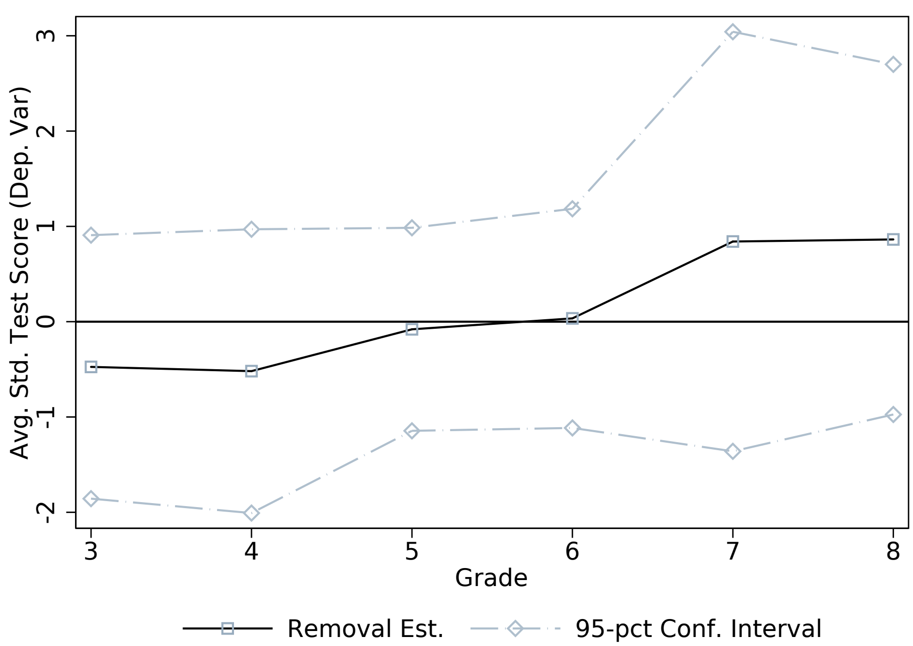
Notes: This table reports results for the impact of removal on outcomes for older girls (Panel A) and older boys (Panel B). Older is defined as being investigated at ages 6 or later (up to age 18). All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. The school index is constructed based on standardized measures of whether an investigated child was ever retained, ever participated in special education (i.e., has an IEP), and the average number of days absent during grades 3-8. All outcomes are measured after the first investigation. Standard errors in parentheses are two-way clustered at the family and CPI level. Significance reported as $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Figure A1: Impact of Removal on Test Scores of Young Children, by Grade

(a) Young Girls



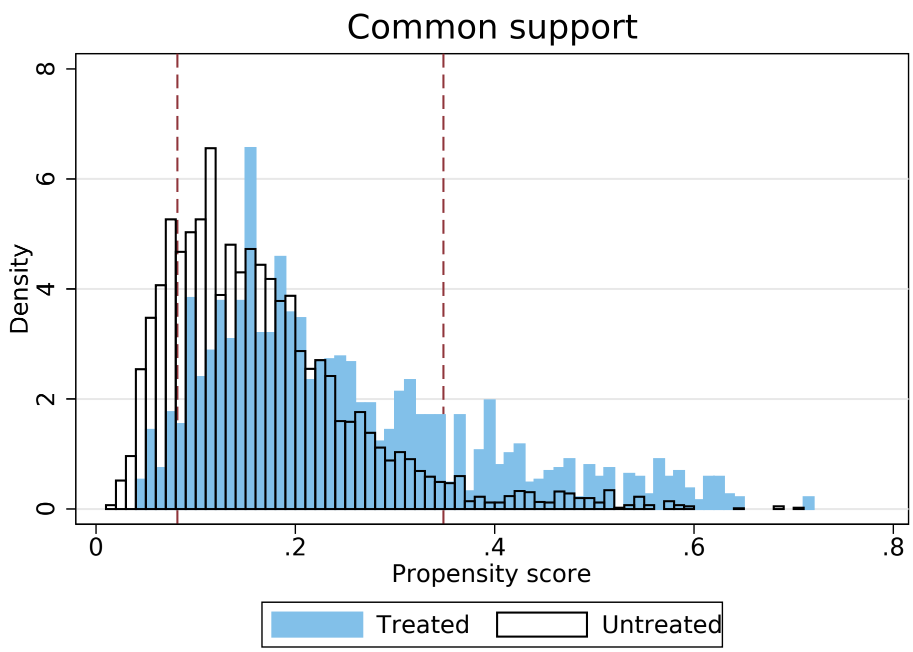
(b) Young Boys



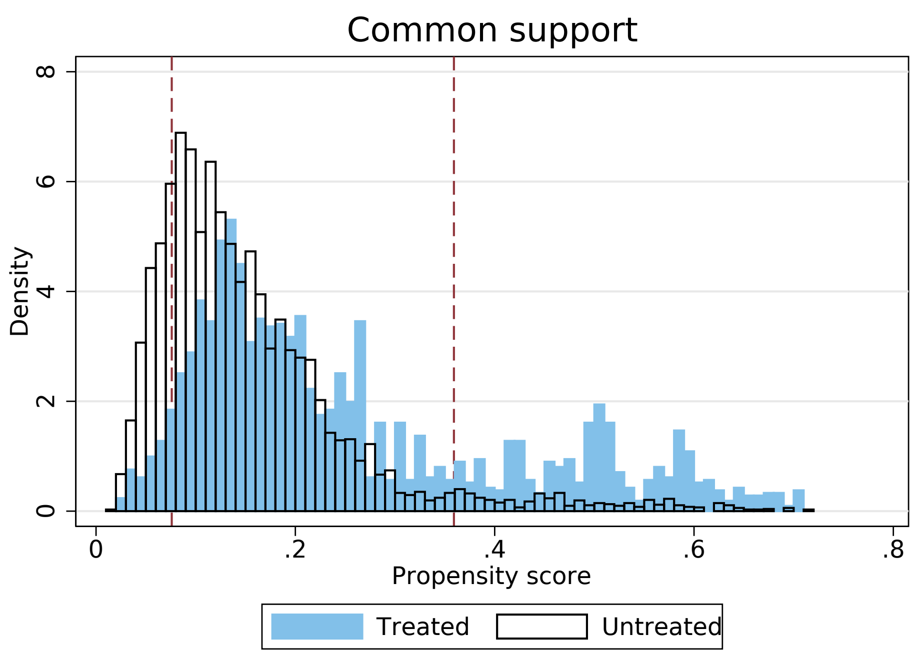
Notes: These figures show results for the impact of removal on test scores estimated in separate regressions for grades 3-8 for young girls (Panel A) and young boys (Panel B). All results are from two-stage least squares models with the leave-out measure of CPI removal tendency as an instrument for removal. All models include controls for the case characteristics in Table 1 and investigation year fixed effects. Confidence intervals are based on standard errors that are two-way clustered at the family and CPI levels.

