Algorithms and Decision-making: Evidence from Child Maltreatment Reports[1]

Maria D. Fitzpatrick[2] Katharine Sadowski[3] Christopher Wildeman[4]

Abstract

Around 40% of children experience maltreatment (Finkelhor et al. 2013), with harmful outcomes and high social costs. Child protection decisions are complex, and potentially biased and/or prone to errors due to underfunding and overload. Our randomized controlled trial showed algorithmic tools sped up decision-making but didn't significantly change child outcomes, perhaps because COVID disruptions limited outcome analysis. Results are suggestive that risk aversions played a role: high-risk cases flagged by the tool were more likely screened in, while low-risk cases weren't more likely screened out. Time savings from the tool could enable caseworkers to spend more time directly with families.

JEL Codes: C93, J13, and I38

[1] We would like to thank the Douglas County Human Services team, particularly Daniel Makelky, Ruby Richards, and Nicole Adams, and the Centre for Data Analysis at Auckland University of Technology team, particularly Rhema Vaithianathan and Emily Putnam-Hornstein for their work to design and implement the Douglas County Decision Aid tool in the field. Invaluable comments and suggestions were shared by Jason Baron, Doug Miller, and other colleagues. We also thank seminar participants at Cornell University, University of California – Merced, Clemson University, the NBER Children Summer Institute 2022, the NBER Artificial Intelligence Conference Fall 2022, Purdue University, Duke University, and the 2024 ASSA Conference.

[2] Maria Fitzpatrick is Professor of Economics and Public Policy in the Jeb E. Brooks School of Public Policy at Cornell University and Research Associate at National Bureau of Economic Research. Email: maria.d.fitzpatrick@cornell.edu.

[3] Katharine Sadowski is an Assistant Professor in the Graduate School of Education at Stanford University and an Affiliate at the Center for Early Childhood.

[4] Christopher Wildeman is a Professor of Sociology and Sanford School of Public Policy at Duke University and an Affiliate of the Center for Child and Family Policy.

doi:10.3368/jhr.0224-13437R2

Disclosure Statements

Funding Organizations: Fitzpatrick and Wildeman each received $50,000 from Douglas County Human Services to conduct this evaluation. Sadowski received no funding for this evaluation.

Support: Fitzpatrick has received support from the Children's Bureau for her role as co-Director of the National Data Archive on Child Abuse and Neglect. Fitzpatrick received support from the New York City Office of Student Health for research on child health in schools. Fitzpatrick also received research support from the Hoover Institution at Stanford University.

Wildeman received support from the Children's Bureau for his role as Director of the National Data Archive on Child Abuse and Neglect. He also received support from the ROCKWOOL Foundation for research and training activities. Wildeman has also received research grants from the following organizations over the past three years: the ROCKWOOL Foundation, Russel Sage Foundation, Social Science Research Institute, New Yorkers for Children, FWD.us, Duke Population Research Institute, Eunice Kennedy Shriver National Institute of Child Health & Human Development, National Institute of Aging, National Institute of Justice, Casey Family Programs, National Science Foundation, and Conrad N. Hilton Foundation.

Sadowski received support from the New York City Office of Student Health for research on child health in schools. Sadowski received data from the Texas Education Agency and the United States Census Bureau with funding from the Horowitz Foundation for Social Policy, the Russell Sage Foundation, and the W.E. Upjohn Institute for Employment Research.

Positions: None to report.

Relatives/Partners: None to report.

Review of the Work: The Douglas County Human Services team had the right to review the work prior to its circulation.

Institutional Review Board (IRB) Approval: This project was approved by the Cornell University IRB under Protocol Number IRB0009859.

Data Replication Statement

The data used in this research study are administrative data from the Douglas County, Colorado Department of Human Services. They are confidential in nature and cannot be made publicly available. The data can be obtained by filing a request directly with the Douglas County Human Services. The authors are willing to assist (maria.d.fitzpatrick@cornell.edu).

I. Introduction

Child maltreatment, which encompasses both neglect and physical, sexual, and emotional abuse, is a widespread problem in the United States. In 2019, there were roughly 3.5 million reports of maltreatment, 656,000 of which were confirmed (U.S. Department of Health and Human Services 2021). Recent estimates indicate that the lifetime costs of maltreatment are as high as 4 percent of GDP (Fang et al. 2012). Both over-investigation (investigating children who are not maltreated, type I error) and under-investigation (not investigating children who are maltreated, type II error) are core challenges in the child welfare system. Government agencies tasked with managing child maltreatment reporting, investigation, and intervention are perennially underfunded and understaffed, leading to caseworkers managing nearly double the recommended caseload of 12 to 15 (Hughes and Lay 2012; AFSCME 2016; Child Welfare Information Gateway 2016). Interviews with state child welfare officials state these administrative burdens reduce caseworkers' ability to effectively conduct timely investigations and make decisions (AFMSCE 2016). Yet, little causal evidence about cost-effective ways to either improve system burden or decrease the risk of both types of errors exists.

In this study, we conducted a randomized controlled trial (RCT) in Douglas County, Colorado, to evaluate the impact of a new algorithmic decision-making tool on child welfare screening decisions and processing time. The tool, known as the Douglas County Decision Aid (DCDA), was adopted by the Douglas County, Colorado Department of Human Services (DCDHS) in 2019 to assist child welfare agency teams in determining which child maltreatment reports should be investigated. The DCDA presents i) a risk score from 1 to 20 indicating where a referral falls in the distribution of predicted likelihood of removal from the home within 2 years and ii) the average probability of removal for referrals across the score distribution. In our study,

caseworker teams – who were already randomly assigned referrals through the county's existing process – were randomly assigned to receive access to the DCDA or not.[i] Those assigned to view the DCDA were instructed to do so only after reaching an initial consensus on whether to recommend an investigation. This design preserved caseworker autonomy but may have limited the tool's influence if workers were reluctant to reverse initial decisions, particularly due to concerns about the costs of Type II errors (i.e., reversing an initial decision to investigate can be costly). Seventy percent of caseworker teams viewed the algorithm when assigned.

In our analysis, we find that investigations increased as the result of the DCDA availability. During the period of the RCT, the availability of the DCDA increased investigations by 1.9 percentage points (5 percent). However, this result is not statistically significant, and, across the whole sample, we do not find statistically significant results for the other child welfare outcomes available to us, including whether a referral was substantiated, re-referred, or led to a placement. The DCDA did, however, lead teams to make decisions more quickly. Each decision was, on average, made half a minute (~5 percent) faster. Given the number of decisions and workers, full adoption of the DCDA for a year would save 2 weeks of caseworker time per year.

The limited effects of the DCDA across the whole sample may mask different types of changes for low-risk and high-risk children. To investigate whether this was the case, we examined the relationship between the risk score of a referral, the use of the DCDA tool, and the outcomes of interest. Although we take great caution in making conclusions based on these results because they are underpowered, there is evidence that (1) initial decisions to investigate were unlikely to be reversed when the DCDA indicated low risk but (2) initial decisions not to investigate were more likely to be reversed when the DCDA indicated high risk. It appears that

the DCDA tool accelerated identification of maltreatment in high-risk environments in ways that may allow for earlier intervention and the prevention of further CPS involvement.

