Algorithms, Humans and Racial Disparities in Child Protection Systems: Evidence from the Allegheny Family Screening Tool*

Katherine Rittenhouse†
University of California, San Diego

Emily Putnam-Hornstein
UNC Chapel Hill

Rhema Vaithianathan
Auckland University of Technology

June 19, 2023

Abstract

We ask whether providing decision-makers with an algorithm can reduce racial disparities. Our context is the implementation of the Allegheny Family Screening Tool (AFST), a predictive risk model that aims to help child protection workers decide which allegations of maltreatment to screen in for investigation. Using difference-in-differences and triple difference designs, we find that the AFST reduced disparities in screening decisions, as well as case opening and home removal rates for investigated referrals involving Black vs. white children.

JEL Codes: J12, J13, J15, J18

*We are grateful for helpful comments from David Arnold, Lindsey Buck, Julie Cullen, Sarah Font, Max Gross, Katherine Meckel, Chris Mills, Marie Pascale-Grimon and David Simon. Although this analysis was not commissioned, Vaithianathan and Putnam-Hornstein wish to disclose that they were contracted by Allegheny County to build the AFST and continue to work with the county on other projects. Putnam-Hornstein also acknowledges support from NICHD P50HD096719. The opinions, findings, and conclusions or recommendations expressed in the paper are those of the authors and do not necessarily reflect the view of any agency or funding partner. All errors are our own.

†Contact e-mail: krittenh@ucsd.edu. Contact address: 9500 Gilman Drive 0508, La Jolla CA 92093.

1 Introduction

Machine learning tools, actuarial tools and predictive risk models ("algorithms") can help human systems make better decisions (Kleinberg et al. 2018). As such, algorithms are increasingly promoted as useful complements to human decision making in a wide variety of settings, including bail decisions (Chohlas-Wood 2020), resume screening (Raghavan et al. 2020), health care (Price 2019), and, more recently, the child protection system. As their prevalence grows, so too do concerns that algorithms may entrench or exacerbate existing disparities in system interactions, and in particular disparities across race.[1] However, human bias is also well-established, and has been shown to cause racial disparities in many of society's institutions.[2] In this paper, we ask whether the introduction of a predictive risk model increases or decreases disparities, relative to the relevant counterfactual of human decision making.

We address this question in the context of the child protection system, an institution which interacts with approximately one third of U.S. children by the time they reach 18 (Kim et al. 2017). The child protection system is comprised of several high-stakes decision points, including whether to investigate alleged maltreatment, and whether to remove children from abusive or neglectful homes. This is also a setting with large racial disparities – Black children are 88 percent more likely than white children to be investigated for maltreatment, and over twice as likely to enter foster care by the time they reach 18 (Kim et al. 2017; Wildeman and Emanuel 2014). These observed disparities may be driven by differences in the rates of maltreatment across race, and/or by biased decision-making. Improving racial equity is a top priority for many child welfare agencies.[3] As more agencies consider using predictive risk modeling in their decision processes, the potential impact of such algorithms on racial disparities is a primary concern.[4]

We study the implementation of the Allegheny Family Screening Tool (AFST), the first automated predictive risk model used to aid decision-makers in the child protection system. This algorithm aims to help workers decide whether to "screen in" referrals alleging that a child is be-

[1] Several high-profile media reports have drawn attention to the potential for algorithms to discriminate (Barry Jester, Casselman, and Goldstein 2015; Angwin et al. 2016). Academic research has in many cases validated these concerns, documenting and exploring the consequences of algorithmic bias in a variety of contexts, including healthcare (Obermeyer et al. 2019) and the criminal justice system (Arnold, Dobbie, and Hull 2021).

[2] See, for example, Goncalves and Mello (2021), Antonovics and Knight (2009), Rehavi and Starr (2014), Arnold, Dobbie, and Yang (2018), and Abrams, Bertrand, and Mullainathan (2012).

[3] See, for example, Gateway (2021) and Thomas and Halbert (2021).

[4] Oregon's Department of Human Services halted use of an algorithm after concerns were raised about its potential to disproportionately affect Black families (Ho and Burke 2022b). California's Department of Social Services abandoned plans to implement a predictive risk tool in part due to worries about the effects on racial equity (Ho and Burke 2022a). Allegheny County, PA (the context for this study) has faced widespread criticism for its use of a predictive risk model and the potential disparate impacts by race.

ing maltreated. The AFST uses information about the referred families from linked administrative data to predict the risk that the child will be removed from their home if screened in, and shows the human decision-maker a risk score ranging from 1 to 20. For referrals with the highest risk scores, the AFST defaults to a screen-in recommendation. In all cases, the call-screening supervisors remain the ultimate decision-maker, and may choose to override the default. Screened-in referrals proceed to an investigation, conducted by a different caseworker. Investigations may result in a case opening (requiring further contact with caseworkers and possibly mandated services). In severe cases where the investigating caseworker determines that a child's safety is at risk, investigations may result in a child being removed from their home and placed in foster care. We ask how the implementation of the AFST affected: (1) the differential probability of screening in referrals involving Black vs. white children; (2) the differential probability of opening a child welfare case for referrals involving Black vs. white children; and (3) the differential probability of home removal for referrals involving Black vs. white children.

Our setting is well-suited to studying the effects of implementing the algorithm, as we observe outcomes for referrals made both before and after the AFST came online. We obtain data on referrals made to Allegheny County's office of Children, Youth and Families (CYF) between 2010 and 2020. For each referral, we observe the AFST score (retroactively calculated for referrals made prior to implementation), the screening decision, and the case opening decision. Referrals are also associated with one or more children, for whom we observe demographic information. We then link these individual children to the universe of foster care records between 2010 and 2020, in order to study home removals.

To study the effects of the AFST on disparities in screening rates we employ a difference-in-differences design, comparing referrals involving Black vs. white children, made before vs. after the AFST was implemented. We validate the necessary assumption of parallel trends in screening rates across race visually and with an event study. We find that the AFST reduced disparities in screening rates by 19%. If the higher screen-in rate for referrals involving Black children reflects differences in the riskiness of referred Black vs. white children, it may not be the case that reducing those disparities is welfare enhancing. However, if the algorithm is equally predictive across race, screening disparities for children within the same risk bin are likely unwarranted. By including a fixed effect for algorithm score in our difference-in-differences model, we show that the AFST reduces within-score disparities by 2.7 percentage points, or 46% of the pre-existing gap.

Next, we study the effects of the AFST on downstream outcomes for screened-in referrals including whether a case is opened for the family, and whether a child on the referral is removed from home. For this analysis, we use two empirical strategies. First, we again use a difference-

in-differences design comparing referrals involving Black vs. white children, before vs. after the AFST was implemented. Second, we use a triple difference design, comparing across referrals which were "treated" vs. not "treated" by the AFST. Specifically, under Pennsylvania law, referrals which include certain allegations are automatically screened in for investigation, and as such were not affected by the screening algorithm. This design enables us to control for any trends across time which might differentially affect Black vs. white families. One drawback of the triple difference design is that the control-group referrals are expunged from the data prior to 2015, giving us a shorter timeframe to study. Both designs yield similar results, finding that the AFST reduced disparities in case openings for screened-in referrals by 5-6 percentage points (80-90% of the pre-existing gap), and home removals within three months by 3 percentage points (70% of the pre-existing gap).

There are several possible causal channels through which the implementation of the AFST might affect these downstream outcomes. Investigating caseworkers do not observe the algorithm score assigned at the time of referral. As such, any changes in outcomes for screened-in cases must be driven by changes in the composition of screened-in referrals, along the dimensions of race and riskiness. This composition change could include one or more of the following: (1) change in probability of screen in for referrals involving high-risk white children; (2) change in the probability of screen in for referrals involving high-risk Black children; (3) change in the probability of screen in for referrals involving low-risk white children; and (4) change in the probability of screen in for referrals involving low-risk Black children. Each of these causal channels may have a different welfare interpretation, depending in part on the relative costs and benefits of case openings and removals across race and riskiness. This paper does not speak to the welfare effects of reducing disparities, or of the algorithm overall.

This paper adds to the growing literature which assesses the ways in which humans interact with algorithms within high-stakes decision making, and the effects of that interaction on observed disparities. Several high-profile media reports have drawn attention to the potential for algorithms to discriminate.[5] Academic research has in many cases validated these concerns, documenting and exploring the consequences of algorithmic bias in a variety of contexts, including healthcare (see, for example, Obermeyer et al. (2019)) and the criminal justice system (see, for example, Arnold, Dobbie, and Hull (2021)). For policy-makers, a relevant question is whether algorithms *worsen* disparities, relative to the human-only systems they replace. Previous work addressing this question is limited, and has found mixed results. Stevenson and Doleac (2021) study the adoption of algorithmic risk assessments on judge decisions in felony sentencing in Virginia, and

[5]See Barry Jester, Casselman, and Goldstein (2015) and Angwin et al. (2016).

do not find any evidence of effects on racial disparities. Albright (2019) studies these issues in the context of pretrial bond decisions in Kentucky, and finds evidence that judges respond differently to revealed risk assessments, depending on the race of the defendant. This differential application of the risk score by race caused an increase in racial disparities in non-financial bond rates. Howell et al. (2021) study racial disparities in small business loans, and find that disparities decrease when the process is more automated, relative to when humans are more involved. In highly related work, Grimon and Mills (2022) provide child welfare workers with randomized access to an algorithmic risk score. They find that providing access to the score reduced child injury hospitalizations, and reduced racial disparities in child welfare contact. They show that providing access to the algorithm score allows screeners to focus on other salient aspects of the allegations.