Figure A2: Common Support of CPI Removal Tendency

(a) Young Girls



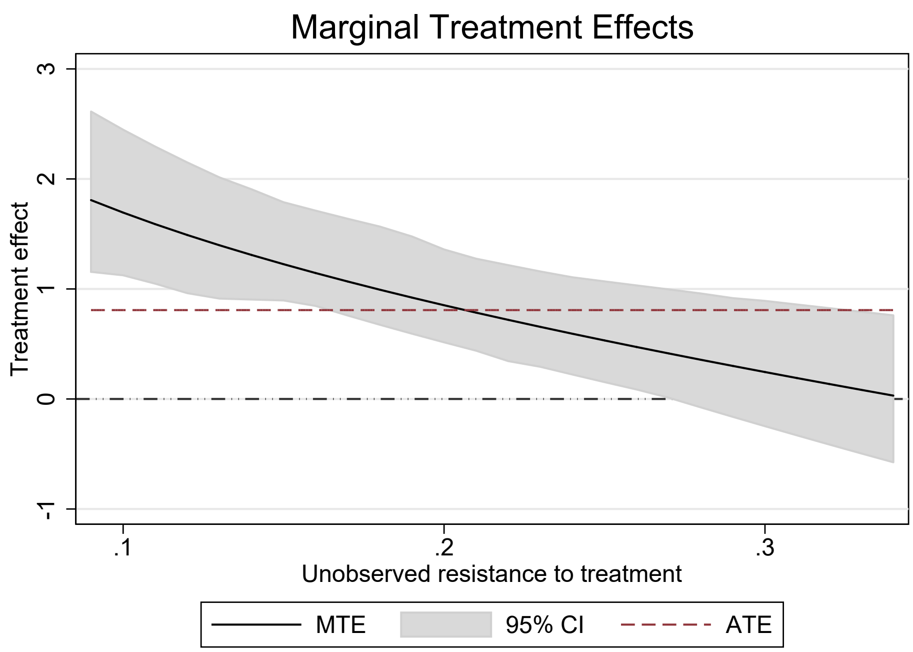
(b) Young Boys



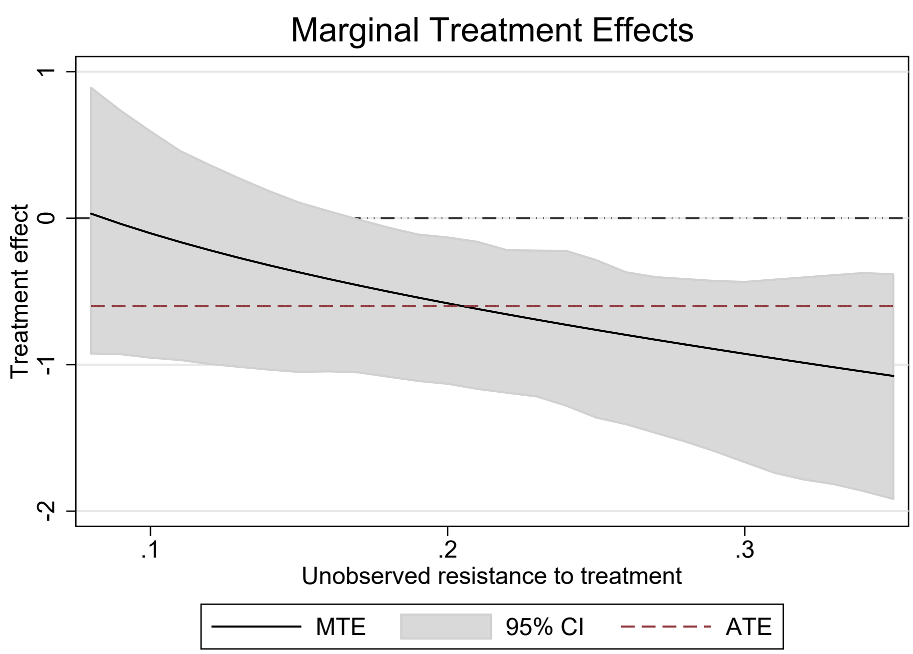
Notes: These figures show the distribution of the propensity score for treated (removed) and non-removed (non-removed) children. The dashed red lines in each figure indicate the upper and the lower points of the propensity score with common support (based on 5 percent trimming).

Figure A3: MTE for Test Scores of Young Children

(a) Young Girls



(b) Young Boys



These figures plot MTEs for the impact of removal on young children based on a local instrumental variables (IV) approach using a global quadratic polynomial specification for the trimmed sample with common support. Standard errors are constructed based on 100 bootstrap replications.

B Data Appendix

This section describes the data sources, data files, and samples that we use for the analysis of the main text.

B.1 Data Sources

Our analysis relies on data from several administrative sources. Table B1 lists each administrative source, files provided, and the time period covered by the associated files.

Table B1: List of Data Sources

Source	Data	Time Period
RI Dept. of Children, Youth, and Families	Child Protective Services (CPS) files – CPS report (allegations) – substantiated investigations – case assignments (field CPIs)	2000-2017
	Adoption and Foster Care Analysis and Reporting System (AFCARS) – foster care placements	2000-2017
	Juvenile delinquency records – sentences to the Rhode Island Training School (RITS) – placement on probation	2000-2016
RI Dept. of Corrections	Criminal justice records – criminal charges – incarceration history	1995-2017
RI Dept. of Education	End-of-Year enrollment records – school, enrollment dates, grade – Individualized Education Program (IEP), free/reduced price lunch status, grade retention, absences – high school graduation	2003-2016
	Standardized testing records – testing school and year – NECAP reading and math test scores (grades 3-8, school years 2005-2013) – PARCC reading and math test scores (grades 3-8, school years 2014-2016)	2005-2016
National Student Clearinghouse	Postsecondary enrollment records – college-going	2004-2015
RI Dept. of Health	Vital records – teen births	2000-2016
RI 360 Database	Demographics – birth date, gender, race	1997-2016

Notes: This table lists data sources, files, and the time period covered by the associated files.

B.2 Description of Files

B.2.1 Child Protective Services and foster care placement files

Child Protective Services (CPS) files (2000-2017) identify victims and perpetrators of child abuse or neglect. These data contain the CPS reports created at the time a suspected abuse or neglect allegation is reported via the Rhode Island (RI) Department of Children, Youth and Families (DCYF) hotline. Note that CPS functions as the investigative arm of DCYF. The CPS files report family structure, primarily language, reporter type, and allegation type for each victim-perpetrator combination, and designated investigation level. The investigation and placement files include all substantiated investigations resulting from CPS reports, as well as the assignment history of investigations to field Child Protection Investigators (CPIs). The Adoption and Foster Care Analysis and Reporting System (AFCARS) data file contains information on all children in foster care in RI.

B.2.2 Juvenile delinquency records

The DCYF houses the Division of Juvenile Corrections, which oversees youth located at the Rhode Island Training School (RITS) or sentenced to probation by the RI Family Court. The RI Family Court handles wayward or delinquent offenses for youth ages 10-17, while youth can remain at RITS through age 18. Records of juvenile delinquency (2000-2016) contain the dates of sentencing for each person.