In the absence of the DCDA, child welfare agency workers likely make decisions based on information they consider correlated with the risk of maltreatment. For example, caseworkers tend to consider referrals from law enforcement officers to be higher risk than those from other reporters and referrals related to infants as higher risk than those for older children. The DCDA provides a measure of risk based on more comprehensive information, which may change how caseworkers assess the risks related to these proxies for risk that they used before. To see if this is the case, we explored how relative rates of screening a child in for investigation changed for referrals based on reporter type, risk score, and availability of the DCDA. We also did the same based on the age of the child. The results are suggestive that DCDA availability leads to more investigation of high-risk referrals with reporter types (e.g., school staff) and child age that would have suggested they were low-risk and less investigation of low-risk referrals whose reporter type (e.g., law enforcement) and age would have suggested were high risk. However, the results are underpowered and we cannot rule out that there are no differences in how DCDA availability affected screen-in rates across these groups.

Our results represent an important contribution to the limited literature that identifies the effects of interventions designed to improve child welfare reporting, investigation, and intervention processes. To date, most causal evidence in child welfare has focused on the effects of foster care placement on child outcomes (e.g., Doyle 2007; Doyle 2008; Doyle 2013; Baron and Gross 2022; Gross and Baron 2022; and summary in Bald et al. 2022a). Concurrent to our study, others have begun investigating the use of algorithmic decision-making tools to improve the processes to investigate or substantiate child maltreatment allegations (Putnam-Horenstien et

al. 2013; Grimon and Mills 2025). Most relevant is the ongoing research of Grimon and Mills (2025), who present results from an RCT in Larimer County, Colorado using an algorithmic decision-making tool that is similar to the DCDA and designed by the same team. They build on our findings by examining the effects of the tool on child hospitalizations. Some of their experimental design choices were different (e.g., randomization level, whether to require an initial decision to view the score from the tool, whether to include all referrals or just the first for each child), but Grimon and Mills (2025) also find no statistically significant effects on CPS outcomes. They do, however, find that access to the tool reduced injury-related hospitalizations by 29 percent and reduced disparities in child health outcomes and CPS surveillance. Importantly, although Grimon and Mills (2025) do not find statistically significant effects on the amount of time spent discussing referrals for investigation, their estimates are not statistically different from ours. Taken together, these studies suggest that algorithmic decision-making tools may improve the efficacy and speed of decision-making in the child welfare system.

We also contribute to the broader literature on the effectiveness of algorithmic decision-making tools on changing human behavior and downstream outcomes. In recent years, the use of algorithmic tools to improve decision-making has become ubiquitous in the private sector. With this, the potential of algorithmic tools to improve public government decision-making has also received much attention. To date, most of the research on the use of these tools in the public sphere has focused on the potential improvement (or not) that using these tools would have on outcomes of interest such as, recidivism, health, hiring, and bias (Kleinberg et al. 2018; Obermeyer et al. 2019; Arnold, Dobbie, and Hull 2021). Little research has explored how algorithms are used in practice by humans and the causal effects on actual outcomes of adopting algorithmic decision-making tools to inform decisions about government-related processes.[ii]

II. Background

Child maltreatment is a significant societal problem with a range of negative effects on children's health (Felitti et al. 1998; Bruce et al. 2009), educational attainment, and productivity, resulting in significant costs for society (Currie and Spatz Widom 2010; Currie and Tekin 2012; Fang et al. 2012). However, maltreatment is often under-detected, with many cases not receiving the intervention children need due to the high legal thresholds for substantiation (Palsuci and Vandervort 2014; Wildeman et al. 2014; Bullinger et al. 2021). At the same time, over-reporting is a concern, especially among children of color, as neglect and poverty are often confounded (Kim et al. 2017; Kim and Drake 2018; Fong 2019; Roberts 2002). These challenges are compounded by key issues in CPS including underfunding and understaffing, which lead to excessive caseloads (Hughes and Lay 2012; AFSCME 2016; Child Welfare Information Gateway 2016), high turnover rates (Kim and Kao 2014; Griffiths et al. 2017; Edwards and Wildeman 2018), and burnout, all of which are associated with poor decision-making (Kim 2011; Fluke et al. 2016). Given these challenges, a key question is whether algorithmic tools like the DCDA can help lessen burden by making decision-making faster while also improving child welfare outcomes in a cost-effective way. In this section, we outline the context of our study, including the child welfare system in which it takes place, the algorithmic decision-making tool used to assist caseworkers, and the randomized controlled trial designed to evaluate its impact.

A. Douglas County Child Welfare System Process

Our study setting is the DCDHS Child Welfare system. Like many agencies, DCDHS receives reports of suspected child maltreatment (also known as referrals), typically by phone,

where a screener takes basic information about the referral, including the nature of the allegation and background on the child(ren) involved. On an average day during the study period, DCDHS received 10 referrals. Referrals are entered into the system, then triaged for response based on the nature of the allegations and the ages of the children; response-time is 'immediate' or within three or five days. Our study focuses on the 93 percent of Child Protection reports that move to team review by caseworkers.[iii] Team reviews are held the day after the referral arrives, to determine whether the child(ren) should be 'screened-in' for investigation or not ('screened-out'). Once a child is screened-in, the referral is assigned to a caseworker who will add it to their caseload. The median caseworker in the United States will manage 55 cases per year (Edwards and Wildeman 2018).

For referrals that involve a team review to make the screen-in decision, DCDHS uses the RED (Read, Evaluate, and Direct) Team process. In this process, a team of caseworkers and supervisors reviews and discusses whether a referral should be screened-in for investigation.[iv] Prior to the development and use of the DCDA, the information reviewed by these RED Teams included information taken in by the screener at the referral point, as well as any additional information the RED Team accessed from a state database of all previous relevant child protection system interactions and Colorado Court records. This data provided information related to past child welfare system interactions, a child's age and gender, and active court cases. If certain factors are present in a referral (e.g., broken bones or other physical signs of abuse), then by law the referral must be investigated. Otherwise, the RED Team aims to use the information available to come to a consensus about whether to investigate a referral or not.

If a referral is investigated, it is assigned to a caseworker for investigation. An investigation involves direct contact with the family and others to gather information. If a referral is screened-out by the RED Team, that concludes the process of the referral.

B. The Predictive Risk Model: the Douglas County Decision Aid

DCDHS was interested in determining whether including predictive risk model information based on an array of data in the agency's decision-making process would improve decisions about investigating child maltreatment reports and change child and family trajectories. They worked with a separate external research team at the Centre for Social Data Analytics at the Auckland University of Technology to develop an algorithm trained to predict children's risk of experiencing removal from the home within 24 months of a referral (Vaithianathan et al. 2019a; Vaithianathan et al. 2019b).

The predictive risk score was developed using an augmented dataset containing linked historical child welfare records, public welfare program data, and case management records from the Client Benefits Management System (CBMS) from Denver Metro Area and Larimer County from January 2015 through September 2016. This dataset provided a broader picture of a child's history with CPS than what was typically available, including information on family ties and interactions with public assistance programs. Researchers at Auckland University of Technology used this enriched dataset to train and test a predictive model using LASSO regularized logistic regression. The model, which excluded Douglas County data during training, relied on 460 predictors—excluding race-specific measures—to generate predictive risk scores. Most of the model's predictive power comes from the child welfare agency related predictors. Researchers

building the model used standard procedures for building the model on a training dataset and tested its predictive power on a separate dataset (see Vaithianathan et al. (2019a) for details).[v]

Once the model was created, they created the DCDA as a way of presenting information from the predictive risk model to caseworkers involved in the decision-making process about whether to investigate. In Figure 1, we present a screen shot of what caseworkers see when using the DCDA. In its middle panel, the DCDA presents information on the placement rates by predicted risk score in the training sample. Lower-risk referrals, or those with scores below 9, have a less than 5 percent chance of placement within 24 months, while mid-range scores between 9 and 12 have slightly higher rates. Placement rates steadily increase for scores between 13 and 19 before jumping to about 45 percent for referrals with scores of 20. In Appendix B, we present information on the percentage of severe injuries, near fatalities, or fatalities by predicted risk score among the referrals that had previously been reported in the Denver Metro Area and Larimer County. This provides evidence that the risk score, which is calculated by predicting placement in foster care, also has predictive power for more objective, and serious, outcome measures of child maltreatment.