Prior studies in the computer science literature have studied the design and deployment of the AFST, as well as its possible implications for racial equity. Chouldechova et al. (2018) describe the development, validation, fairness auditing and deployment of the AFST. De-Arteaga, Fogliato, and Chouldechova (2020) study the effect of the AFST on call-screeners' decisions, finding that humans updated their behavior to align more closely with the risk score. They also study a period when a technical glitch led to some incorrect scores shown to call screeners, and find that screeners were less likely to adhere to the algorithm's recommendation in this case.[6] Finally, Cheng et al. (2022) compare screening decisions under the AFST to a theoretical setting where the AFST is used without human input. However, this exercise is purely theoretical, as the AFST was not designed to be used without human supervision. In contrast, our paper studies the effects of the AFST as it was used in practice, relative to the prior decision-making process.

This paper also contributes to an interdisciplinary literature which attempts to understand causes of and solutions to racial disparities within child protection systems. While the existence of racial disparities is well-established, the causes for those disparities are not fully understood. It is not clear whether disparities result solely from underlying differences in incidence of maltreatment due to co-occurring risk factors, or if and at which stages bias plays a role. Several studies find that physicians may be more likely to report cases of potential abuse if the child is Black, and miss or overlook cases of abuse if the child is white (Lane et al. (2002), Jenny et al. (1999), Hampton and Newberger (1985)). More recent work suggests that there are significant differences in risk of maltreatment across race, and that this is likely a primary driver of disparities in child

[6]Our main analysis is not affected by this glitch, as we use comparable retroactively-calculated scores from a newer version of the AFST, rather than the observed scores for each given referral. Moreover, the glitch only affected a small fraction of referrals, and in general the shown score was close to the true score. See De-Arteaga, Fogliato, and Chouldechova (2020) for more details on the technical glitch, as well information on the correlation between observed and true scores.

protective service interactions. National differences in the rates of substantiated child abuse by race are largely consistent with racial differences in other public health outcomes, including infant mortality, low infant birth weight, and premature birth (Drake et al. 2011), suggesting that both sets of disparities may be driven by differences in exposure to the same underlying risk factors. Drake et al. (2021) provides an overview of the evidence linking poverty to child maltreatment. Not only is poverty strongly correlated with maltreatment rates, but recent evidence suggests that the link is causal.[7] Several studies show that when income, or proxies for income, are controlled for, disparities in referrals, victimization rates and foster care placement rates are reduced and in some cases even reversed.[8] Reducing racial disparities is an ongoing priority for child welfare agencies, which has led to the adoption of new policies and practices with minimal evidence that they reduce disparities. For example, "blind-removals", where people deciding whether to remove a child do not observe the child's race, are being implemented (New York State Office of Children and Family Services 2020) and championed (Programs 2021), despite a lack of evidence of their effectiveness, and even indications that the practice could negatively affect child safety (Baron, Goldstein, and Ryan 2021).

The rest of the paper proceeds as follows. In Section II we describe the institutional context of the Allegheny County child protection system, as well as the Allegheny Family Screening Tool. Section III describes our data and Section IV our empirical strategies. Section V presents and discusses results, and Section VI concludes.

2 Allegheny County

Allegheny County, Pennsylvania is home to 1.2 million people, and includes the city of Pittsburgh within its boundaries. The Office of Children, Youth and Families in Allegheny County is responsible for investigating allegations of child neglect and abuse. In Pennsylvania, child welfare is a State-supervised, county-administered system.

2.1 Referral Process

Allegations of maltreatment are brought to the attention of the County by mandated reporters and community members. The State of Pennsylvania categorizes each referral under either Child Protective Services (CPS) or General Protective Services (GPS), based on the type of allegation.

[7] See, for example, Berger et al. (2017), Raissian and Bullinger (2017), Cancian, Yang, and Slack (2013), Kovski et al. (2022), and Rittenhouse (2023).

[8] See Putnam-Hornstein et al. (2013) and Maloney et al. (2017).

CPS referrals include an allegation of abuse as defined in state statute, whereas GPS referrals allege neglect, implying a child may be at risk due to inadequate parental care.[9] After this classification is made by state staff, all referrals are sent to the County for further review and possible investigation. Figure 1 presents a simplified depiction of how maltreatment referrals move through the child protection system in Allegheny County. By state law, CPS referrals must always be investigated. For the remainder of referrals (GPS), staff (or "screeners") must decide whether or not to screen in the referral for investigation.

Screened-in referrals are assigned to an investigator operating out of a regional office.[10] The investigator visits the home of the alleged victim, speaks to collateral contacts (e.g., teachers, other family members), and may gather medical and other information to evaluate the allegations of maltreatment and determine whether the child or family is in need of additional monitoring or services. State law requires that the investigation is concluded within 60 days of receipt of the report.

Based on their findings, the investigator and their supervisor decide whether or not to open a case for services. In general, opening a case indicates that the family requires ongoing services or involvement from social workers to ensure the safety of the child. An opened case might result in continued monitoring by a social worker, suggested or mandated participation in services, or, often as a last resort, a court-ordered removal of the child from the home. A court will order removal if there are imminent and unresolved concerns for a child's safety and well-being. Removals can occur at any time during an investigation, or after a case has been opened.

Other than the subset of referrals which are automatically screened in for investigation, a family's involvement with the child protection system is determined by the screener's decision. Prior to August 2016, screeners relied solely on professional judgement to recommend which of the GPS referrals to screen in. In making this recommendation, screeners could use both information from the referral itself (e.g., reporter, allegations, age of children), as well as data on each child and adult included in the referral from the the linked Allegheny Data Warehouse (e.g., a child's history of foster care placements, adult arrest records). The Data Warehouse provides individual-level information on previous child protection system involvement, as well as involvement in a range of other County systems.[11] However, while these data were available and the County expected infor-

[9]The Child Protective Services Law is the relevant Pennsylvania statute which defines child abuse and prescribes the counties' responsibility. Allegheny County provides a brief overview of the two types of referrals here: https://www.alleghenycounty.us/Human-Services/Programs-Services/Children-Families/Protective-Services.aspx.

[10]In some cases, there might be multiple referrals made by different people about the same allegation or incident. The County may in these instances, combine all of these referrals into one referral, i.e. requiring just one investigation.

[11]*Allegheny County Data Warehouse* (2021) provides a detailed description on the linked data systems, and how they feed into the AFST.

mation to be systematically reviewed, there was little guidance for screeners on exactly how those data should be incorporated into their screening decision.[12] There was also no way for the County to confirm whether or not a screener had reviewed data to inform their decision. Call screeners' recommendations are reviewed and approved by a supervisor. Note, after making their recommendation, call screeners would not learn about any outcomes for investigated or screened-out families.[13] As such, there was little opportunity for improvement in decision making.

2.2 Referral Process and the Allegheny Family Screening Tool

In August 2016, Allegheny County implemented the AFST, a predictive risk model to help screeners decide which referrals to screen in for investigation. Note, the AFST score is generated for both CPS and GPS referrals, but is only included in the decision-making process for GPS referrals. CPS referrals are screened in 100% of the time both before and after AFST implementation. Further details on the design and implementation of the algorithm are included in subsection 2.3 below; this section explains how the AFST changed the decision-making process for screeners. After reviewing the referral, as well as other historical information on the family from the Data Warehouse, the screener now runs the AFST, which shows a numerical score between 1 (lowest risk) and 20 (highest risk). Figure 2a shows an example of what the screener would see. The score is generated using only information that the screener has access to, but may not have time to review in detail and may not know how to incorporate into their assessment of safety and risk.

For referrals with a score greater than 17 and at least one child aged 16 or under, the screener sees a "High Risk Protocol" notification, with no numeric score (see Figure 2b). These referrals are recommended to be screened in for investigation, and require explicit supervisor approval to be screened out.

For referrals with a score less than 11 and no children under the age of 12, the screener sees a "Low Risk Protocol" notification, again with no numeric score (see Figure 2c). These referrals are recommended to be screened out, but no supervisor approval is required to screen in these referrals.[14] Low-risk protocols initially made up only 4% of referrals (Vaithianathan et al. 2017), and as such have a limited possible impact on overall screening rates.

[12] Based on conversations with call screeners and supervisors, they primarily focus on age and allegation in order to determine the likely safety of children on referrals.

[13] In some very rare cases where a fatality occurred, screening staff might learn about any mistakes that had been made, e.g., in screening out an at-risk child.

[14] The low risk protocol has changed over time. Prior to 2018 there was no low risk protocol. From November 2018 through October 2019, referrals fell under the low risk protocol if the maximum score was less than 10 and all children were over age 11. In October 2019, the protocol criteria was expanded to include referrals with a maximum score less than or equal to 12 and no children aged 6 or younger.

For referrals which do not meet the criteria for high- or low-risk protocols, the screener observes the numeric score, and there is no screening decision recommendation.

The score is not seen outside of the screening process. That is, investigators and caseworkers do not have access to the results of the predictive risk model, and thus their downstream decisions should not be directly affected by the score. Screeners work in a centralized office, while investigators and caseworkers are based out of regional field offices, so the two groups have little chance to interact.

2.3 Allegheny Family Screening Tool

For each child associated with a referral, the AFST predicts the risk that, if screened in for investigation, that child will experience a court-ordered removal from their home within two years. The model uses data associated with all individuals on the referral, including alleged victims and other children in the household, household members, parents and alleged perpetrators. Data from past referrals and interactions with child welfare, past and present involvement with the courts, jail and other County systems, as well as information from the child's birth record, are all used to generate a risk score.[15] A risk score is generated for each child living in the home of the alleged victim, but the screener only observes the maximum of these scores.[16]

Allegheny County Department of Human Services developed the AFST with the purpose of using existing data to improve the quality and consistency of screening decisions.[17] Importantly, the tool was never meant to replace human decision-making, but rather to inform and improve those decisions. That is, the tool was intended to be complementary to the call screener's professional judgement.