B.2.3 Criminal justice records

The RI Department of Corrections (DOC) records contain the population of charged and incarcerated individuals in the state of Rhode Island (1995-2017). The dates of each unique charge or sentence are observed, as well as the type of charge (e.g., assault, property crime) and the total sentence length.

B.2.4 End-of-year enrollment records

The RI Department of Education (RIDE) maintains records of all students enrolled in RI public and charter schools; we have access to data from school years 2003-04 through 2016-17. These data include enrollment dates, grade and school attended, Individualized Education Program status (which identifies special education status), free/reduced price lunch status, yearly absences, and high school graduation status.

B.2.5 Standardized testing records

RIDE reports standardized mathematics and reading testing results for enrolled students in grades 3-8. Rhode Island administered the New England Common Assessment Program (NECAP) test from school years 2005-06 to 2013-14 and the Partnership for Assessment of Readiness for College and Careers (PARCC) test from 2014-15 to 2016-17. Participation rates for standardized exams in RI have historically been high (more 95 percent of students take exams). In 2014, participation rates fell to roughly 90 percent, but rose to previous levels by 2016.

B.2.6 Post-secondary enrollment records

The National Student Clearinghouse (NSC) reports post-secondary enrollment dates for RI high school students (2004-2015), regardless of high school completion.

B.2.7 Vital (birth) records

The RI Department of Health (DOH) vital (birth) records contain all Rhode Island births (2000-2016) and include identifiers for the mother and father, as well as mother's date of birth.

B.2.8 Demographics

The RI 360 Database joins records associated with an individual across a range of social programs and government services (see Hastings et al. (2019)). The database provides demographic information (birth date, gender, and race) for all children in the DCYF sample born between 1982 and 2015 and appearing in administrative records between 1997 and 2016.

B.3 Samples and Key Outcomes

B.3.1 Sample of DCYF Investigated Children

We use CPS records to construct a sample of children that CPS investigates. As an initial step, we link alleged abuse or neglect investigation records to a file containing assignment records. This allows us to determine the Child Protective Investigator (CPI) assigned to each investigation, and whether the assignment was via the rotation list (see Section 2). We also link investigations to the AFCARS foster care placement history file to determine whether DCYF placed investigated children into foster care as the result of an investigation.

Using the assembled CPS investigation records, we impose the following restrictions to create a final sample of DCYF investigations.

1. Restrictions related to data cleaning:

(a) Restrict to children ages 0-18 with known demographics. We join children in CPS case files to the RI 360 database to obtain a global identifier and verifiable demographic information (see Hastings et al. (2019)). To be included, children must have birth date and gender.

(b) Restrict to allegations reported via the DCYF hotline. Allegations are primarily reported via the hotline.

(c) Restrict to allegations in which the alleged perpetrator is a family member. In the full CPS case files, 93 percent of reports of neglect or abuse are alleged to have been perpetrated by a member of the child's family. The remaining 7 percent involve DCYF providers of care or institutional abuse allegations, but these investigations follow a different set of procedures.[64]

(d) Drop allegations reported after the initial DCYF hotline call.

(e) Drop allegations that do not meet the criteria for investigation (internally designated as "info/referral" reports). These reports would not be forwarded to the Investigative Unit.

(f) Drop investigations that are unfounded (i.e., there was no preponderance of evidence that child abuse or neglect occurred).

(g) Restrict to investigations from 2000 to 2015. We remove investigations that began after 2015 to avoid censored foster care placement outcomes.

(h) Drop children involved in at most one investigation per day. CPS may receive more than one report of abuse or neglect on the same day for the same child; in such instances, the child could be affiliated with more than one CPI. We exclude these cases.

(i) Restrict to investigations matched to a CPI assignment. The link between investigations and case assignment history is imperfect, and we are sometimes unable to identify the CPI assigned to the investigation following the initial hotline call. We ignore these unmatched observations.

B. Restrictions related to data cleaning:

(a) Restrict to investigations assigned via the rotation list. We do not consider investigations that CPIS assigns to CPIs "off-rotation." For example, CPIs can volunteer to take an investigation. In order to identify full-time CPIs who received

[64] Following DCYF Operating Procedure 500.0035.

their daily case assignment via the rotation list, we impose additional restrictions and do not consider investigations where CPIs were working primarily as hotline workers or investigations where CPIs had already received their daily assignment via the rotation list.

(b) Drop investigations based on alleged sex abuse. From conversations with DCYF, we understand that sometimes the Investigative Unit supervisor attempts to assign sex abuse cases to CPIs of the same gender as the child. This violates random assignment, and we therefore do not consider these investigations.

(c) Restrict to the first investigation observed for each child. We do not consider later investigations where the child reappears in the DCYF caseload.

(d) Drop if the associated CPI's removal tendency (see definition in Section 4) is calculated using less than 10 cases. We impose this restriction to avoid concerns regarding small cell sizes.

(e) Drop outliers based on the top or bottom 1 percent of CPI removal tendency.

The items listed (a) – (n) in Table B2 provide the number of distinct allegations, investigations, and children present in CPS case files after imposing the above restrictions. The first row shows that initial CPS records contain 187,023 allegations of abuse or neglect that are associated with 54,199 investigations and 63,351 children (more than one child can be part of the same investigation). The subsequent rows report the remaining number of observations after imposing data restrictions. For example, the row labeled (a) shows that there are 176,034 allegations of abuse or neglect associated with 51,864 investigations and 58,429 children. The last two rows of Table B2 report the final statistics for the number of younger and older children in the final DCYF investigations sample.

The key variables for children in the DCYF investigations sample are:

- Total days in foster care: Days spent in foster care as a result of the child's first investigation, from removal date to discharge date (also applies to days spent with relatives, with foster families, in group homes, and in other care).

- Adopted (=1): Indicator for child adopted upon discharge from foster care.

- Number of placements: Number of foster care placements as a result of the child's first investigation.

- Placed with relative (=1): Indicator for any placement with a relative as a result of the child's first investigation.