The historical relationship between predictions and investigation decisions suggests there may be room for improvement in the investigation decision and broader child welfare process of identifying true underlying maltreatment or risk thereof. In the testing sample, among referrals that have a predicted risk score of 1, 46 percent are screened in for investigation, but only 0.003 percent of those that are investigated are placed within 24 months. At the other end of the risk score spectrum, 43 percent of referrals with a score of 20 are not screened in for investigation, but 35 percent of those are children who are placed within 24 months. This suggests that

investigation decisions are missing a large percentage of the highest risk cases where there ends up being substantiated maltreatment serious enough for the child to be placed in foster care.

Before implementing the DCDA, all part- and full-time staff, child welfare administrators, and intake administrators underwent a 3-hour training designed to provide workers with an overview of predictive risk modeling, an explanation of how the DCDA tool was built, and practice using the DCDA in simulations. Staff were informed 1) the tool would act as an aid during their case discussions, 2) it would not reflect any information from the current referral (i.e., the tool is based only on historical information),[vi] and 3) they would make the final determination, regardless of the score. After the training, staff underwent a four-week trial period where they experimented with discussing the score before and after their initial decisions.

C. The Randomized Control Trial

In collaboration with the DCDHS staff, we designed an RCT to evaluate whether the use of the DCDA changed decisions about whether to investigate a referral alleging maltreatment and the amount of time it took to make those decisions, as well as whether there were subsequent effects on longer run outcomes for children and families (e.g., placement into foster care). The goal was to create an RCT that would allow us to determine the causal effect of using the DCDA tool in a process that closely resembled existing practices and made use of the tool to supplement the existing information utilized and the experienced decision-making of the caseworkers.

Screeners randomly assigned referrals to RED Teams, then used a random number generator to determine which RED Team was in the treatment group. One RED Team on each day in the experimental period was assigned randomly to the control group. The other one or two teams each day were in the treatment group.[vii] [viii]

RED Teams assigned to the treatment group had access to all the same information as those in the control group plus the DCDA score. In their meetings, the treatment group RED Teams were instructed to first conduct their review and discussion of the prior history and current allegation of a referral with the aim of deciding whether to investigate without viewing the DCDA. Then, they were to make an initial recommendation about whether to screen-in the referral for investigation. After the initial recommendation decision, the supervisor was to access the DCDA score and share it with the entire RED Team. The RED Team then discussed the score in the context of the previous information and decided whether to change the initial decision based on the DCDA score.[ix] Scores were sometimes unavailable due to system outages or late-evening referrals; supervisors noted whether the score was visible on their tracking sheets.

The goal in designing the setting this way – a choice by the DCDHS – was to ensure that the DCDA tool was used in addition to, rather than instead of, existing processes and expertise. However, this setting design may mean our experiment is not an accurate estimate of the potential of tools like the DCDA to change decisions or the time it takes to make them if the tool is used at the beginning of the discussion. It will be an underestimate if having access to the information earlier would speed up decision-making (e.g., if knowing a referral was low scoring would prompt a quicker decision not to investigate). It will be an overestimate if knowing they can revise the preliminary decision after viewing the score leads to quicker decisions. Many counties are likely to have similar reasons to Douglas for preferring a process that involves a preliminary decision before viewing the score, so the parameter we estimate is policy relevant.

In the text below, we describe results of the experiment run from its launch on March 5, 2019 through March 4, 2020.[x] It is important to highlight that although our study period stops right as the global coronavirus pandemic began in the United States, some of the outcome period

may have been significantly altered by the pandemic since we evaluate the impact of the DCDA on some outcomes of children within one year of a referral. This is particularly true for outcomes that take time to occur, like removal. As we describe in more detail in Section III, to help deal with this issue, we measure outcomes that materialize quickly using a 30-day window and outcomes that take longer to occur, like removal from the home, in 365-day windows. The former will be less likely to have been affected by the COVID-19 pandemic than the latter.

III. Data

Our main source of data is Colorado's statewide information system on child welfare, SACWIS, locally known as Trails.[xi] The SACWIS system is used to record information from the time a referral is reported through investigation, placement, or services provision. It includes information on the referral, including referral date as well as basic child and allegation information, information on the decision to investigate (or screen-in) a referral, whether a case was opened because an investigation was substantiated or due to an assessed need for services, and information on removal from the home (or placement in foster care). Children are uniquely identified in the data and can be referred by multiple entities on the same day. Each new referral is given a separate identifier. Therefore, children can be followed over time, both within the process of a particular referral and across separate referrals. There can be multiple children on any given referral in the Trails data, each of whom may receive a different DCDA score, a point we return to in our analyses.

We also utilize the tracking sheet data that were created as part of our study. As described above, RED Team supervisors were instructed to use these tracking sheets to record information about which referrals were in a treatment group or control group RED Team, as well as whether

a score was seen or not in both the treatment and control groups. They were also instructed to record information about the start and stop time of the discussion of each referral. Unfortunately, adherence to the tracking sheets was imperfect. Some tracking sheets were never filled out or were lost in the process. Since the tracking sheets were the lone record of whether a referral was assigned to a treatment or control team, we do not include these referrals in our main analysis.

Our final source of data is the score included in the DCDA and created by the predicted risk model. When a referral comes in, it must be matched to the child's history in the SACWIS system before a DCDA score can be produced. However, matches between referrals and SACWIS are only run once a day in the evening. Once the daily referrals are matched, the predictive risk model is run and the DCDA score is added as a point of information in the system available to RED Teams. This means a score was not created in real-time for referrals that came in overnight or those that came in when the system was down for idiosyncratic reasons. Since these referrals without scores could not be viewed, we do not include them in our analysis.[xii]

Combining these data sources, our analysis sample is at the child-by-referral level. Referrals related to an investigation or existing substantiated case were sent to the caseworker currently involved, so we focus our sample on the set of referrals during our experiment period that were not related to an ongoing investigation or substantiated case. There were 5,872 child-referrals from March 5th, 2019 to March 4th, 2020 involving Child Protection referrals without open cases that went to a RED Team for review. We drop the 0.7 percent of these that required immediate attention or were not reviewed by a RED Team, leaving a sample of 5,831. Also, as mentioned, we drop an additional 6 percent of child-referrals during this period that were not recorded on tracking sheets and the 21 percent that did not have real-time scores. Our main analysis sample therefore includes 4,346 observations at the child-by-referral level.[xiii]

Our primary outcomes are the time spent discussing a referral and whether a referral was investigated, as well as whether the child was part of any referral that was substantiated within 365 days, whether the child was re-referred to the system within 30 or 365 days, and whether the child was removed from the home (placed in foster care) within 30 or 365 days.[xiv] We include measures of rereferral and placement within both 30 days and 365 days for two reasons. First, these are outcome measures that can take time to occur (hence the measure at 365 days). Second, we are also concerned that the pandemic may have changed the reporting and child welfare processes in ways that affect outcomes measured well into 2020 or 2021, so we include measures of these outcomes 30 days from the receipt of the referral. Our key independent variable is an indicator for whether the RED Team was assigned to treatment.

Douglas County is a wealthy, highly educated, majority-white county with 368,990 residents, 22 percent of whom are under 18, and is located on the outskirts of Denver, CO. The average median income is $121,393 which is nearly double the national average of $64,994. Nearly all Douglas County residents have graduated from high school (98 percent), while over half of all residents hold a bachelor's degree (59 percent). These rates of education are ten and twenty-six percentage points higher than the national average, respectively. Ninety percent of residents are white compared to 76 percent nationally.