In August 2016 Allegheny County deployed the AFST. Since then, they have updated the predictive risk model and the screening tool twice. In November 2018, the original model was replaced with a LASSO model.[18] In January 2019, the LASSO model was updated in response to a change in the data that were available to the model. Goldhaber-Fiebert and Prince (2019) were contracted to conduct an independent impact evaluation of the original AFST, and studied effects on accuracy, workload, disparities, and consistency.

[15] A full list of the features used in the latest version of the algorithm can be found in Vaithianathan et al. (2017).

[16] For example, two children in the same household may have different histories with child protective services, which leads to different risk scores. However, the screener will only see one score per referral.

[17] See Vaithianathan et al. (2019) for an overview of the development of the original AFST. Additional background and documents related to the AFST are available at: https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

[18] See Vaithianathan et al. (2017).

3 Data

We obtained de-identified administrative data from Allegheny County on the universe of child maltreatment referrals (both CPS and GPS) made between 2015 and 2020. For GPS referrals, we also obtain data on referrals made between 2010 and 2015.[19] Every referral links to unique IDs for each person living in the household, including the alleged victim, other children, parents and alleged perpetrators. For each alleged victim, the referral lists one or more allegations of abuse or neglect. For each individual, we observe demographic information including race, gender, and age. We also observe outcomes within the child protection system at each decision point. There are three decision-points of interest in this study: (i) screening; (ii) case-opening; and (iii) home removals.

Screening and case opening decisions each occur at the referral, rather than individual, level. As such, we collapse data to the referral level in our main analyses. In order to study effects on removals (which occur at the individual level), we create a variable equal to one if any child on the referral is removed from their home within three months of the referral date, and zero otherwise. We choose three months since investigations are required to be completed within 60 days of referrals, and as such any removals associated with a given referral are likely to occur within approximately this time. We also define race at the referral level, classifying a referral as Black if at least one child on that referral is identified as Black or African American, and classifying a referral as white if there are no Black children and at least one white child on the referral. Referrals with no children identified as either Black or white make up approximately 6% of referrals from 2010 through 2020 and are excluded from our analysis sample.

Table 1 presents summary statistics separately for GPS (all and screened-in only) and CPS referrals. We exclude referrals without a CPS or GPS designation (about 3% of referrals from 2010 - 2020). Our sample is comprised of over 100,000 referrals. We do not include referrals for families which, at the time of referral, have an active case with CYF (about 9% of referrals from 2010 - 2020). We also exclude referrals made by truancy courts (about 1% of referrals from 2010 - 2020). While only 13% of the population in Allegheny County is Black, approximately 50% of referrals involve a Black child. This disproportionate representation of Black children and families is reflective of national trends. Note also that 100% of CPS referrals are investigated, reflecting the automatic screen in for this category.

For each referral, we observe one or more algorithm-generated scores. First, we observe the score as calculated by the algorithm in use at the time of referral. In addition, we observe

[19] CPS referrals made prior to 2015 were expunged and are not available for analysis.

retroactively-calculated scores generated by AFST V3, the most recent version of the model as of this writing, for all referrals made after January 2013. Going forward, we primarily focus on the V3 score in order to allow for comparability across years. For referrals made between January 2013 and July 2019, this comparable score was retroactively calculated, and can be thought of as the score that the screener *would have* seen, had this version of the algorithm been deployed.[20] Scores are strongly correlated across AFST versions. Figure A1 shows a heatmap of the correlation between retroactively-calculated V3 scores and each of V1 and V2 scores.

4 Empirical Framework

4.1 Screening

To test whether the implementation of the algorithm differentially affected the screening rate for Black vs. white children, we use a difference-in-differences approach, comparing GPS referrals involving Black children to those involving white children before and after the algorithm was implemented.

We begin by estimating the following regression Equation:

$$y_{it} = \beta_0 Black_{it} + \beta_1 Post_{it} + \beta_2 Black \times Post_{it} + \beta_3 X_{it} + \gamma_m + \gamma_y + \epsilon \tag{1}$$

Where y_{it} is an indicator equal to one if referral i, made in month t, is screened in for an investigation; $Black_{it}$ is in indicator equal to one if there are any Black children listed on referral i, and zero if there are no Black children, and at least one white child, listed on referral i; and $Post_{it}$ is in indicator variable equal to one if referral i in month t was made after the implementation of the AFST, and zero otherwise. We include fixed effects for Year (γ_y) and Month-of-Year (γ_m), to control for variation across time and seasonality. Finally, X_{it} is a vector of referral-level controls, which includes allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol exposure.[21] We include these variables to control for any differential trends across race and time. For example, substance exposure might differ across both race and time, as the opioid crisis has

[20] The way in which predictive features are retrospectively coded, it is as close to what those features would have been at the time that the call came in. It is possible that some features may have changed due to subsequent data entering the data-warehouse–for example, demographic data might be updated, but these are in the minority.

[21] Allegation categories are listed in Panel C of Table 1. Reporter categories are listed in Panel D of Table 1. The indicator for exposure to drugs or alcohol is equal to one if any allegation on the referral mentions drugs or alcohol, and zero otherwise.

affected primarily white communities.

The identification assumption for this model requires that there are no differential trends in screening rates for referrals involving Black vs. white children. This assumption may be violated, for example, if there are other efforts to reduce disparities in screening decisions around the time that AFST was introduced, or if Black and white families are differently affected by changing local conditions. We address this concern first by plotting monthly referral numbers and screen-in rates separately for referrals involving Black and white children, in Figure A2. Visually, there does not appear to be any obvious differential trends in the pre-period. We also directly test this parallel trends assumption using an event-study specification of Equation 1.

Finally, effects may differ across the range of algorithm scores. Given the low-risk and high-risk protocols, we might expect that effects on screening would be concentrated at these ends of the risk score distribution. To explore this heterogeneity, we plot screening decisions across the range of risk scores, before and after the AFST deployment and separately for referrals involving Black vs. white children, in Figure 3. We show this comparison first using the entire post period in Figure 3b, and then defining the post period as only the time when AFST V3 was in place (i.e. after July 2019) in Figure 3c. Note that the change to the racial gap in screen-in rates seems particularly stark for high-risk referrals. The AFST appears to have increased the screen-in rates for high-risk referrals involving white children. Motivated by this observation, we look for heterogeneous effects according to algorithm risk scores, by interacting the indicator variables in Equation 1 with indicators for each algorithm score decile $S^i, i \in \{2, 10\}$, estimating the equation:

$$\begin{aligned} y_{it} = &\sum_{j \neq 6}^{10} \beta_j S_{it}^j + \sum^{10} \beta_{j+9} S_{it}^j \times Black_i + \sum^{10} \beta_{j+19} S_{it}^j \times Post_t \\ &+ \sum^{10} \beta_{j+29} S_{it}^j \times Black_i \times Post_t + \beta_{40} X_{it} + \gamma_m + \gamma_y + \epsilon \end{aligned} \tag{2}$$

Where y_{it} is an indicator equal to one if a referral is screened in for investigation, and zero otherwise. This approach allows us to test how the algorithm affected disparities in screen-in rates differently by comparable risk score. The first decile, or scores 1-2, are excluded due to small sample size.

4.2 Downstream Outcomes

We next turn to the downstream outcomes of case opening and removal. For this analysis we conduct two empirical exercises, with different strengths and weaknesses. We start with a difference-

in-differences approach. We estimate Equation 1, where y_{it} is set to each of: (1) an indicator equal to one if screened-in referral i, made in month t, is associated with a case opening; and (2) an indicator equal to one if any child associated with screened-in referral i, made in month t, is removed from their home within three months. The identification strategy requires the parallel trends assumption that case openings and removals evolve over time similarly for referrals involving Black vs. white children. We plot these outcome variables of interest over time in Figure A3, and test this assumption directly using an event-study specification of Equation 1.

We also estimate a triple differences model, using CPS referrals as the control group. For this analysis we use a sub-sample of referrals made between 2015 and 2020, as CPS referrals made before 2015 are expunged in our data. To illustrate and motivate the triple differences approach, we plot our two outcome variables of interest across each of the three differences (Pre vs Post, Black vs. white, CPS vs. GPS), in Figure 4. Figures 4c and 4d show the evolution of racial disparities in case openings and removals, before and after algorithm implementation. Note, in Figures 4a and 4c, that while the disparities in case opening rates seem stable across time in the control group, they fall in the treated group after the implementation of the algorithm. In Figures 4b and 4d, the pattern is similar for removal rates.

The triple difference model is estimated with the following Equation:

$$\begin{aligned} y_{it} = \beta_0 Black_{it} + \beta_1 GPS_{it} + \beta_2 Post_{it} + \beta_3 Black \times GPS_{it} + \beta_4 Black \times Post_{it} \\ + \beta_5 GPS \times Post_{it} + \beta_6 Black \times PostXGPS_{it} + \beta_7 X_{it} + \gamma_m + \gamma_y + \epsilon \end{aligned} \tag{3}$$

Where everything is defined as in Equation 1, and GPS_{it} is an indicator equal to one if referral i falls under GPS, and zero if referral i falls under CPS. The coefficient of interest, β_6, tells us how case opening and removal rates change for Black children, relative to white children, in the treated group relative to the control group of referrals. Note, a triple difference specification is not possible for estimating effects on screening decisions, as all CPS referrals are screened in both before and after algorithm implementation.

The identification assumption required for this triple differences specification is weaker than that for difference-in-differences, and requires common trends in racial disparities for CPS and GPS referrals. We directly test this assumption in an event-study specification.