Table B2: Summary and Statistics for Data Restrictions

		(1) Allegations	(2) Investigations	(3) Children
Full DCYF data		187,023	54,119	63,351
1. Restrictions related to data cleaning				
a.	Restrict to children ages 0-18 with known demographics	176,034	51,864	58,429
b.	Restrict to the first allegations reported via the DCYF hotline	154,809	51,585	56,508
c.	Restrict to allegations involving a family	146,372	49,103	54,427
d.	Drop additional info. allegations	134,684	48,943	54,079
e.	Drop allegations not investigated	102,005	48,026	46,036
f.	Drop unfounded investigations	81,134	38,120	38,730
g.	Restrict to investigations from 2000 to 2015	71,451	33,492	34,364
h.	Drop if child in multiple investigations on the same date	71,278	33,418	34,348
i.	Restrict to investigations matched to a CPI assignment	70,039	32,845	33,971
2. Restrictions related to the research design				
j.	Restrict to investigations assigned via the rotation list	57,986	27,050	29,286
k.	Drop investigations involving sex abuse	54,697	25,312	27,798
l.	Restrict to the first investigation for each child	39,813	19,838	27,606
m.	Drop if the CPI removal tendency is calculated with $>$ 10 obs.	39,636	19,758	27,484
n.	Drop outliers in CPI removal tendency	38,957	19,330	26,954
Final DCYF investigation sample statistics				
	Young children (age $<$ 6)	18,611	11,509	13,834
	Older children (age $\geq$ 6)	20,346	9,853	13,120

Notes: This table summarizes the data restrictions and the resulting number of allegations, investigations and children present in the CPS case files after imposing the associated restriction.

- Police notified (=1): Indicator for whether police were notified during the investigation.

Note that we focus on these outcomes measured for investigations from 2000-2015 to ensure an uncensored foster care placement measure. For children still in care as of Jan. 1, 2018, foster care variables (e.g., total days in care, days spent with relatives) are measured as of Jan. 1, 2018.

B.3.2 Sample of Investigated Children with Test Scores

We standardized RIDE math and reading test scores to have a zero mean and a standard deviation equal to one by school year and grade. We also created an average standardized test scores that averages math and reading scores. Then, we join the DCYF sample to RIDE standardized testing records for grades 3-8, available as a yearly panel from 2005 to 2016. There are several reasons why a child from the DCYF sample may not have testing records for grades 3-8. Children born before 1995 or after 2008 will not have observations in the panel because they are either too old or too young to be enrolled in the testing grades during the sample period. Children who enroll in private school will not have observations in the panel, nor will children who leave the state of Rhode Island. The DCYF-RIDE matched sample for test scores contains 2,721 young girls and 3,148 young boys who have both standardized math and reading testing records in at least one year.

The key variables for children in the test-score sample are:

- Reading z-score: Reading test score, standardized with mean equal to zero and standard deviation equal to one at the grade and year level among the full population of tested students in Rhode Island.

- Math z-score: Math test score, standardized with mean equal to zero and standard deviation equal to one at the grade and year level among the full population of tested students in Rhode Island.

- Average z-score: The mean of a child's reading and math z-scores.

Note that we standardized these scores to maintain comparability across testing years.

B.3.3 Sample of Investigated Children with Non-test-score Outcomes (Grade Retention, Special Education (IEP), and Absences)

We join the DCYF sample to RIDE public school enrollment records (2003-2016) to construct additional schooling outcomes at the child level. We create measures defined for outcomes that occur during grades 3-8. Similar to the test score sample, children born before 1989 or after 2008 will not have observations because they are too old or too young. We do not condition on test taking when constructing these measures. The DCYF-RIDE matched sample for non-test score outcomes contains 2,778 young girls and 3,225 young boys.

The key variables for children in the non-test score sample are listed below.

- Retention (=1): Indicator for ever repeating a grade over two consecutive years in grades 3-8. (This is missing for students not observed in two consecutive years.)

- IEP (=1): Indicator for any enrollment in an Individualized Education Program (IEP; i.e., special education) in grades 3-8.

- Absences: Average yearly absences (excused and unexcused) in grades 3-8. We set the top percentile in school absences to missing as these students were likely not enrolled .

- School index: The mean of the retention, IEP, and absences outcomes, where standardize each outcome by gender and age group (e.g., those younger than 6 years old at the time of an investigation).

For consistency, we restrict analysis in Section 5.2 to children with complete information for retention, special education participation, and absences.

B.3.4 Sample of Investigated Children for Enrollment and Test-taking Analysis

For children in the DCYF sample, we create a panel from 2003 to 2016 that is balanced at the child and school-year level. This panel includes children who are expected to be enrolled in grades 3-8 based on their age. We join this panel to RIDE public school records to determine effective enrollment in public school during grades 3-8. Similar to the test score and other schooling samples, children who were born before 1989 or after 2008 will not have observations because they are too old or too young. We consider only post-investigation years and create an indicator for whether a child was enrolled in a year.

We conduct the analogous exercise with RIDE testing records to create an indicator for whether a child took a standardized test during ages 8-13. The DCYF-RIDE matched sample for enrollment and test-taking contains 4,101 young girls and 4,750 young boys.

The key variables for children in the enrollment and test-taking sample are:

- Enrolled (=1): Indicator for enrollment in RI public school, defined as a panel outcome for children who are ages 8-13 or 9-14 (depending on date of birth) in a given school year.

- Tested (=1): Indicator for having taken a standardized test, defined as a panel outcome for chidlren who are ages 8-13 or 9-14 (depending on date of birth) in a given school year.

B.3.5 Sample of Investigated Children for Enrollment and Test-taking Analysis

For children in the DCYF sample, we create a panel from 2003 to 2016 that is balanced at the child and school-year level. We join this panel to RIDE public school records to generate measures of mobility (e.g., school change) and the characteristics of schools attended during grades 3-8. Similar to the schooling sample, children born before 1989 or after 2008 will not have observations because they are too old or too young. If a child attends multiple schools in a year, we consider the characteristics only of the first school attended. The DCYF-RIDE matched sample for school mobility and characteristics contains 2,855 young girls and 3,325 young boys.

The key variables for children in the mobility and school characteristics sample are:

- Moved Schools (=1): Indicator for changing schools.

- School value-Added: We construct a school-level value-added measure that considers tests taken by RI students in grades 4-8. We restrict to students not in the DCYF sample. We exclude test scores for students repeating grades and for students missing any of the baseline controls used in the value-added estimation. We estimate a school's value-added measure (μ) from the following student-level regression:

$$A_{ijt} = X_{ijt}\beta + \nu_{ijt}$$

where

$$\nu_{ijt} = \mu_j + \epsilon_{ijt}.$$

For each child i in school j in year t, we observe the dependent variable A_{ijt} as the child's test score (standardized by grade and year). We include a vector of control variables X_{ijt} that includes race, gender, special education status, English learner status, free/reduced price lunch status, and a cubic in lagged test scores. The residual v_{ijt} is composed of the school j's value-added measure (μ) and an error term. To match the students in the DCYF sample to measures of school value-added, we assign the value-added measure to the first school attended in each of grades 3-8. The final outcome is the mean of the school value-added measure for schools that a child attends in grades 3-8.

- School avg. test scores: The raw average standardized test score for each school, used in the calculation of the value-added measure described above. We restrict to students not in the DCYF sample.
- School % Minority: Fraction of non-white students in the child's school, measured at the school-year level. We restrict to students not in the DCYF sample.
- School % FRL: Fraction of students with free/reduced price lunch subsidies at the child's school, measured at the school-year level. We restrict to students not in the DCYF sample.