That this is an affluent county may make the results from this RCT less generalizable. For example, in affluent counties there are fewer child welfare reports and less interaction with the juvenile justice system. Since the DCDA score is heavily reliant on historical interactions with the child welfare system, including their interactions with the juvenile justice system, it may be less accurate for children with limited histories. Our discussions with caseworkers revealed that they understood this intuitively and may therefore have been less likely to revise their

preliminary investigation decisions if the score was inconsistent with their initial assessment for referrals where the child did not have a prior history with the child welfare system and/or juvenile justice system. In contrast, in a less affluent county, a low score due to low child welfare and juvenile justice contact is more informative because both are more frequent than in more affluent counties. In another example, child welfare agencies in affluent counties may be less overburdened or face different structural or institutional constraints, such as higher rates of legal involvement, which may alter how caseworkers utilize the tool when it is available, perhaps changing its influence on time to decision and on child welfare outcomes.

A. *Validation of the Randomization Process*

To determine whether our randomization process was successful at creating balanced samples in the treatment and control group, in Table 1, we present average characteristics of the treatment and control groups separately, as well as p-values for a t-test of whether the characteristics are the same across the two groups.[xv] We do not find a statistically significant difference between the means of most of these characteristics across the control and treatment groups, including on the risk score calculated by the DCDA.[xvi] The exception is that there are statistically significant differences in the types of reporters that make referrals across groups. Referrals in the control group were slightly more likely to come from medical professionals or family, friends, and neighbors, while those in the treatment group were slightly more likely to come from social services agency related reporters. The differences are small, but to assuage concerns that differences between the treatment and control groups are driving any estimated effects, in some specifications we control for these characteristics to test robustness.

IV. Methods

We use traditional methods for an RCT setting to evaluate the effect of the DCDA tool availability on the outcomes of interest. The relationship between the DCDA tool availability and the outcomes of interest can be summarized by the following equation:

$$Y_{crd} = \alpha + \beta Treat_{rd} + \rho_d + \varepsilon_{crd}.$$

Here, *Y* is the outcome of interest for child *c* on referral *r* that was discussed by a RED Team on day *d*. *Treat* is an indicator for whether the referral was part of a RED Team that was assigned to view the DCDA. Randomization was conducted at the day-level, so we include fixed effects for each RED Team day, noted by ρ_d. These control for any seasonal or day-of-the-week variation in the types of cases seen at any time during our experiment.[xvii] We cluster the standard errors at the level of the RED Team day to control for any serial correlation in the error term across referrals reviewed by the same RED Team.[xviii] We interpret β as the effect of the score being available on investigation decisions.

Another estimate of potential interest in this setting is the effect of viewing the DCDA on investigative decisions. In our RCT setting, β would translate into the causal effect of seeing the score if everyone assigned to the treatment group looked up the score and the score was always available. However, we know that some scores were not seen because the DCDA system was down for maintenance, or the report came in after the DCDA model was run for the day. There also may have been user infidelity. In about 88 percent of treated child-referrals where a score was available caseworkers confirmed they viewed the score. Because user error and system-related error are common in the implementation of any algorithmic decision-making tool, the Intent-to-Treat (ITT) estimate, β, may be closest to what another county would experience if

they implemented a similar program. Because it is also of interest, below, we present an instrumental variables estimate of the effect of seeing the DCDA on investigative decisions.

V. Results

In columns 1 and 2 of Table 2, we present estimates of the ITT effect of the team being assigned to view the DCDA. In the table, each set of rows presents coefficient estimates and standard errors from a set of regressions with a different dependent variable: investigation, time spent in discussion, substantiation, re-referral at 30 days, re-referral at 365 days, placement into foster care at 30 days, and placement into foster care at 365 days. In the Table, the first column for each outcome presents the estimates from estimating equation (1). In the second column, we report the coefficient estimates of β from a specification where we add the set of controls presented as part of our balance tests in Table 1.[xix] In the third and fourth columns, we also present the results from an instrumental variables (IV) regression, with and without controls, respectively. In the first stage, the random assignment to view the score is an instrument for viewing the score. The estimated coefficients on viewing the DCDA represent the local average treatment (LATE) effect of viewing the DCDA on a given outcome for referrals where viewing the score was influenced by the assignment.

The results in Table 2 show that being assigned to see the DCDA increased the probability of investigation by about 1.9 percentage points (5 percent).[xx] However, the estimate is not statistically significant at conventional levels. Similarly, none of the other child welfare system outcomes are statistically significant at conventional levels. Because the impact of the tool may have been different for children with different levels of risk, below we explore whether these treatment effects are different across children with different levels of risk.

There is one aspect of RED Team decisions where we detect a statistically significant change because of the DCDA: the number of minutes spent discussing a referral. Because discussion happens for all children on the referral at the same time, these results are run at the referral level (rather than the child-by-referral level as for the other outcomes). The coefficient estimate, which is statistically significant at the 10 percent level, suggests that RED Teams spent about half a minute less discussing the referral. Relative to an average discussion time of 11 minutes, this represents a 5 percent decrease. Recall that the RED Teams were not allowed to look at the score before reaching an initial recommendation decision about whether to investigate. This means that just knowing that they would access the score after their initial discussion decreased the amount of time discussing referrals. Given that there are at least three employees per RED Team and that there were 3,296 referrals within our one-year sample period, this translates into 76 hours, or 2 full-time-equivalent weeks, of worker time over the year.

If the DCDA tool provides a more accurate signal of the likelihood of underlying child maltreatment than the information previously available to RED Teams, then we would hypothesize that referrals with lower scores would be less likely to get investigated, while higher risk cases would be more likely to get investigated. However, if RED Team workers place more emphasis on avoiding type II errors (deciding not to investigate a true instance of maltreatment) than on type I errors (investigating when there is not true maltreatment), then the relationship between the use and availability of the DCDA and the investigation decision may be strictly positive. This is even more likely to be the case in our setting where RED Teams were instructed to make a preliminary decision before viewing the DCDA and only then incorporate the DCDA into their final decision. In this setup, treatment RED Teams deciding to investigate before seeing the score made that decision based on the evidence available in the referral and may have

thought that evidence was compelling enough to merit investigation even when the DCDA risk score was low.

To investigate whether this was the case, we examined the relationship between the risk score of a referral, the use of the DCDA tool, and the outcomes of interest. We categorized risk scores by their quartile of the overall risk score. The bins are as follows: scores between 1 to 5, 6 to 10, 11 to 15, and 16 to 20.[xxi] We added to our main specification controls for dummy variables for each of these bins and the interaction between these bin variables and our measure of whether a referral was part of the treatment group. (For ease of interpretation, we fully saturated the model with these dummy variable interactions, so there is no omitted group.) We present the coefficients and related confidence intervals for each of these outcome variables in Figure 2. In each panel of the figure, the light gray bars present estimates of the average likelihood of each outcome for the control group of a given risk score bin and the dark gray bars present estimates for the treatment group.

Although only one of the coefficients (that for time spent in discussion in Panel B) is statistically significant at conventional levels, some interesting patterns emerge from the point estimates. In Panel A, we see that the availability of the DCDA increases the likelihood of investigations for the highest risk referrals but has little effect on the likelihood of investigation for the lowest risk referrals. We interpret this as evidence that RED Teams were unlikely to reverse their preliminary decisions to investigate referrals because of the signal of low risk from the DCDA score. This may be because they weight type II errors more heavily than type I errors, there was compelling information in the referral not captured in the risk score, or both.