4.3 Interpretation

Effects of the AFST on the downstream outcomes of conditional case openings and removals have a somewhat complex interpretation. Recall that caseworkers and any other downstream decision makers do not observe the AFST score assigned at the time of referral. Any changes in mean outcomes must thus be driven by AFST-driven changes in the composition of screened-in referrals. For example, an AFST-driven reduction in the racial disparity in case opening (removal) rates could be driven by any or some combination of the following:

1. increased probability of screen in for referrals involving white children at high risk of having a case opened (being removed),

2. decreased probability of screen in for referrals involving Black children with a high risk of having a case opened (being removed),

3. decreased probability of screen in for referrals involving white children at low risk of having a case opened (being removed),

4. increased probability of screen in for referrals involving Black children with a low risk of having a case opened (being removed).

We attempt to distinguish between these channels in two analyses. First, we ask where in the AFST-defined risk distribution effects are concentrated, by estimating Equation 2 with y_{it} set to each downstream outcome of interest. Next, we ask whether any changes in downstream disparities are driven by referrals involving Black children, white children, or a combination. To do this, we estimate another difference-in-differences model comparing outcomes across time (Pre vs. Post) and treatment status (GPS vs. CPS). That is, separately for referrals involving Black children and referrals involving white children, we estimate the equation:

$$y_{it} = \beta_0 GPS_{it} + \beta_1 Post_{it} + \beta_2 GPS_{it} \times Post_{it} + \beta_3 X_{it} + \gamma_m + \gamma_y + \epsilon \tag{4}$$

Where everything is defined as in Equation 3. The coefficient on $GPS_{it} \times Post_{it}$ will tell us whether the AFST increased or decreased the average "riskiness" of screened-in referrals involving children of a given race.

Each of the potential channels described above may have different welfare implications, which depend in part on the costs and benefits of having a case opened or a child removed for each of high-risk and low-risk referrals involving Black vs. white children. In this paper, we make no argument as to the welfare implications of changes in disparities.

5 Results and Discussion

5.1 Screening Decision

In Table 2 we report results from estimating Equation 1, or the effects of the algorithm on disparities in screen-in rates. Column (1) reports results from a difference-in-differences specification excluding fixed effects and controls, Column (2) adds year and month-of-year fixed effects, Column (3) adds referral-level controls, Column (4) replicates Column 3 for the time period in which we have AFST V3 scores, and Column (5) adds an additional fixed effect for the underlying AFST V3 score. First, note that referrals involving Black children are more likely to be screened in than referrals involving white children. There is a 10.9 percentage point gap in screen-in rates before controlling for referral-level characteristics, which is reduced to 5.9 percentage points after controlling for referral-level characteristics and underlying risk scores. The algorithm reduces the unconditional gap by 2.0 percentage points (Column 1), or 19%. It reduces the conditional gap by 2.7 percentage points (Column 5), or 46%.

Figure A4 presents coefficients from an event study version of Equation 1, where each of $Post_{it}$ and $Black \times Post_{it}$ are interacted with quarterly indicator variables. While the estimated coefficients are noisy, there are no obvious violations of the parallel trends assumption.

Motivated by the patterns in screening disparities observed in Figure 3, we next study how effects on screening decisions vary across algorithm scores. Figure 5 shows coefficients and 95% confidence intervals from estimating Equation 2. Figure 5a shows a positive relationship between algorithm score bin and screen-in rate. Figure 5b shows that referrals involving Black children are more likely to be screened in at every risk score. Figure 5c suggests that the algorithm primarily increased screen-in rates for referrals with the highest risk scores. Finally, Figure 5d graphically presents the coefficients on $Black \times Post \times Score^i$. The effect on screening disparities is negative for every score bin, with varying magnitudes and levels of statistical significance. For referrals in the highest risk bin, the algorithm reduced the gap in screening rates across race by 8.8 percentage points, or 83% percent of the pre-existing disparity in this score bin.[22]

Referrals within this highest-risk decile are most often defaulted to be screened in through the high risk protocol under AFST (see Section 2.2).[23] Recall, although referrals in this category may still be screened out, this decision requires a supervisor's override. This result suggests that the

[22] In unreported results from this regression, the coefficient on $Black \times Post \times Score^{10}$ is -0.088 and the coefficient on $Black \times Score^{10}$ is 0.106 (each significant at the 1% level).

[23] Since we use the comparable AFST V3 score, the highest risk decile does not correspond exactly with the high-risk protocol. That is, there may be referrals with an AFST V3 score of 19-20, but a screener-observed score (from an earlier algorithm version) below 17.

protocol plays an important role in changing screening outcomes. One might ask why we do not see similarly stark effects for the referrals with the lowest risk scores, which fall under the low-risk protocol and are defaulted to be screened out. However, the low-risk protocol was initially implemented in quite a weak way — only 4% of the referrals were expected to meet the initial low-risk protocol (Vaithianathan et al. 2017).

5.2 Downstream outcomes

We next turn to the effects of the algorithm on the downstream outcomes of case opening and removal. We start by estimating a difference-in-differences model, using only the GPS referrals. Results from estimating Equation 1, where y_{it} is an indicator for case opening conditional on screen-in, are reported in Column (1) of Table 3. Screened-in referrals involving Black children are on average 5.9 percentage points more likely to have a case opened (15 percent of the sample mean). After the implementation of the algorithm, case opening rates increased overall by 8.8 percentage points, consistent with an increase in the average severity or accuracy of screened-in referrals. The coefficient on $Black \times Post$ is negative and significant, suggesting that the implementation of the algorithm reduced disparities in case opening rates by 5.2 percentage points, or 87% of the pre-existing Black-white gap. The corresponding results for removals are reported in Column (1) of Table 4. Screened-in referrals involving Black children are on average 3.9 percentage points more likely to involve a child who is removed from their home within 3 months (44 percent of the sample mean). After the implementation of the algorithm, removal rates for screened-in referrals increase by 3.0 percentage points, again consistent with a change in the composition of screened-in referrals. The coefficient on $Black \times Post$ suggests that the implementation of the algorithm reduced disparities in removal rates by 3.1 percentage points, or 80% of the pre-existing Black-white gap. We also estimate effects on removals *unconditional* on screen in, reported in Table A1. Effects are attenuated (in line with lower average 3-month removal rates for this group), but negative and statistically significant at the one percent level. We do not estimate effects on unconditional case opening rates, as the case opening decision is associated with a given referral.

Figure A5 presents coefficients from a event study versions of Equation 1, where y_{it} is either an indicator for whether a screened-in referral has a case opened, or a child removed within 3 months. There do not appear to be any obvious violations of the parallel trends assumption for either outcome.

Next, we incorporate CPS referrals as a control group in a triple differences specification. For this analysis, we must restrict our sample to referrals made in or after 2015. For easier comparison across specifications, we also run a difference-in-differences model using this shorter time span.

Results from estimating Equation 1 on this restricted sample are reported in Column (2) of Tables 3 and 4, and are statistically indistinguishable from the results in Column (1).

Results from estimating Equation 3, where y_{it} is an indicator for case opening conditional on screen-in, are reported in Columns (3) through (5) of Table 3. Column (3) reports results from a basic triple differences specification, Column (4) reports results from a regression which adds year and month-of-year fixed effects, and Column (5) reports results from a regression which adds referral-level controls (our preferred specification). Our coefficient of interest, on the triple interaction $Black \times Post \times GPS$, is significant at the 1% level, and stable across specifications. In our preferred specification (reported in Column 5), the coefficient implies that the introduction of the algorithm reduced the relative likelihood of having a case opened for screened-in Black children by 5.6 percentage points. This accounts for 91% of the pre-existing difference between GPS referrals involving Black and white children (6.1 percentage points).[24]

Results from estimating Equation 3 on the conditional likelihood of removal within three months are reported in Columns (3) through (5) of Table 4. The coefficient of interest (on the triple interaction $Black \times Post \times GPS$) is negative, statistically significant, and is robust to adding fixed effects and referral-level controls. According to our preferred specification (Column 5), the introduction of the algorithm reduced the racial disparity in three-month removal rates by 3.1 percentage points, or 77% of the pre-existing difference.[25] Again, effects on unconditional 3-month removals are reported in Table A1, and are attenuated but negative and statistically significant at the 10 percent level.

Event study coefficients and 95% confidence intervals for each of these downstream outcomes are shown in Figure A6. We do not find any evidence of a violation of the parallel trends assumption.

5.3 Interpretation

As discussed above, effects on downstream outcomes can be interpreted as changes in the composition of screened-in referrals across race and riskiness.

To study where in the risk distribution downstream effects are concentrated we estimate Equation 2, setting the outcome variable equal to either an indicator for case opening conditional on screen in, or removal within three months conditional on screen in. For this analysis we restrict our sample to referrals made in or after 2013, as we have comparable scores for these dates. The

[24] To calculate the pre-existing difference in case opening rates for GPS referrals involving Black vs. white children, we sum the coefficients on $Black$ and $Black \times GPS$. 0.0268 + 0.0345 = 0.0613.