B.3.6 Samples for Older Investigated Children

We also create different samples to analyze outcomes of older investigated children (ages 6-18 at the time of an investigation). For short-run outcomes, we examine test score and non-test score school outcomes for older children. The matched DCYF-RIDE sample for test scores contains 2,581 older girls and 2,965 older boys who have both standardized math and reading testing records in at least one year. The matched DCYF-RIDE sample for non-test-scores (grade retention, participation in special education, and absences) contains 3,029 young girls and 3,440 young boys. Note that this sample comprises only older children who have records in years *after* their first investigation. For example, we *do not* study a child's third grade standardized test score if the child was enrolled and took an exam in grade 3 at the time of the DCYF investigation. Instead, we focus on their post-investigation exams in grades 4-8.

For older children investigated at age 6-18, we also create samples to study the following later-life outcomes: delinquency, high school graduation, teen births, and college enrollment. We construct a different sample for each outcome, based on the time period available for each outcome and the expected age of the investigated children during the time period. The restrictions ensure that outcomes are uncensored and that children are observable in the post-investigation period. (See list below for further details on restrictions for each outcome.)

The variables used in the analysis of outcomes for older investigated children are:

- Average z-score: The mean of a child's reading and math z-scores. All measures are based on scores observed after the year of the investigation.

- School index: The mean of the retention, IEP, and absences outcomes, where each outcome has been standardized by gender and age group (e.g., less than age 6 at the time of an investigation). All of the components of the index are based on outcomes observed after the year of the investigation.

- Delinquent (=1): Indicator for RITS enrollment or probation for wayward or delinquent offenses at ages 12-18. Eligible children are those investigated prior to the age of 12 and are born between 1988 and 1998 so that they are observable at ages 12-18 in juvenile delinquency records.

- HS Grad. (=1): Indicator for graduation from a RI high school at ages 18-19. Eligible children are investigated prior to the age of 18 and are born between 1985 and 1997 so that they are observable at ages 18-19 in RIDE public school records.

- Teen Birth (=1): Indicator for presence in the DOH vital records as a teen parent at ages 15-19. Eligible children are investigated prior to the age of 15 and are born between 1985 and 1997 so that they are observable at ages 15-19 in vital records.

- College (=1): Indicator for any post-secondary educational institution enrollment at ages 18-20. Eligible children are investigated prior to the age of 18 and are born between 1986 and 1995 so that they are observable at ages 18-20 in NSC records.

B.3.7 Sample of Parent Perpetrators

For nearly all children in the DCYF sample (99 percent), we observe the set of perpetrators associated with allegations of abuse or neglect. We focus on parent perpetrators, which make up 95 percent of the perpetrators for children in the DCYF sample.[65] We join the sample of DCYF parent perpetrators to criminal justice records (1995-2017). The outcomes that we consider are whether the perpetrators were ever charged with a crime or incarcerated in the post-investigation years. We construct a four-year post-investigation measure that is partially censored for perpetrators investigated in 2014 or later, as well as a two-year measure that is uncensored.

The variables used in the analysis of perpetrators are described below.

- Charged/incar., 2-year post: Indicator for whether the parent perpetrators of abuse or neglect were charged with a crime or incarcerated in the 2-year post-investigation period.

- Charged/incar., 4-year post: Indicator for whether the parent perpetrators of abuse or neglect were charged with a crime or incarcerated in the 4-year post-investigation period.

B.4 Description and Statistics for Child Protection Investigators (CPI)

As detailed in Section 3 and Appendix B.3.1, we create a sample of XXXX young children subject to a substantiated (founded) DCYF investigation. There are 102 child protection investigators (CPIs) associated with these investigations. Table B3 reports statistics for the first and repeat investigations assigned to the 102 CPIs. First refers to whether the investigation is the initial case that we see for the associated child. We provide statistics for first and repeat investigations because we use both in the preferred definition of our instrumental variable. By using first and repeat investigations, we have more information to use to infer removal tendencies.

To summarize, the average CPI handles investigations in eight of the years covered by the DCYF records (2000-2015). The average CPI makes decisions for 387 children and removed 70 children over the entire period that we observe them. The average CPI is first observed (in the administrative records) in 2003, and the median CPI is first observed in 2000. The average CPI is last observed in 2011, and the median CPI is last observed in 2013.

Note that we calculate the main instrument separately for the 2000-2007 and 2008-2015 periods. Table B3 also provides the average CPI statistics in each of these eight-year periods. (When a CPI is not observed in one of the two periods, we include a 0 in computing these summary statistics. There are 30 CPIs who only appear in 2000-2007 period. There are 12 CPIs who only appear in the 2008-2015 period.) In each period, the average CPI handles nearly 200 cases and removed around 35 children.

[65] Note that restricting the sample to children with parent perpetrators does not imply perpetrators live in the same location (or home) as the child.

Table B3: CPI Summary Statistics

	(1) Mean	(2) p10	(3) p50	(4) p90	(5) N
All Years					
# Years	8.13	2.00	8.00	14.00	102
# Children	387.09	55.00	304.50	796.00	102
# Removed Children	69.56	10.00	60.00	142.00	102
Year Start	2003	2000	2000	2009	102
Year End	2010	2004	2013	2015	102
Period 2000-2007					
# Children	188.51	0.00	182.50	384.00	102
# Removed Children	35.81	0.00	32.50	73.00	102
Period 2008-2015					
# Children	198.58	0.00	66.00	575.00	102
# Removed Children	33.75	0.00	13.50	100.00	102

Notes: This table presents summary statistics for the sample of 102 CPIs that are associated with the children in the DCYF investigations sample.

B.5 Sibling Statistics

As detailed in Section 4 of the main text and Appendix Section Y, the sample created from the DCYF investigations records contains 13,834 children investigated before age 6. These children are associated with 9,749 cases. In 6,778 of these cases (70 percent), there is only a single young child. The remaining 2,971 cases contain siblings. At the case level, the average number of young children is 1.42.

C Complier Calculations

This section provides details on how we estimate the characteristics and outcomes of compliers in our sample.

C.1 Estimating Complier Characteristics

In the child protective service context, we define compliers as children whose removal decision would have been different if they had been assigned to the most lenient (i.e., less likely to recommend a removal from home) instead of the strictest investigator (CPI). We follow the approaches developed by Abadie (2003), Dahl et al. (2014), and Dobbie, Goldin and Yang (2018) to characterize compliers in the sample of investigated children.

Let $\overline{z}$ denote the maximum value of the instrument (the most strict investigator) and $\underline{z}$ denote the minimum value of the instrument (the most lenient investigator). By the monotonicity and independence assumptions, we define the share of compliers as:

$$p_c = Pr(R_i = 1|Z_i = \overline{z}) - Pr(R_i = 1|Z_i = \underline{z}) = Pr(R_i(\overline{z}) > R_i(\underline{z})), \tag{C1}$$

where R_i is an indicator for removal. In practice, we assign the top percentile of our instrument to $\overline{z}$ and the bottom percentile of our instrument to $\underline{z}$. As discussed in Dahl et al. (2014) and Dobbie, Goldin and Yang (2018), the share of compliers can be directly estimated as $p_c = \alpha$, where α is the coefficient on the instrument from the first stage regression (Equation 2).