In Appendix D we present results for other key outcomes. Panel A, the increase in the likelihood of a substantiation on any referral within the following year is positive for referrals

with scores between 6 and 10 and negative for referrals with higher risk scores. In Panel B, across all risk scores, except the lowest, rates of referral fall. And, in Panel C, risks of placement in foster care increase for referrals with risk scores between 6 and 10 but fall for higher risk scores. In fact, for referrals with the highest risk scores, rates of placement fall by 70 percent of the control group level. Although we take caution in making conclusions based on these underpowered results, they may be suggestive that the DCDA tool may accelerate identification of maltreatment in high-risk environments in ways that allow for earlier intervention and the prevention of further involvement with the child welfare system.

A. Are Patterns Different by the Quality of the Original Information or Risk of Type II Error?

To further investigate the role of the information from the DCDA tool in decisions to investigate, we now examine how the relationship between the DCDA score and investigation decisions varied with DCDA availability based on characteristics that workers may have previously considered to be proxies for risk of maltreatment. Conversations with child welfare professionals suggested to us that 1) referrals by law enforcement officials are considered more reliable than those by teachers and other education professionals and 2) referrals for children under 2 were considered more serious. If this is the case, we should see different investigation propensities for these different types of reporters and referral characteristics. Appendix E and F show point estimates that suggest the DCDA leads to 1) more screen in of high-risk referrals not typically considered as risky (those from school-based and other reporters), 2) less screen in of low-risk referrals that are generally considered riskier (those from law enforcement and medical professionals, 3) more screen in of older children with riskier scores, and 4) less screen in of

younger children with less risky scores. However, these results should be interpreted with caution as our experiment is underpowered to detect statistical differences.

VI. Conclusion

We conducted one of the first RCTs of the use of algorithmic decision-making tools based on predicted risk models in the child welfare process. These tools have the potential to change decisions to i) improve child outcomes, ii) reduce bias in the process, and/or iii) speed up decision-making, freeing up valuable staff time for other activities to support child wellbeing.

Despite our experimental design requiring teams of caseworkers to make a preliminary decision on a referral before viewing a risk score, which may have served to increase time spent on decisions about child maltreatment investigations, we find that agency staff worked faster in the presence of the risk score than in the absence of the risk score. In a system as overburdened and underfunded as the child welfare system, the ability of algorithmic decision-making tools to provide agency workers with time for supporting families may be incredibly valuable.

Although designed to examine whether the algorithmic decision-making tool affected child welfare outcomes, our experiment was underpowered to detect any effects. However, we find some suggestive evidence that child welfare agency workers may be risk averse. This suggests algorithmic tools may have the ability to reduce type II errors.

In part, our inability to detect meaningful differences was because the pandemic interrupted the experiment and significantly altered child welfare outcomes measured within the experiment. It also may be the case that there is little effect because caseworkers placed limited weight on the tool, especially after having come to an initial decision. Also, because the setting of our study is a homogeneous county and because measures of race and ethnicity in child

welfare data are unreliable,[xxii] we are unable to speak to whether the use of algorithmic decision-making tools affects bias in the process or related outcomes for children from different backgrounds. This is an important area for future research. Important related research by others (Grimon and Mills 2025) has found a similar tool in a similar setting decreased injury-related hospitalizations and disparities in these hospitalizations and in CPS surveillance. Taken together, these studies imply that algorithmic decision-making tools have the potential to lessen the burden on CPS and ensure more at-risk children are identified and receive necessary intervention.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffery M. Wooldridge. 2023. “When Should You Adjust Standard Errors for Clustering?” *The Quarterly Journal of Economics* 138(1): 1-35.

AFSCME. 2016. “Cost savings from reasonable child welfare workloads.” Washington, DC: AFSCME Department of Research and Collective Bargaining Services. Retrieved May 16, 2025, from https://ncwwi.org/files/Job_Analysis__Position_Requirements/Cost_Savings_from_Reasonable_Child_Welfare_Workloads.pdf.

Albright, Alex. 2019. “If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions.” Working Paper. Retrieved May 16, 2025, from https://thelittledataset.com/about_files/albright_judge_score.pdf.

Arnold, David, Will Dobbie, and Peter Hull. 2021. “Measuring racial discrimination in algorithms.” *AEA Papers and Proceedings* 111: 49-54.

Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett. 2022a. “The Causal Impact of Removing Children from Abusive and Neglectful Homes.” *Journal of Political Economy* 130(7): 1717-1991.

Bald, Anthony, Joseph J. Doyle Jr., Max Gross, and Brian A. Jacob. 2022b. "Economics of Foster Care." *Journal of Economic Perspectives* 36(2): 223-246.

Baron, E. Jason, Ezra G. Goldstein, Cullen T. Wallace. 2020. “Suffering in silence: How COVID-19 school closures inhibit the reporting of child maltreatment.” *Journal of Public Economics* 190: 104258.

Baron, E. Jason, and Max Gross. 2022. “Is There a Foster Care-to-Prison Pipeline? Evidence from Quasi-Randomly Assigned Investigators.” *NBER* Working Paper No. 29922.

Brown, Samantha M., Rebecca Orsi, Pang Ching Bobby Chen, Courtney L. Everson, and John Fluke. (2022). “The Impact of the COVID-19 Pandemic on Child Protection System Referrals and Responses in Colorado, USA.” *Child Maltreatment* 27(1): 3-11.

Bruce, Jacqueline, Philip A. Fisher, Katherine C. Pears, and Seymour Levine. 2009. “Morning cortisol Levels in preschool-aged foster children: Differential effects of maltreatment type.” *Developmental Psychobiology* 51(1): 14-23.

Bullinger, Lindsey Rose, Kerri M. Raissian, Megan Feely, and William J. Schneider. 2021. “The neglected ones: Time at home during COVID-19 and child maltreatment.” *Children and Youth Services Review* 131: 106287.

Cameron, A. Colin, and Douglas L. Miller. 2015. “A practitioner’s guide to cluster-robust inference.” *Journal of Human Resources* 50(2): 317–372.

Child Welfare Information Gateway. 2016. “Caseload and workload management.” Washington, DC: Children’s Bureau, Issue Brief. Retrieved May 16, 2025, from https://ncwwi.org/files/Job_Analysis__Position_Requirements/case_work_management.pdf.

Clarke, Damian, Joseph P. Romano, and Michael Wolf. 2020. "The Romano-Wolf Multiple Hypothesis Correction in Stata." *The Stata Journal* 20(4): 812-843.

Currie, Janet, and Cathy Spatz Widom. 2010. "Long-Term Consequences of Child Abuse and Neglect on Adult Economic Well-Being." *Child Maltreatment* 15(2): 111-120.

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

Doyle, Joseph J. 2007 "Child Protection and Child Outcomes: Measuring the Effects of Foster Care." *American Economic Review* 97(5): 1583-1610.

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

_____. 2013. "Causal Effects of Foster Care: An Instrumental-Variables Approach." *Children and Youth Services Review* 35(7): 1143–51.

Edwards, Frank, and Christopher Wildeman. 2018. "Characteristics of the front-line child welfare workforce." *Children and Youth Services Review* 89: 13-26.

Fang, Xiangming, Derek S. Brown, Curtis S. Florence, and James A. Mercy. 2012. "The economic burden of child maltreatment in the United States and implications for prevention." *Child Abuse & Neglect* 36(2): 156-165.

Felitti, Vincent J., Robert F. Anda, Dale Nordenberg, David F. Williamson, Allison M. Spitz, Valerie Edwards, Mary P. Koss, and James S. Marks. 1998. "Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study." *American Journal of Preventive Medicine* 14(4): 245-258.