[25] Similarly to above, to calculate the pre-existing difference in 3-month removal rates for GPS referrals involving Black vs. white children, we sum the coefficients on $Black$ and $Black \times GPS$. 0.0278 + 0.0147 = 0.0425.

coefficients on $Score_i$, $Black \times Score_i$, $Post \times Score_i$ and $Black \times Post \times Score_i$ are shown in Figure 6 (case openings) and Figure 7 (placements). In Figure 6a, case opening rates among screened-in referrals are generally increasing in algorithm score. In Figure 7a, conditional removals are relatively flat on the lower end of the risk distribution, and increase only in the top three deciles of the risk score. This is consistent with removals being indicators of severe or chronic problems (i.e., associated with higher risk). In Figure 6b, note that screened-in Black children are generally more likely to have a case opened across the range of risk scores. In Figure 7b, the race differential in likelihood of removal is most stark for the highest-risk referrals. In Figure 6c, we see that conditional case openings are higher across the risk distribution in the Post period (consistent with an increase in accuracy of screening decisions). Figure 7c shows that conditional removals are also generally higher in the Post period, except for the highest-risk referrals. This likely reflects a reduction in average severity for screened-in referrals in this risk bin, consistent with the high-risk protocol increasing screen-in rates for less-severe referrals. Finally, for case openings, the effect on disparities is relatively evenly distributed across the risk score distribution, as shown in Figure 6d. For removals, the effect on disparities is somewhat concentrated in the highest risk decile, as shown in Figure 7d.

To study whether reductions in downstream outcome disparities are driven by changes for referrals involving white children, referrals involving Black children or both, we estimate Equation 4 separately by race. Results are reported in Table 5 for both case openings (Columns 1-2) and removals (Columns 3-4). For referrals involving Black children, the implementation of the AFST has no statistically significant effect on conditional case openings, and significantly reduces conditional removals by 1.9 percentage points. For referrals involving white children, the implementation of the AFST has increase conditional case openings by 3.5 percentage points, and has no statistically significant effects on conditional removals. For case openings, these results imply that the effect on disparities is driven largely by an increase in the average severity of screened-in referrals involving white children. For removals, in contrast, the effect on disparities is driven by a decrease in the average severity of screened-in referrals involving Black children. One explanation for such a change in composition could be that the AFST is leading to earlier detection of maltreatment for high-risk Black children, allowing for in-home prevention services in place of removals. Another explanation could be that the AFST is flagging as "high risk" referrals involving Black children which actually have low risk of short-term removal and which workers would otherwise have screened out.

Again, it is not clear from this analysis whether reducing disparities in screening or downstream disparities is welfare enhancing. We do not have estimates for the welfare effects of investigations,

case openings and home removals in this setting, or how those effects differ across race and risk level. That said, if policy-makers are concerned that predictive risk models will exacerbate existing racial disparities, this analysis suggests that algorithms can be implemented in such a way as to narrow disparities in child protection system involvement.

6 Conclusion

Predictive risk models are increasingly used to assist decision makers in a wide variety of settings. Academics, activists, and policy makers have rightfully raised concerns that such algorithms may exacerbate existing biases, or even create new ones. We show that, for one particular algorithm and institutional context, predictive risk models can also serve to reduce racial disparities, relative to human decision makers.

Using a difference-in-differences and a triple difference design, we study the effects of the Allegheny Family Screening Tool on racial disparities in screening decisions, case opening rates and home removals. Relative to the prior decision-making protocols, the implementation of the AFST reduced disparities in each of these outcomes.

We cannot and do not speak to either optimal investigation and removal rates, or the welfare consequences of reducing disparities in these outcomes. Future work is needed to address these crucially important questions. However, policy-makers and communities are concerned about racial disparities in the child protection system, and are actively working to improve racial equity. Our work shows that predictive risk models may be a useful tool for reaching this specific policy goal.

References

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do judges vary in their treatment of race?" *The Journal of Legal Studies* 41 (2): 347–383.

Albright, Alex. 2019. "If you give a judge a risk score: evidence from Kentucky bail decisions." *Working Paper.*

Allegheny County Data Warehouse. 2021. Technical report. Allegheny County Department of Human Services.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *Propublica.*

Antonovics, Kate, and Brian G Knight. 2009. "A new look at racial profiling: Evidence from the Boston Police Department." *The Review of Economics and Statistics* 91 (1): 163–177.

Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring racial discrimination in algorithms." In *AEA Papers and Proceedings,* 111:49–54.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics* 133 (4): 1885–1932.

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova. 2020. "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–12.

Baron, E Jason, Ezra G Goldstein, and Joseph Ryan. 2021. "The Push for Racial Equity in Child Welfare: Can Blind Removals Reduce Disproportionality?" *Available at SSRN 3947210.*

Barry Jester, Anna Maria, Ben Casselman, and Dana Goldstein. 2015. "The New Science of Sentencing." *The Marshall Project.*

Berger, Lawrence M, Sarah A Font, Kristen S Slack, and Jane Waldfogel. 2017. "Income and child maltreatment in unmarried families: Evidence from the earned income tax credit." *Review of Economics of the Household* 15 (4): 1345–1372.

Cancian, Maria, Mi-Youn Yang, and Kristen Shook Slack. 2013. "The effect of additional child support income on the risk of child maltreatment." *Social Service Review* 87 (3): 417–437.

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. "How child welfare workers reduce racial disparities in algorithmic decisions." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 1–22.

Chohlas-Wood, Alex. 2020. *Understanding the Child Welfare System in California: A Primer for Service Providers and Policymakers.* Technical report. Brookings Institution.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." In *Conference on Fairness, Accountability and Transparency,* 134–148. PMLR.

Drake, Brett, Jennifer M Jolley, Paul Lanier, John Fluke, Richard P Barth, and Melissa Jonson-Reid. 2011. "Racial bias in child protection? A comparison of competing explanations using national data." *Pediatrics* 127 (3): 471–478.

Drake, Brett, Melissa Jonson-Reid, Hyunil Kim, Chien-Jen Chiang, and Daji Davalishvili. 2021. "Disproportionate need as a factor explaining racial disproportionality in the CW system." In *Racial disproportionality and disparities in the child welfare system,* 159–176. Springer.

Gateway, Child Welfare Information. 2021. *Child Welfare Practice to Address Racial Disproportionality and Disparity.* Technical report. Children's Bureau.

Goldhaber-Fiebert, Jeremy D., and Lea Prince. 2019. *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office.* Technical report. Allegheny County Analytics, March.

Goncalves, Felipe, and Steven Mello. 2021. "A few bad apples? Racial bias in policing." *American Economic Review* 111 (5): 1406–41.

Grimon, Marie Pascale, and Christopher Mills. 2022. "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial." *Working Paper.*

Hampton, Robert L, and Eli H Newberger. 1985. "Child abuse incidence and reporting by hospitals: significance of severity, class, and race." *American Journal of Public Health* 75 (1): 56–60.

Ho, Sally, and Garance Burke. 2022a. "An algorithm that screens for child neglect raises concerns." *Associated Press* (April 2, 2022). https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1.

———. 2022b. "Oregon dropping AI tool used in child abuse cases." *Associated Press* (June 2, 2022). https://apnews.com/article/politics-technology-pennsylvania-child-abuse-1ea160dc5c2c203fdab456e3c2d97930.

Howell, Sabrina T, Theresa Kuchler, David Snitkof, Johannes Stroebel, and Jun Wong. 2021. *Racial disparities in access to small business credit: Evidence from the paycheck protection program.* Technical report. National Bureau of Economic Research.

Jenny, Carole, Kent P Hymel, Alene Ritzen, Steven E Reinert, and Thomas C Hay. 1999. "Analysis of missed cases of abusive head trauma." *Jama* 281 (7): 621–626.

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, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The quarterly journal of economics* 133 (1): 237–293.

Kovski, Nicole L, Heather D Hill, Stephen J Mooney, Frederick P Rivara, and Ali Rowhani-Rahbar. 2022. "Short-Term Effects of Tax Credits on Rates of Child Maltreatment Reports in the United States." *Pediatrics.*

Lane, Wendy G, David M Rubin, Ragin Monteith, and Cindy W Christian. 2002. "Racial differences in the evaluation of pediatric fractures for physical abuse." *Jama* 288 (13): 1603–1609.

Maloney, Tim, Nan Jiang, Emily Putnam-Hornstein, Erin Dalton, and Rhema Vaithianathan. 2017. "Black–White differences in child maltreatment reports and foster care placements: A statistical decomposition using linked administrative data." *Maternal and child health journal* 21 (3): 414–420.

New York State Office of Children and Family Services. 2020. *Administrative Directive: The Blind Removal Process.* https://www.acf.hhs.gov/cb/research-data-technology/statistics-research/child-maltreatment.

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.

Price, W. Nicholas. 2019. *Risks and remedies for artificial intelligence in health care.* Technical report. Brookings Institution.

Programs, Casey Family. 2021. *How Did the Blind Removal Process in Nassau County, NY Address Disparity Among Children Entering Care?* Technical report. Casey Family Programs.

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.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In *Proceedings of the 2020 conference on fairness, accountability, and transparency,* 469–481.

Raissian, Kerri M, and Lindsey Rose Bullinger. 2017. "Money matters: Does the minimum wage affect child maltreatment rates?" *Children and youth services review* 72:60–70.

Rehavi, M Marit, and Sonja B Starr. 2014. "Racial disparity in federal criminal sentences." *Journal of Political Economy* 122 (6): 1320–1354.

Rittenhouse, Katherine. 2023. "Income and Child Maltreatment: Evidence from a Discontinuity in Tax Benefits." *SSRN Working Paper.*

Stevenson, Megan T, and Jennifer L Doleac. 2021. "Algorithmic risk assessment in the hands of humans." *Available at SSRN 3489440.*

Thomas, Krista, and Charlotte Halbert. 2021. *Transforming Child Welfare: Prioritizing Prevention, Racial Equity, and Advancing Child and Family Well-Being*. Technical report. National Council on Family Relations.