This is useful for studying the characteristics of compliers. For binary characteristic x_i, we know that:

$$\begin{aligned}\frac{Pr(x_i = 1|R_i(\overline{z}) > R_i(\underline{z}))}{Pr(x_i = 1)} &= \frac{Pr(R_i(\overline{z}) > R_i(\underline{z})|x_i = 1)}{Pr(R_i(\overline{z}) > R_i(\underline{z}))} \\ &= \frac{\mathbb{E}(R_i|Z_i = \overline{z}, x_i = 1) - \mathbb{E}(R_i|Z_i = \underline{z}, x_i = 1)}{\mathbb{E}(R_i|Z_i = \overline{z}) - \mathbb{E}(R_i|Z_i = \underline{z})}\end{aligned} \tag{C2}$$

This expression shows that the relative characteristics of compliers can be recovered by computing a ratio where the numerator is obtained by estimating the first stage coefficient for the subgroup $x_i = 1$ and constructing $\alpha_x(\overline{z}-\underline{z})$. The denominator is constructed similarly using the entire sample to estimate a first stage equation.

C.2 Estimating Complier Outcomes When Not-Removed

Our IV estimates are the causal impact of removal for compliers (i.e., the children whose removal decision would have been different if they had been assigned the most lenient instead of the strictest investigator). In other words, the estimates tell us about the impact of removal for a child on the marginal case. To better understand this impact, it is helpful to have a benchmark comparison by estimating complier outcomes when removal *does not* occur. To answer this question, we need to estimate the untreated potential outcome (denoted Y_{i0}) for compliers:

$$\mathbb{E}(Y_{0i}|R_i(\overline{z}) > R_i(\underline{z})) \tag{C3}$$

As discussed in Dahl et al. (2014), this can be obtained by examining children who are assigned to lenient and strict investigators: For non-removed children (i.e., those with $R_i = 0$) assigned to $\underline{z}$, we know:

$$\begin{aligned}\mathbb{E}(Y_i|R_i = 0, Z_i = \underline{z}) = &\left(\frac{p_c}{p_c + p_n}\right)\mathbb{E}(Y_{0i}|R_i(\overline{z}) > R_i(\underline{z})) \\ &+ \left(\frac{p_n}{p_c + p_n}\right)\mathbb{E}(Y_{0i}|R_i(\overline{z}) = R_i(\underline{z}) = 0)\end{aligned} \tag{C4}$$

where Y_i is the observed outcome, p_c is the share of compliers, and p_n is the share of never-takers (i.e., children who would never be removed by the most or least strict investigator).

The outcomes for never-takers can be inferred from the outcomes of the non-removed children who are assigned the strictest investigator:

$$\mathbb{E}(Y_{0i}|R_i(\overline{z}) = R_i(\underline{z}) = 0) = \mathbb{E}(Y_i|R_i = 0, Z_i = \overline{z}) \tag{C5}$$

Equation C5 allows us to disentangle the mixture from Equation C4. Specifically, we can re-write Equation C4 as:

$$\begin{aligned}\mathbb{E}(Y_{0i}|R(\overline{z}) > R(\underline{z})) = &\left(\frac{p_c + p_n}{p_c}\right)\mathbb{E}(Y_i|R_i = 0, Z_i = \underline{z}) \\ &- \left(\frac{p_n}{p_c}\right)\mathbb{E}(Y_i|R_i = 0, Z_i = \overline{z})\end{aligned} \tag{C6}$$

To evaluate this expression, we estimate the share of always-takers, never-takers and compliers in the sample.[66] With these quantities, we solve Equation C6 by estimating a linear model for Y_i and z_i in the subsample of non-removed children (i.e., $R_i = 0$). In this specification, we control for investigation year fixed effects.

[66] Recall that $p_a = Pr(R_i = 1|Z_i = \underline{z})$ and $p_n = Pr(R_i = 0|Z_i = \overline{z})$.

D Details on the Machine Learning Approach

As discussed in Section 5.5, we use conduct a complementary approach where we use a machine learning (ML) approach and re-estimate the impact of removal on school outcomes. Following Mueller-Smith (2015), this approach relaxes the monotonicity assumption from our main identification strategy by selecting measures of CPI removal tendencies that vary with case characteristics. We use LASSO regressions to select from a set of potential instruments and use the selected instruments in our two-stage least squares models. The following sections describe the implementation of this approach in more detail.

D.1 Constructing Flexible Leave-out Measures for Machine Learning

As a first step, we build removal tendency measures that vary with case characteristics. We focus on the following six characteristics:

1. gender;
2. minority (ethnic/race) status (non-minority and minority, respectively);
3. marital status;
4. reporter type;
5. allegation type;
6. investigation level.

For each characteristic, we define mutually exclusive groups of children and calculate the leave-out measure of removal tendency based on the CPIs tendency for the group.[67] For example, each CPI will have a leave-out removal tendency calculated separately for minority (non-white) and non-minority (white) children. We do this for six characteristics and create six versions of leave-out measures of removal tendency. (We do not consider any interactions between case characteristics.)

To parallel our main measure of CPI removal tendency, we create the case characteristic-specific instruments over two eight-year periods (2000-2007 and 2008-2015). By calculating the measure using an eight-year period, we address concerns that a CPI may see relatively few children who have a given case characteristic in a shorter period (e.g., one year). Table D1 on page Appendix - 35 provides statistics for CPIs on the number of investigated children

[67]Each removal tendency by case characteristics uses the categories included in Table 1, except for minority status which divides white and non-white ethnicities/races into two categories. We exclude family language as a characteristic since 97 percent of children belong to English speaking families, resulting in a tendency measure close to our standard measure.

by types of case characteristics. For example, the first rows show that the average CPI investigates about 387 children over the entire 2000-2015 period. In addition, the average CPI investigated 150 and 237 and minority (non-white) and non-minority (white) children. The means in Column 1 show that the average CPI always sees more than 10 cases for a given type of case characteristic. Column 2 shows that some CPIs investigate relatively few children for some of the groups (e.g., "other" types of reporters). In our implementation, we address concerns over small-sizes by defining a given CPI tendency measure to be missing when there are fewer than 10 children available to construct the leave-out measure. For example, if a CPI investigates only 9 children whose cases were at the emergency level over the relevant period, we define the instrument that varies at the investigation level to be missing for this CPI.

D.2 Machine Learning Implementation Details

As discussed in Section 5.5, we test the robustness of our main results using an alternative machine-learning (ML) approach that allows the instrument to vary with case characteristics. We consider six types of case characteristics: (1) gender, (2) minority (ethnic/race) status (non-minority and minority, respectively), (3) marital status, (4) reporter type, (5) allegation type, and (6) investigation level. We have six potential instruments to use in our first-stage removal equation. Following Mueller-Smith (2015), we use LASSO to select the instruments with the greatest predictive power (Belloni et al., 2014).