Finkelhor David, Heather A. Turner, Anne Shattuck, and Sherry L. Hamby. 2013. "Violence, crime, and abuse exposure in a national sample of children and youth: an update." *JAMA Pediatrics* 167(7): 614-21.

Fluke, John D., Tyler W. Corwin, Dana M. Hollinshead, and Erin J. Maher. 2016. "Family preservation or child safety? Associations between child welfare workers' experience, position, and perspectives." *Children and Youth Services Review* 69: 210-218.

Fong, Kelley. 2019. "Neighborhood inequality in the prevalence of reported and substantiated child maltreatment." *Child Abuse & Neglect* 90: 13-21.

Griffiths, Austin, David Royse, Kaylee Culver, Kristine Piescher, and Yanchen Zhang. 2017. "Who stays, who goes, who knows? A state-wide survey of child welfare workers." *Children and Youth Services Review* 77: 110-117.

Grimon, Marie-Pascale, and Christopher Mills. 2025. "Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection." Working paper. Retrieved May 16, 2025, from https://www.dropbox.com/scl/fi/cmwjmik400s8sqbp7dwct/GrimonMills.pdf?rlkey=l9la4ede9ww5v01ha30393r48&e=1&st=8d62ld3t&dl=0.

Gross, Max, and E. Jason Baron. 2022. “Temporary Stays and Persistent Gains: The Causal Effects of Foster Care.” *American Economic Journal: Applied Economics* 14(2): 170-199.

Hughes, Sean, and Suzanne Lay. 2012. “Direct services workers’ recommendations for child welfare financing and system reform.” Washington, DC: Child Welfare League of America. Retrived May 16, 2025, from https://www.cwla.org/wp-content/uploads/2014/05/DirectServiceWEB.pdf.

Kim, Hansung. 2011. “Job conditions, unmet expectations, and burnout in public child welfare workers: How different from other social workers?” *Children and Youth Services Review* 33(2): 358–367.

Kim, Hyunil, and Brett Drake. 2018. “Child maltreatment risk as a function of poverty and race/ethnicity in the USA.” *International Journal of Epidemiology* 47(3): 780-787.

Kim, Hyosu, and Dennis Kao. 2014. “A meta-analysis of turnover intention predictors among U.S. child welfare workers.” *Children and Youth Services Review* 47(3): 214-223.

Kim, Hyunil, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. 2017. “Lifetime Prevalence of Investigating Child Maltreatment Among US Children.” *American Journal of Public Health* 107(2): 274–280.

Kleinberg, John, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133(1): 237–293.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Ragavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting racial bias in an algorithm used to manage the health of populations.” *Science* 366(6464): 447-453.

Palusci, Vincent J., and Frank E. Vandervort. 2014. “Universal reporting laws and child maltreatment report rates in large U.S. counties.” *Children and Youth Services Review* 38: 20-28.

Putnam-Hornstein, Emily, Barbara Needell, Bryn King, and Michelle Johnson-Motoyama. 2013. “Racial and Ethnic Disparities: A Population-Based Examination of Risk Factors for Involvement with Child Protective Services.” *Child Abuse & Neglect* 37(1): 33–46.

Renov, Veronica, Lauren Risser, Rachel Berger, Tammy Hurley, Andres Villaveces, Sarah DeGue, Abigail Katz, Cynterria Henderson, Kelly Premo, Janine Talis, Judy C. Chang, and Maya Ragavan. 2022. “The impact of the COVID-19 pandemic on child protective services caseworkers and administrators.” *Child Abuse & Neglect* 130: 105431.

Roberts, Dorothy. 2002. “Shattered Bonds: The Color of Child Welfare.” *Civitas Books*.

Roberts, Kelsey V. 2019. “Foster Care and Child Welfare.” PhD dissertation Clemson University. Retrieved May 16, 2025, from https://open.clemson.edu/all_dissertations/2375/.

Stevenson, Megan T., and Jennifer L. Doleac. 2024. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy* 16(4): 382-414.

U.S. Department of Health and Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. 2021. "Child Maltreatment 2019." Retrieved May 16, 2025, from https://acf.gov/cb/report/child-maltreatment-2019.

Vaithianathan, Rhema, Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur, Athena Ning, and Diana Benavides Prado. 2019a. "Implementing a Child Welfare Decision Aide in Douglas County Methodology Report." *Centre for Social Data Analytics Report.* Retrieved May 16, 2025, from https://csda.aut.ac.nz/__data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf.

Vaithianathan, Rhema, Emily Putnam-Hornstein, Matthew Stagner, and Elizabeth Weigensberg. 2019b. "Predictive Risk Modeling for Child Protection." Washington, DC: Mathematica Policy Research. Retrieved February 9, 2025, from https://www.mathematica.org/publications/predictive-risk-modeling-for-child-protection.

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

Weigensberg, Elizabeth, Nuzhat Islam, Jean Knab, Mary Grider, Jeremy Page, and Sarah Bardin. 2022. "State Child Abuse and Neglect (SCAN) Policies Database Codebook 2019 Version 2." Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved May 16, 2025 from https://www.ndacan.acf.hhs.gov/datasets/pdfs_user_guides/dataset245-scan-2019-codebook.pdf.

Wildeman, Christopher, Natalia Emanuel, John M. Leventhal, Emily Putnam-Hornstein, Jane Waldfogel, and Hedwig Lee. 2014. "The Prevalence of Confirmed Maltreatment Among US Children, 2004 to 2011." *JAMA Pediatrics* 168(8): 706–713.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Inference of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134(2): 557-598.

Table 1

Baseline Characteristics by Treatment Group

	Control		Treatment		
	Mean	Standard Deviation	Mean	Standard Deviation	p-value
DCDA Score - Child-Level	6.67	5.93	6.64	5.85	0.86
Age	9.92	5.06	9.95	5.30	0.84
Race: Hispanic	0.07	0.26	0.07	0.25	0.61
Race: Black	0.05	0.21	0.04	0.20	0.33
Reason: Physical Abuse	0.36	0.48	0.33	0.47	0.16
Reason: Sexual Abuse	0.07	0.26	0.08	0.27	0.66
Reason: Domestic Violence	0.17	0.37	0.18	0.38	0.25
Reason: Neglect	0.17	0.38	0.19	0.39	0.14
Reason: Other	0.39	0.49	0.39	0.49	0.95
Reporter Type: School	0.57	0.49	0.58	0.49	0.68
Reporter Type: Law Enforcement	0.10	0.30	0.12	0.32	0.16
Reporter Type: Medical professionals	0.13	0.34	0.10	0.30	0.01
Reporter Type: Social Services related	0.08	0.28	0.10	0.30	0.04
Reporter Type: Family, Friend, Neighbor	0.04	0.18	0.02	0.14	0.00
Reporter Type: Other	0.08	0.27	0.08	0.27	0.73
Number of Observations	1,701		2,645		

Note: Based on the authors' calculations using administrative data from Douglas County. Sample includes referrals assigned to RED Teams between March 5, 2019 and March 4, 2020. The

treatment of viewing the DCDA score was randomly assigned to RED Teams. Mean characteristics of observations at the child by referral level are presented in the table along with standard deviations. The last column contains p-values for a test of the sample means being identical. The sample includes only referrals where the real time score was available.