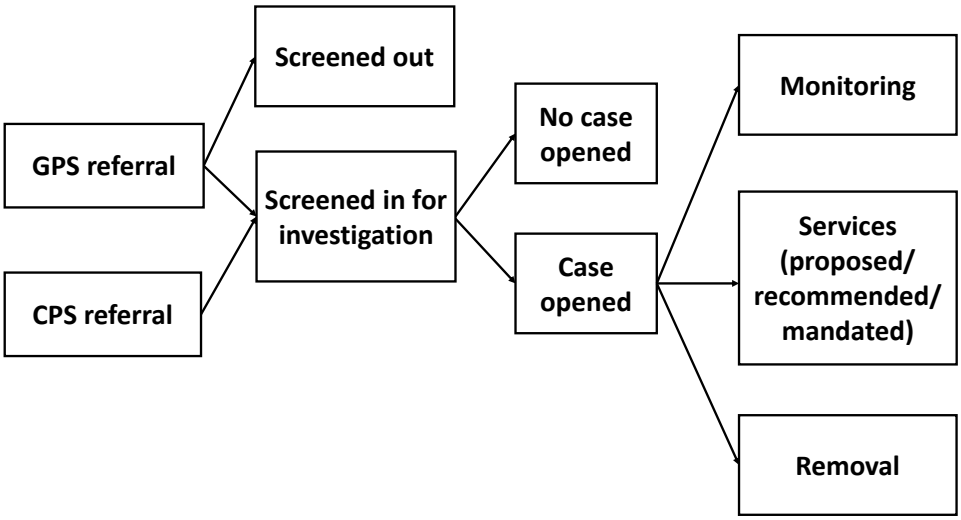
Vaithianathan, Rhema, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. 2019. *Allegheny Family Screening Tool: Methodology, Version 2.* Technical report. Centre for Social Data Analytics.

Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation.* Technical report. Centre for Social Data Analytics.

Wildeman, Christopher, and Natalia Emanuel. 2014. "Cumulative risks of foster care placement by age 18 for US children, 2000–2011." *PloS one* 9 (3): e92785.

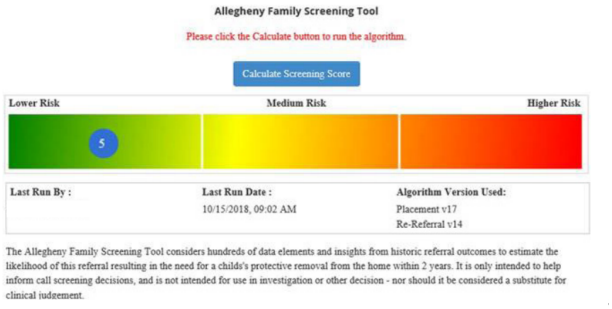
Figures

Figure 1: Referral process

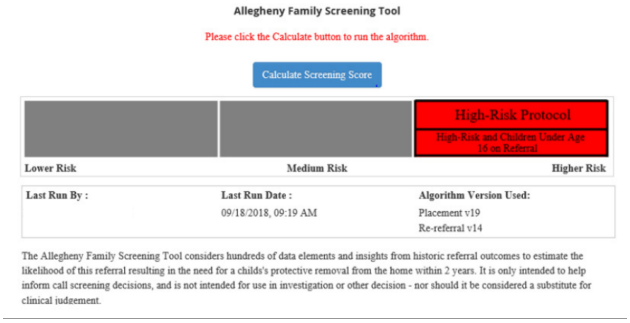


This figure presents the steps by which referrals to CYF move through the child welfare system in Allegheny County, PA.

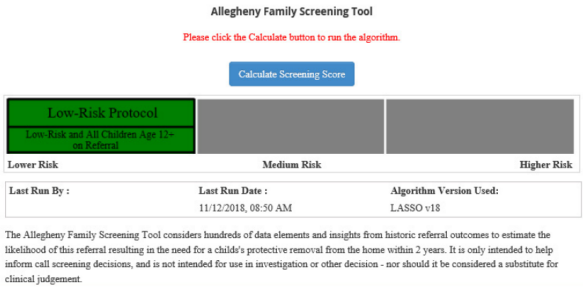
Figure 2: Screener View of AFST Output



(a) No protocol



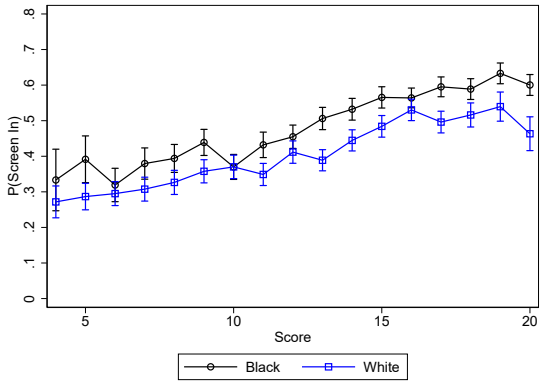
(b) High-risk protocol



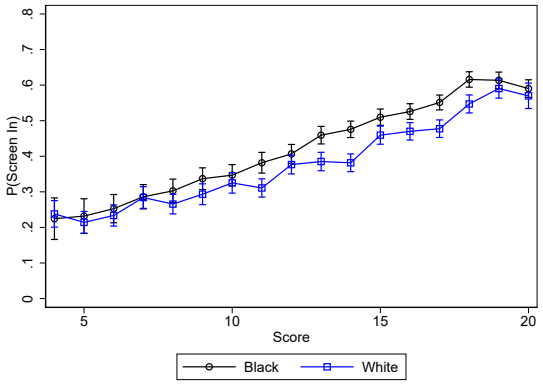
(c) Low-risk protocol

This figure presents the view that a screener has after running the algorithm, in three different scenarios. Panel (a) shows the output when neither the high-risk protocol nor the low-risk protocol is in place. Panel (b) shows the screener's view if the high-risk protocol is implemented (i.e. any child under age 16 and at least one score above 17). Panel (c) shows the screener's view if a low-risk protocol is implemented (i.e. all children are above a given age and all scores are below a given cutoff – the exact age and score cutoffs have changed over time).

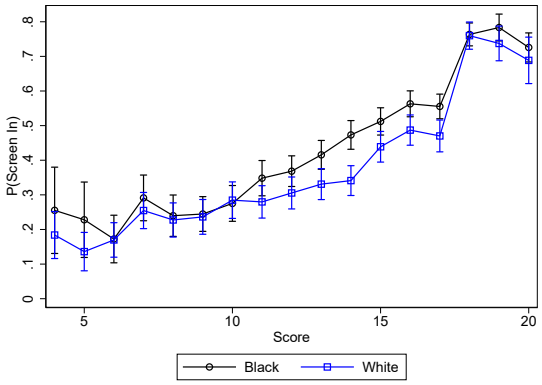
Figure 3: Screen In Disparities



(a) Pre-deployment (V3 Score)



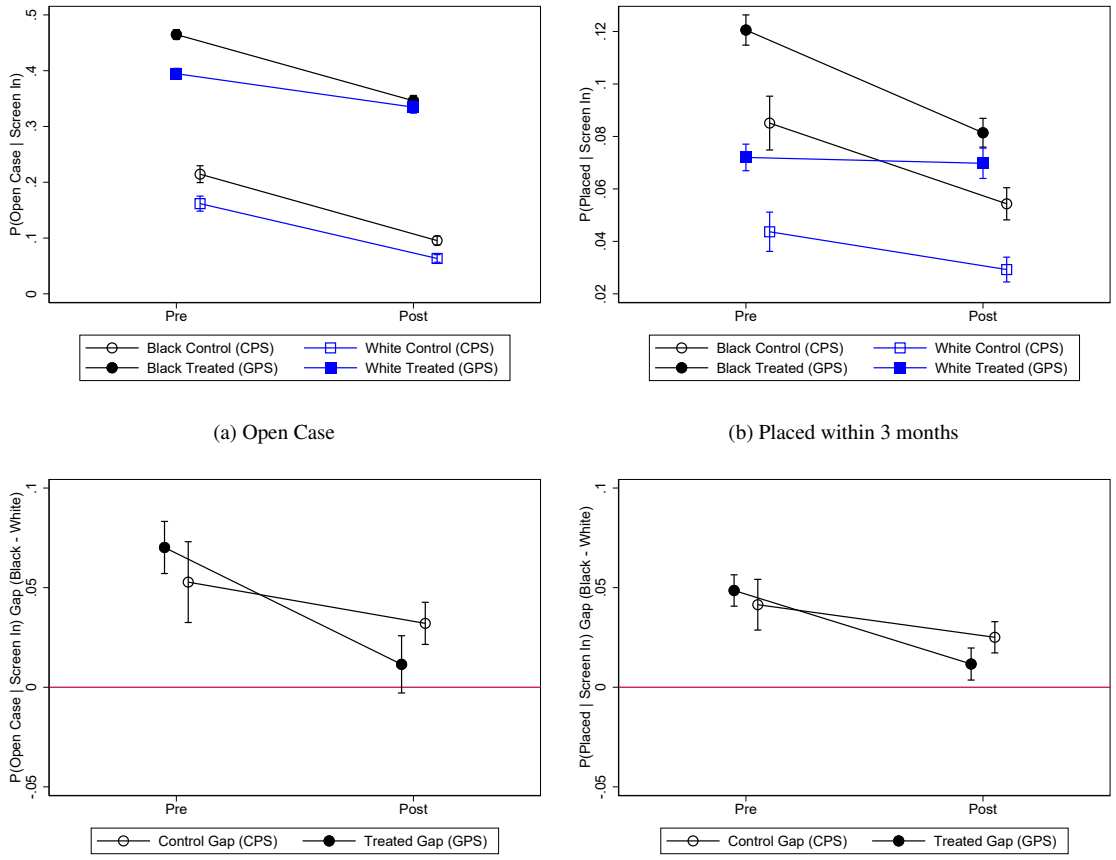
(b) Post-deployment (V3 Score)



(c) Post-deployment AFST V3 (Observed Score = V3 Score)

This figure reports average screen-in rates by algorithm score and race. Screen-in rate is defined as the share of referrals which are screened in for an investigation. See the notes to Table 1 for a description of the analysis sample. The sample is further restricted to referrals made in or after January 2013, due to limited availability of retroactively calculated scores. For each score and race, 95% confidence intervals are shown. Each sub-figure presents results from a different time period. Panel (a) presents screen-in rates from the period before the algorithm was implemented, from Jan. 2013 through Jun. 2016. The algorithm score in this panel was retroactively calculated, using AFST V3. Panel (b) presents screen-in rates from the period after the algorithm was implemented, from July 2016 through Dec. 2020. The algorithm score in this panel was also calculated using AFST V3. Panel (c) presents screen-in rates from the period after AFST V3 was implemented, i.e. for July 2019 through Dec. 2020. The algorithm score in this panel uses the AFST V3 score as seen by the call screeners.