In Appendix Table A10, we present our main results based on the ML approach. For each outcome, we estimate separate LASSO regressions of removal on the six removal tendencies by case characteristics to select instruments for use in the first stage. Each of these regressions always specifies investigation year fixed effects and case characteristic controls as variables always selected, as these controls are included in our two-stage least squares specifications. Table D2 reports which instruments are selected by LASSO for each outcome for young girls and young boys. For young girls, the test score outcomes (average, math, and reading) use instruments that vary based on gender, minority status, and reporter type. For young girls, the school index, retention, and IEP outcomes use instruments that vary based on gender and minority status. For young boys, the test score outcomes (average, math, and reading) use the instruments that vary by minority status and investigation level. For young boys, the school index, retention and IEP outcomes use only the instrument that varies by minority status.

Table D1: CPI Summary Statistics by Case Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	2000-2015 (# CPIs=102)				2000-2007 (# CPIs=90)				2008-2015 (# CPIs=72)			
	Mean	p10	p50	p90	Mean	p10	p50	p90	Mean	p10	p50	p90
All children	387.1	55.0	304.5	796.0	213.6	47.0	201.0	390.5	281.3	12.0	205.5	638.0
Girls	193.6	27.0	149.5	405.0	108.1	24.0	103.0	196.5	139.1	5.0	98.0	390.0
Boys	193.5	28.0	152.5	395.0	105.5	21.0	101.5	199.5	142.3	6.0	106.5	328.0
Non-Minority	236.7	34.0	196.0	488.0	137.7	28.0	134.5	258.5	163.2	8.0	119.5	351.0
Minority	150.4	18.0	108.0	352.0	75.9	16.5	68.5	139.0	118.2	2.0	83.0	275.0
Married couple	63.5	8.0	56.0	129.0	40.5	5.5	38.0	75.0	39.4	2.0	26.5	94.0
Unmarried couple	90.0	8.0	55.0	221.0	35.4	4.0	30.5	69.5	83.2	3.0	56.5	193.0
Single/Other	233.6	40.0	202.0	468.0	137.8	33.0	131.5	262.0	158.7	5.0	121.0	349.0
Neglect	288.6	38.0	208.5	624.0	149.9	33.5	138.5	267.5	221.5	10.0	151.5	505.0
Physical neglect	14.3	1.0	12.0	31.0	9.3	0.5	9.0	18.0	8.5	1.0	6.5	20.0
Physical abuse	69.8	11.0	64.0	141.0	44.3	9.5	43.5	85.5	43.5	1.0	37.5	96.0
Professional reporter	309.5	37.0	227.0	629.0	164.0	32.5	157.0	310.5	233.5	12.0	164.0	517.0
Other reporter	20.5	4.0	19.0	43.0	14.1	1.0	11.0	27.0	11.5	0.0	8.0	27.0
Family/friend reporter	57.0	9.0	49.0	115.0	35.5	7.0	35.0	66.0	36.0	0.0	27.5	88.0
Routine	111.4	14.0	81.0	235.0	60.7	11.5	55.0	119.0	81.9	4.0	55.0	174.0
Immediate	237.9	33.0	186.5	534.0	128.2	26.5	123.0	237.5	176.7	8.0	130.5	393.0
Emergency	37.8	7.0	32.5	70.0	24.7	6.0	23.5	46.0	22.7	1.0	19.0	50.0

Notes: This table presents summary statistics for the sample of 102 CPIs that are associated with the children in the DCYF investigations sample. The rows provide summary statistics based on case characteristics. For example, the second row provides summary statistics for the number of girls involved in a CPI's cases during different time periods. Column 1 shows that the average CPI had 193 girls in their cases during 2000-2015.

Table D2: Instrument(s) Selected by LASSO for ML Approach

	Young Girls (Age < 6)		Young Boys (Age < 6)	
	(1)	(2)	(3)	(4)
Instruments by case characteristics:	Test score outcomes	Schooling outcomes	Test score outcomes	Schooling outcomes
Gender	Yes	Yes		
Minority	Yes	Yes	Yes	Yes
Marital status				
Reporter type	Yes			
Allegation type				
Investigation level			Yes	

Notes: This table reports the version of the removal tendencies by case characteristic instruments selected in each LASSO regression. Columns 1 and 2 report the selected instruments (denoted by "Yes") for the test score and schooling outcomes of young girls. Columns 3 and 4 reports the selected instruments for the test score and schooling outcomes of young boys. The LASSO regressions always specify investigation year fixed effects and case characteristic controls as variables always selected.

E Additional Discussion of Impacts for Older Children

As discussed in Section 7, we hope to estimate the causal impact of home removal for older children investigated at ages 6-18. To assess the validity of our IV approach, we examined the relationship CPI removal tendency and the case characteristics for older children. The randomization test results in Appendix Table A14 show that, while we do not reject our null hypothesis of no joint significance of case characteristics in the sample of older boys, we reject the null hypothesis in the sample of older girls. Examining the regression results for older girls in Column 2 shows that there are four case characteristics (out of fourteen) that have coefficients that are significant at the 10-percent level or lower. The largest statistically significant coefficient is equal to roughly one quarter of a standard deviation of CPI removal tendency.

To help assess whether this imbalance threatens the validity of IV estimates for older children, we conduct two tests. First, we examine estimates of the impact of removal with and without controls for case characteristics. Altonji et al. (2005) suggest that assessing whether point estimates are sensitive to the inclusion of controls can provide informative on the extent of selection bias in some situations. In Table E1, we restrict our analysis to older (ages 6-18) investigated children and present IV estimates for test scores, the school index (which is based on retention, special education participation and absences) and post secondary schooling attendance with and without case characteristic controls. For older girls, there is no strong pattern of coefficient sensitivity for these outcomes. For example, the point estimates for the school index are -0.398 and -0.411, respectively. The point estimates for older boys also display no strong pattern of sensitivity, which is expected given that the results in Appendix Table A14 provide no evidence of a relationship between case characteristics and CPI removal tendency for older boys.

In our second test, we assess the validity of our IV approach by examining test scores in the periods before an investigation begins for older children.[68] Due to the random assignment of cases, we expect that there should be no statistically significant relationship between removal (and our instrument) and the "pre-treatment" test score outcomes. To conduct this test, we construct a panel of test scores for older investigated children that includes observations from school years that *precede* the year of the first investigation. For most older removed children, we observe two test scores that precede the year of the first investigation. For the purpose of comparison, we also include observations in the panel for the year of the investigation and three school years that follow. Using the panel of test scores for older children, we estimate separate IV models where the dependent variable is the average of

[68]Note that we cannot conduct this type of analysis for young children because their first investigation occurs before they enter testing grades.

standardized test scores in a given school year. We estimate seven models starting with observations that are two years before the year of a DCYF investigation and ending with the school year that is three years after a DCYF investigation. Figures E1 displays the point estimates and confidence intervals associated with these estimates. The y-axis displays the year relative to investigation. For example, the left-most point estimate for older girls shows that there is an insignificant 0.19 standard deviation impact of removal on test scores that occur two years prior to the investigation. Across the school years that we examine, there are no statistically significant impacts of removal (including in the years that follow an investigation). The results for test scores that occur two years and one year before an investigation provide no strong evidence that CPI removal tendency is correlated with child characteristics, although the standard errors associated with our estimates are large and the confidence intervals span from -1 to 1 standard deviation.