Table 2

Effects of Assignment to Treatment to View the Douglas County Decision Aide on Caseworker's Decision to Intervene in Child Maltreatment Referrals

	(1)	(2)	(3)	(4)	(5)
	ITT Effect	ITT Effect	IV Effect	IV Effect	Mean Outcome in Control Group
Investigation	0.019	0.017	0.022	0.019	0.371
	(0.025)	(0.022)	(0.028)	(0.025)	
Time Spent (minutes)	-0.503*	-0.517*	-0.574*	-0.592*	11.224
(referral level)	(0.272)	(0.267)	(0.310)	(0.306)	
Substantiation	0.002	0.001	0.002	0.002	0.089
	(0.013)	(0.012)	(0.015)	(0.014)	
Re-referral within 30 days	0.016	0.023	0.019	0.026	0.164
	(0.020)	(0.019)	(0.023)	(0.022)	
Re-referral within 365 days	-0.016	-0.011	-0.018	-0.012	0.483
	(0.025)	(0.022)	(0.029)	(0.025)	
Placement within 30 days	0.001	0.002	0.001	0.002	0.003
	(0.002)	(0.003)	(0.002)	(0.003)	
Placement within 365 days	-0.001	-0.001	-0.002	-0.001	0.012
	(0.005)	(0.005)	(0.006)	(0.006)	
Controls Included	No	Yes	No	Yes	

Number of child-referrals	4,346
Number of referrals	2,073

Note: Based on the authors' calculations using administrative data from Douglas County. Sample includes referrals assigned to RED Teams between March 5, 2019 and March 4, 2020. The measure of interest is whether a referral was assigned to the treatment group for viewing the DCDA score. The treatment of viewing the DCDA score was randomly assigned to RED Teams based on their meeting date, so we have included calendar day fixed effects for the RED Team meetings. Columns (1) and (2) show the intent-to-treat effect where the independent variable is whether a team was assigned to treatment. Twelve percent of teams assigned to treatment did not see the score. Therefore, columns (3) and (4) show a local average treatment effect from an instrumental variables approach where the first stage predicts treatment status by whether the score was seen by the team. The dependent variable is a dummy variable for (1) the referral leading to an investigation (screen-in), (2) the amount of time the RED Team spent discussing the referral, (3) the child being involved in a substantiated case opening within 365 days, (4 and 5) the child being involved in another referral within 30 or 365 days, or (6 and 7) the child being removed from the home (placed in foster care) within 30 or 365 days. Standard errors are clustered at the level of the day of the RED Team meeting. Significance stars correspond to *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

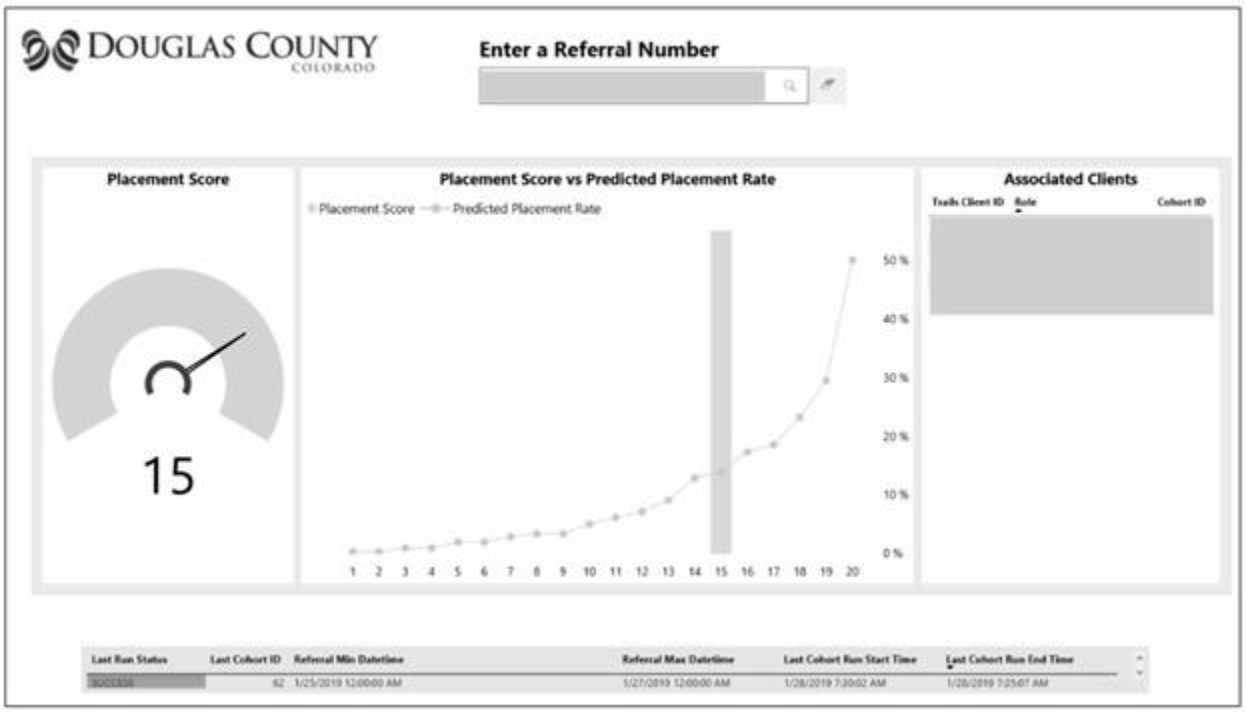

Figure 1

Douglas County Decision Aid Tool

Notes: The above image is a black and white screen shot of the Douglas County Decision Aid tool as seen by RED Team members during meetings to discuss whether to open an investigation. Information about the child and referral has been redacted. On the left-hand-side of the screen is the score with a dial representing its place in the range of possible scores. In the middle of the screen is a figure showing the placement score and the related predicted placement rate at that score relative to predicted placement rates throughout the score distribution.

Panel A. Panel B.

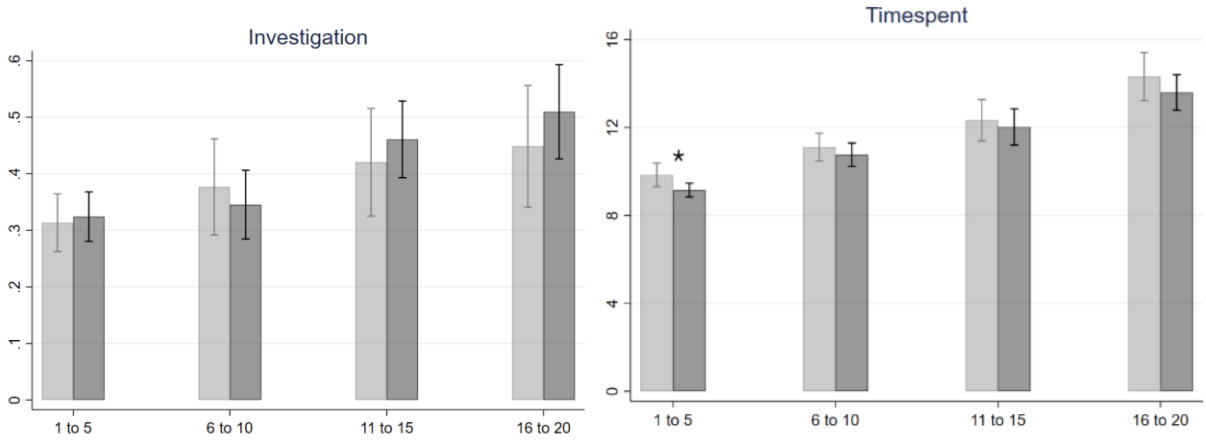

Figure 2

Effects of Assignment to Treatment to View the Douglas County Decision Aide on Caseworker's Decision to Intervene in Child Maltreatment Referrals, by Risk Score Category

Notes: Based on the authors' calculations using administrative data from Douglas County. Sample includes referrals assigned to RED Teams between March 5, 2019 and March 4, 2020. The measure of interest is whether a referral was assigned to the treatment group for viewing the DCDA score, with light gray representing the control group and dark gray representing the treated group. The x-axis shows risk score bins, and the y-axis shows the effect size. The treatment of viewing the DCDA score was randomly assigned to RED Teams based on their meeting date, so we have included calendar day fixed effects for the RED Team meetings. The dependent variable is a dummy variable for (A) the referral leading to an investigation (screen-in) and (B) the time spent by caseworkers discussing the referral. Standard errors are clustered at the level of the day of the RED Team meeting. Significance stars correspond to *** $p<0.01$, ** $p<0.05$, * $p<0.1$. 95% confidence interval bars are displayed.