Figure 4: Downstream Disparities



(a) Open Case

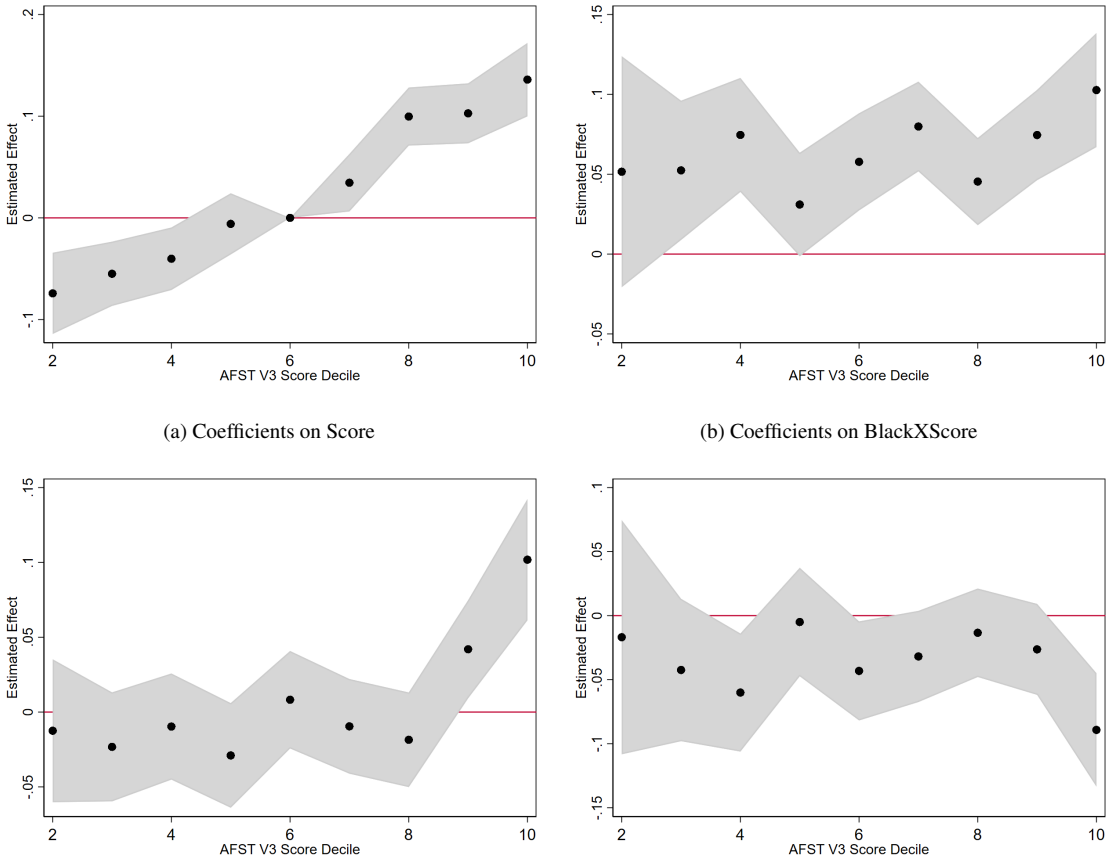
(b) Placed within 3 months

(c) Disparity - Open Case

(d) Disparity - Placed within 3 months

This figure presents our main outcome variables of interest, across time period, treatment/control group, and race of children on the referral. See the notes to Table 1 for a description of the analysis sample. The sample is further restricted to referrals made between January 2015 and September 2020 (due to expungement of CPS data and to allow for a 3 month follow up for each referral) Panel (a) reports the share of referrals which result in an open case, separately across time period, treatment/control group, and race. Panel (b) reports the share of referrals which involve a child who is removed from their home within three months of the screening decision, again separately across time period, treatment/control group, and race. Panel (c) shows the Black-white gap in case opening rates, separately across time periods and treatment/control groups, and Panel (d) shows the Black-white gap in 3-month removal rates, separately across time periods and treatment/control groups. For all data points, 95% confidence intervals are shown.

Figure 5: Screening by Score



(a) Coefficients on Score

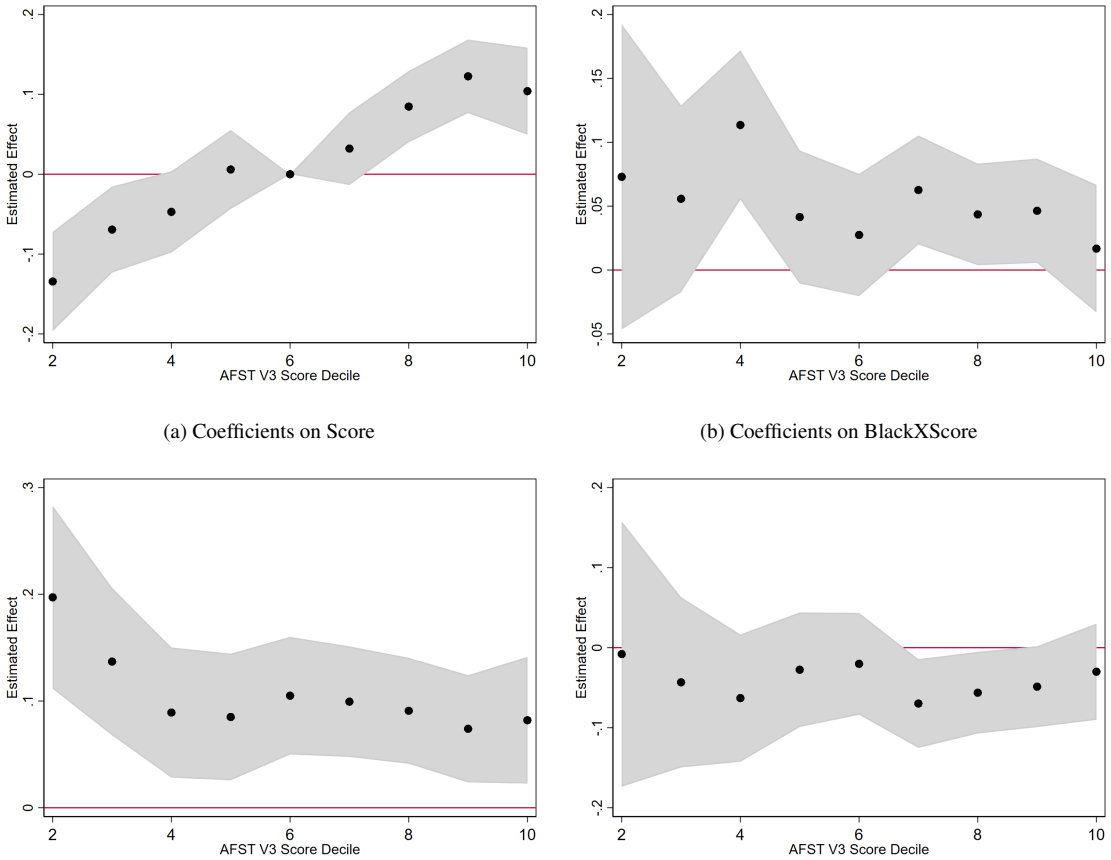
(b) Coefficients on BlackXScore

(c) Coefficients on PostXScore

(d) Coefficients on BlackXPostXScore

This figure graphically presents coefficients and 95% confidence intervals from estimating Equation 2, where the outcome variable is equal to one if the referral is screened-in, and zero otherwise. See notes to Table 1 for a description of the analysis sample. The sample is further restricted to referrals made in or after January 2013, when AFST V3 scores are available.