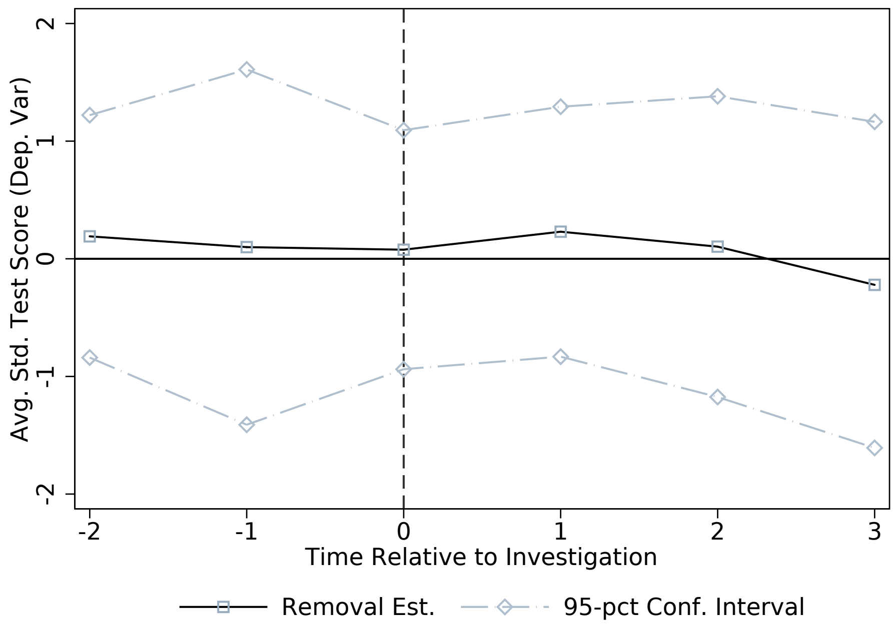
Table E1: Impact of Removal on Outcomes of Older Children, Sensitivity Test

	Panel A. Older Girls (Age $\geq$ 6)					
	School-age outcomes		Later-life outcomes			
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Average z-score		School Index		College (=1)	
Removed (= 1)	-0.109	-0.230	-0.341	-0.373	0.155	0.133
	(0.625)	(0.582)	(0.347)	(0.326)	(0.228)	(0.222)
Mean of dependent variable	0.068	0.068	-0.005	-0.005	0.303	0.303
Complier mean if not removed	-0.337	-0.337	0.138	0.138	0.302	0.032
Case controls	No	Yes	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	24.050	22.129	37.376	35.092	36.367	38.718
N	7,517	7,517	3,029	3,029	3,326	3,326
Individuals	2,581	2,581	3,029	3,029	3,326	3,326
	Panel B. Older Boys (Age $\geq$ 6)					
Dependent variable:	Average z-score		School Index		College (=1)	
Removed (= 1)	-0.250	-0.237	0.352	0.323	-0.147	-0.127
	(0.458)	(0.429)	(0.219)	(0.216)	(0.194)	(0.187)
Mean of dependent variable	0.053	0.053	-0.003	-0.003	0.239	0.239
Complier mean if not removed	-0.414	-0.414	-0.297	-0.297	0.367	0.367
Case controls	No	Yes	No	Yes	No	Yes
Investigation year FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic (instrument)	27.910	30.911	30.435	34.273	37.069	41.145
N	8,838	8,838	3,440	3,440	2,953	2,953
Individuals	2,965	2,965	3,440	3,440	2,953	2,953

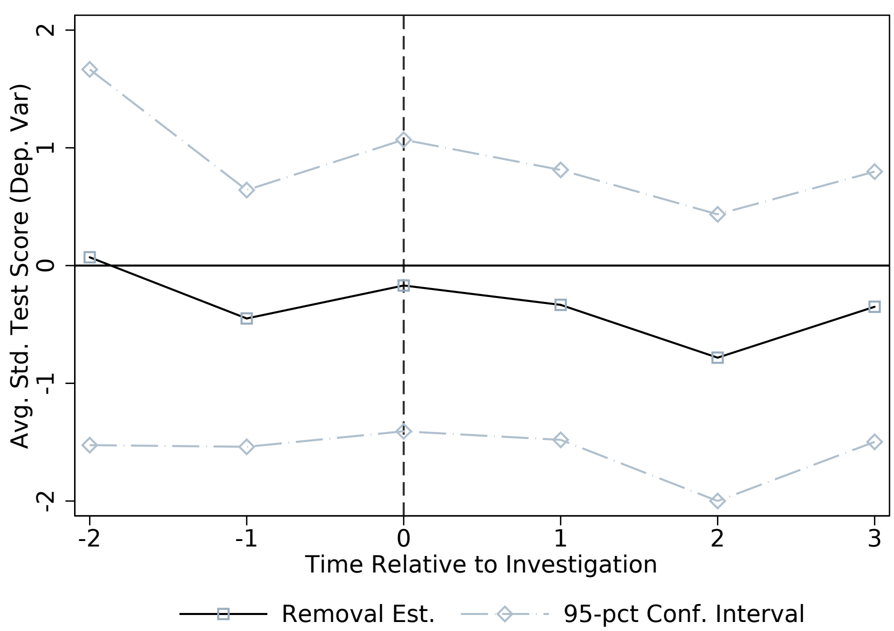
Notes: This table reports results for the impact of removal on outcomes for older girls (Panel A) and older boys (Panel B). Older is defined as being investigated at ages 6 or later (up to age 18). All results are two-stage least squares models with the standard leave-out measure of CPI removal tendency as an instrument for removal. The school index is constructed based on standardized measures of whether an investigated child was ever retained, ever participated in special education (i.e., has an IEP), and the average number of days absent during grades 3-8. Standard errors are two-way clustered at the family and CPI level in parentheses. Significance reported as ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Figure E1: Impact of Removal on Test Scores of Older Children, by Time Relative to Investigation

(a) Young Girls



(b) Young Boys



Notes: These figures show results for the impact of removal on test scores estimated in separate regressions by time relative to the year of investigation for young girls (Panel A) and young boys (Panel B). All results are two-stage least squares models with the standard leave-out measure of CPI removal tendency as an instrument for removal. All models include controls for the case characteristics in Table 1 and investigation year fixed effects. Confidence intervals are based on standard errors that are two-way clustered at the family and CPI levels.