[i] As detailed below, certain case types bypass this process. In Douglas County, like 13 percent of CPS offices nationwide, caseworker teams decide whether to investigate reports (Weigensberg et al. 2022). The tool may have a different impact in settings where individuals, rather than teams, make decisions about whether to investigate referrals. For example, overworked individual caseworkers may become overly reliant on the predictive risk score and treat is as definitive. This could reduce the accuracy of assessments, as they might underweight important context or details that a group of people would discuss. The results from this study may not generalize to a context where individual caseworkers are provided the tool in their decision-making process.
[ii] Exceptions in criminal justice are studies by Albright (2019) and Stevenson and Doleac (2024) of the use of minimalist risk assessment modeling as part of the process for judicial sentencing of individuals convicted of felonies and for bail decisions. There are no exceptions that we are aware of in child welfare.
[iii] 86 percent of our referrals are categorized as Child Protection (PA-5), 11 percent are Youth-in-Conflict (PA-4), and 3 percent are handled through other processes (See also Appendix A).
[iv] Each RED Team is comprised of at least three members: one supervisor and two caseworkers. Each of the 12 supervisors and 32 caseworkers sign up for seven and five RED Teams per month, respectively. Each RED Team is composed of a different combination of caseworkers.
[v] The modelling team took several steps to verify the fairness of the model, particularly given known potential bias in reporting, investigation, and intervention. They tested the model's functionality with and without measures for race and ethnicity, ultimately leaving these measures out of the model. They also present information that the model is "well-calibrated" (see especially Kleinberg, Mullainathan, and Ragavan 2017) by showing that children with the same risk scores but different racial backgrounds, ages, and genders have similar rates of being placed within two years.
[vi] Since the time of this experiment, the DCDA tool has been updated to incorporate current referral information.
[vii] We considered randomizing each referral to be in the treatment or control group. However, one concern we had in setting up this experiment was learning-by-doing on the part of RED teams. In other words, we were worried that caseworkers would create their own mental model versions of the predictive risk model (based on the scores they saw) as their exposure to the score grew. This would have happened more quickly if scores were visible in every RED Team meeting. Additionally, we decided that a setting where all referrals in a RED Team meeting had scores visible or did not more appropriately mirror the comparison between two work environments where caseworkers either are or are not able to access scores for all referrals.
[viii] Typically, caseworkers can be assigned as investigators on referrals that are screened-in from RED Teams they were a part of. During the experiment, however, caseworkers on a RED Team that reviewed a referral were not assigned to investigate those referrals. This was done to prevent information about a referral's score being used to determine the outcome of the investigation.
[ix] If a RED Team decides to screen out a case with a score of 18 or higher, the case is reviewed by an additional supervisor (not the supervisor that served on the relevant RED Team).
[x] The experiment was scheduled to run for one year but was extended through April 30, 2020 to increase power. Soon after, the pandemic led to widespread shutdowns and a massive shift in maltreatment allegations and assessment processes. Reports of child maltreatment decreased as government-mandated stay-at-home orders began in March 2020 (e.g., Baron, Goldstein, and

Wallace 2020, Brown et al. 2022). Although investigation guidelines and criteria for substantiation did not change, child welfare system processes adjusted to allow functions remotely, when possible, or adjust to the safety requirements, if in-person (e.g., Renov et al. 2022). These changes may have affected the collection of relevant information for a substantiation decision. We have examined the results for the entire study period. There are some differences in the size of point estimates and statistical power, but overall, the estimates from the full sample are similar to results herein.

[xi] SACWIS is the same CPS data used to create the DCDA, though we only have access to the information on referrals during our period, not the historical data used to create the DCDA.

[xii] The likelihood of a real-time score existing is 2 percentage points higher in the treatment group than the control group. This relationship is statistically significant with a p-value of 0.086.

[xiii] Results are similar when we include the referrals not on tracking sheets (and therefore missing information on treatment status) and code the referrals as either all in the treatment group or all in the control group.

[xiv] The literature examining the effects of foster care placement on children has produced mixed results, likely because of different institutional contexts and different types of families with differing levels or types of maltreatment (Doyle 2007; Doyle 2008; Doyle 2013; Warburton et al. 2014; Roberts 2019; Bald et al. 2022a; Bald et al. 2022b; Baron and Gross 2022; Gross and Baron 2022).

[xv] Because the score was not viewed for 12% of the referrals assigned to treatment, we also examine baseline balance between treated referrals where the score was viewed and those where it was not in Appendix C. We find meaningful differences for roughly half (8 of the 18) of the referral characteristics.

[xvi] The average DCDA score in the control and treatment groups is just over 6, rather than at the middle of the risk score distribution (10). This is because the DCDA score is predicted using data on all the Denver Metro area while our sample only includes Douglas County, which has a population with a lower risk of future system involvement. It may also be driven by the fact that the highest risk cases in Douglas are not assigned to the RED Team.

[xvii] These fixed effects make little difference to the qualitative conclusions herein. Results available upon request.

[xviii] Abadie et al. (2023) suggest a method for bootstrapping standard errors in this type of setting that is based on the design of the experiment. In practice, these standard errors are slightly smaller than those we report from standard clustering at the day level. Since these are more conservative and almost all of our conclusions remain unchanged using the Abadie et al. (2023) method, we present the more conservative standard errors using Cameron and Mill (2015). Additionally, we estimated a specification in which we clustered the standard errors at the RED team day by referral level. Again, this specification leads to slightly smaller standard errors. Finally, we also conducted tests for inference when testing multiple hypotheses using both Romano-Wolf (Clarke, Romano, and Wolfe (2020)) and Young (2019) methods, and the statistical significance of our results remains. Results for these three specifications are available upon request.

[xix] Children can be referred multiple times over the study period. In our sample, 80 percent of children are referred once. Of the remaining 20 percent, 48 percent have different treatment statuses over the course of the experiment. Using the sample of only first referrals for each child

in the sample, we find similar point estimates for the outcome of time spent in discussion as well as similar, though qualitatively smaller, estimates for all other outcomes. We present the model run on the full sample of students, as crossover between treatment could occur in a similar study. Results are available upon request.

[xx] The point estimates between the ITT results in columns (1) and (2) and the IV/LATE results in columns (3) and (4) are largely consistent with one another, though the IV/LATE results are larger in magnitude.

[xxi] We used score bins because we do not have enough power to identify the relationship for individual risk scores. Specifically, a score of 15 only occurs 3% of the time in our sample which represents 214 observations. Recall that because Douglas County has lower rates of child maltreatment than the greater Denver Metro Area used to calculate the predicted risk score, these are not quartiles of the risk distribution within Douglas County.

[xxii] In the CPS data for our setting, measure of race and ethnicity are missing for at least half of children. This is likely to be the case in all child welfare system data, which relies on reporters to volunteer information about a child's race and ethnicity when they report.