Figure 6: Open Case by Score



(a) Coefficients on Score

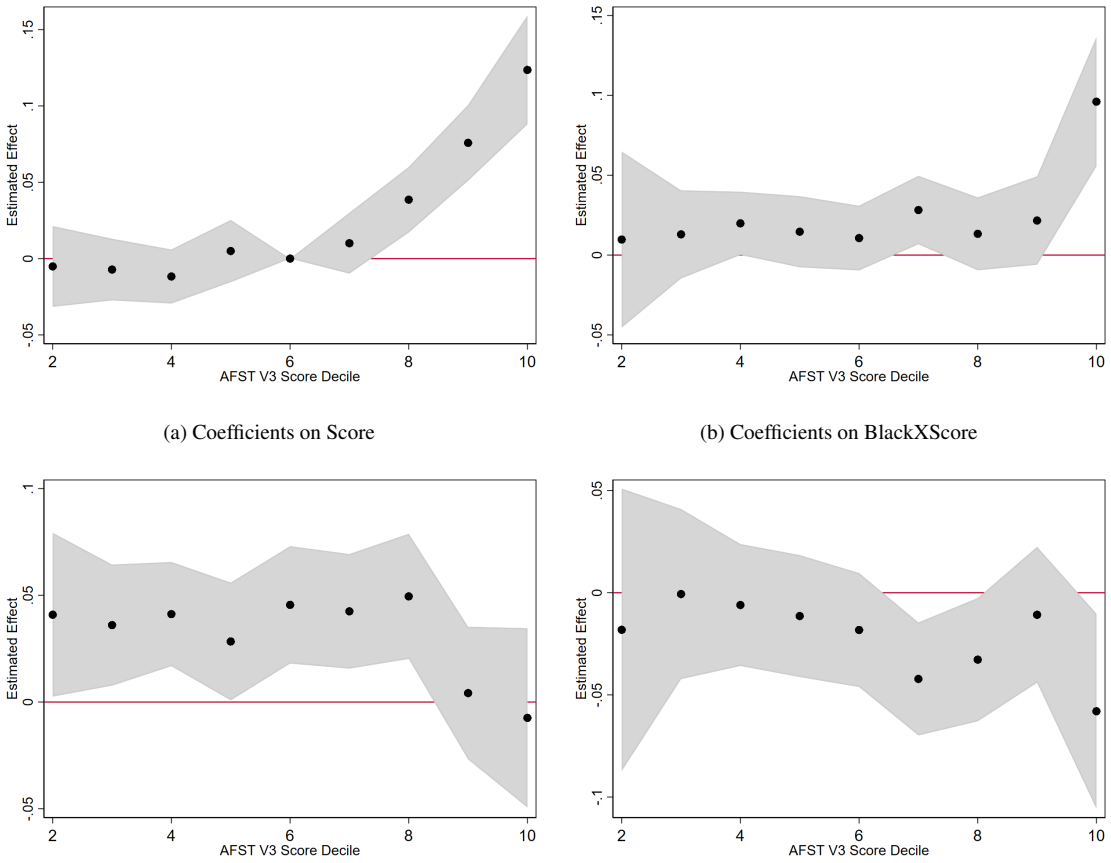
(b) Coefficients on BlackXScore

(c) Coefficients on PostXScore

(d) Coefficients on BlackXPostXScore

This figure graphically presents coefficients and 95% confidence intervals from estimating Equation 2, where the outcome variable is equal to one if the referral results in a case opening, and zero otherwise. See notes to Table 1 for a description of the analysis sample. The sample is further restricted to referrals made in or after January 2013, when AFST V3 scores are available.

Figure 7: Results: Removal (3m) by Score



(a) Coefficients on Score

(b) Coefficients on BlackXScore

(c) Coefficients on PostXScore

(d) Coefficients on BlackXPostXScore

This figure graphically presents coefficients and 95% confidence intervals from estimating Equation 2, where the outcome variable is equal to one if any child associated with the referral is placed within 3 months, and zero otherwise. See notes to Table 1 for a description of the analysis sample. The sample is further restricted to referrals made before October 2020, to allow for a 3 month follow up, and to referrals made in or after January 2013, when AFST V3 scores are available.

Tables

Table 1: Summary Statistics

	All GPS Mean	***Screened-in GPS*** Mean	***All CPS*** Mean
Panel A: Demographics			
Black	0.501	0.556	0.511
Any Infant	0.149	0.223	0.116
Any Child Age 1-5	0.452	0.477	0.443
Any Child Age 6-12	0.584	0.586	0.632
Any Child Age 13-17	0.391	0.394	0.443
Number of Children	2.261	2.409	2.379
Panel B: Referral Outcomes			
Screened In	0.443	1.000	0.997
Case Opened	0.172	0.388	0.117
Removed within 3 months	0.051	0.088	0.050
Panel C: Allegation Categories			
Abandonment	0.010	0.012	0.003
Caregiver Behavioral Issues	0.040	0.053	0.013
Caregiver Substance Abuse	0.219	0.295	0.036
Causing Death of Child	0.004	0.005	0.006
Child Behaviors	0.059	0.065	0.022
Domestic Violence	0.070	0.085	0.026
Exposure to Risk	0.109	0.117	0.024
Failure to Protect	0.054	0.052	0.014
Imminent Risk	0.034	0.038	0.014
Inadequate Physical Care	0.280	0.254	0.031
Medical Neglect	0.037	0.049	0.016
Mental Health	0.059	0.048	0.024
Mental Injuries	0.026	0.032	0.021
Neglect	0.108	0.108	0.015
No/Inadequate Home	0.103	0.133	0.012
Parent/Child Conflict	0.047	0.041	0.018
Physical Altercation	0.014	0.015	0.024
Physical Maltreatment	0.101	0.054	0.656
Sexual Abuse or Exploitation	0.020	0.011	0.180
Sexual Contact Between Children	0.040	0.024	0.004
Truancy	0.045	0.069	0.006
Unwilling or Unable to Provide Care	0.065	0.073	0.011
Youth Substance Abuse	0.012	0.011	0.003
Panel D: Reporter Categories			
Agency	0.227	0.247	0.280
Anonymous	0.120	0.087	0.032
Community	0.052	0.044	0.025
Family	0.167	0.157	0.070
Law	0.116	0.179	0.094
Medical	0.090	0.101	0.152
School	0.131	0.120	0.150
Self	0.004	0.003	0.010
Therapist	0.086	0.053	0.185
N	91203	40405	16366

This table reports means of referral characteristics separately for all GPS, screened-in GPS, and all CPS referrals. Note, since 100% of CPS referrals are screened in, averages for screened-in CPS referrals are identical to those in the full CPS sample. The sample is all referrals made between Jan. 2010 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor white children, and referrals stemming from truancy courts. The variable *Removed within 3 months* is set to missing for referrals made after September 2020, in order to allow for a three month follow-up period for each referral.

Table 2: Screen In

	(1) DD	(2) +FE	(3) +Controls	(4) Restricted Sample	(5) +Risk Score
Post	-0.0122***	0.0129	0.00286	0.00476	0.00552
	(0.00460)	(0.0118)	(0.0107)	(0.0110)	(0.0110)
Black	0.109***	0.108***	0.0848***	0.0869***	0.0589***
	(0.00446)	(0.00446)	(0.00419)	(0.00558)	(0.00563)
BlackXPost	-0.0207***	-0.0203***	-0.0225***	-0.0258***	-0.0269***
	(0.00657)	(0.00656)	(0.00606)	(0.00705)	(0.00705)
Year FE	No	Yes	Yes	Yes	Yes
Month-of-Year FE	No	Yes	Yes	Yes	Yes
Score FE	No	No	No	No	Yes
Controls	No	No	Yes	Yes	Yes
Data Span	2010 - 2020	2010 - 2020	2010 - 2020	2013 - 2020	2013 - 2020
Mean	0.443	0.443	0.443	0.442	0.447
Obs.	91203	91203	91203	69751	68510

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reports coefficients and standard errors from estimating four specifications of Equation 2. The sample is described in the notes to Table 1. In Columns 4 and 5, the sample is further restricted to referrals made in or after January 2013, when AFST V3 scores are available. The outcome variable in each regression is an indicator equal to one for referrals which are screened in, and zero otherwise.

Table 3: Open Case | Screen In

	(1) DD	(2) DD Restricted Sample	(3) DDD Base	(4) +FE	(5) +Controls
Post	0.0884***	0.0908***	-0.00883	0.0484***	0.0396***
	(0.0177)	(0.0190)	(0.00686)	(0.0127)	(0.0126)
Black	0.0593***	0.0608***	0.0323***	0.0331***	0.0268***
	(0.00657)	(0.0125)	(0.00933)	(0.00934)	(0.00930)
BlackXPost	-0.0518***	-0.0524***	0.000127	0.000369	0.00123
	(0.00946)	(0.0142)	(0.0107)	(0.0107)	(0.0107)
GPS			0.237***	0.236***	0.130***
			(0.0108)	(0.0108)	(0.0128)
BlackXGPS			0.0346**	0.0334**	0.0345**
			(0.0157)	(0.0157)	(0.0155)
GPSXPost			0.0267**	0.0304**	0.0368***
			(0.0125)	(0.0125)	(0.0124)
BlackXPostXGPS			-0.0569***	-0.0531***	-0.0559***
			(0.0180)	(0.0180)	(0.0178)
Year FE	Yes	Yes	No	Yes	Yes
Month-of-Year FE	Yes	Yes	No	Yes	Yes
Controls	Yes	Yes	No	No	Yes
Data Span	2010 - 2020	2015 - 2020	2015 - 2020	2015 - 2020	2015 - 2020
Mean	0.388	0.335	0.239	0.239	0.240
Obs.	40405	23754	38044	38044	37925

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reports coefficients and standard errors from five separate regressions estimating Equations 1 (Columns (1) and (2)) and different specifications of Equation 3 (Columns (3)-(5)). The sample is described in the notes to Table 1. The sample is further restricted to screened-in referrals. In Columns (2) - (5), the sample is further restricted to referrals made in or after January 2015. The outcome variable in each regression is an indicator equal to one for referrals which result in a case being opened, and zero otherwise. Controls include allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol exposure.

Table 4: Placed within 3 months | Screen In

	(1) DD	(2) Restricted Sample	(3) DDD Base	(4) +FE	(5) +Controls
Post	0.0295***	0.0322***	0.00293	0.0295***	0.0255***
	(0.0103)	(0.0110)	(0.00443)	(0.00779)	(0.00766)
Black	0.0391***	0.0427***	0.0289***	0.0291***	0.0278***
	(0.00386)	(0.00729)	(0.00652)	(0.00652)	(0.00636)
BlackXPost	-0.0313***	-0.0334***	-0.00379	-0.00410	-0.00314
	(0.00558)	(0.00830)	(0.00763)	(0.00762)	(0.00745)
GPS			0.0343***	0.0345***	0.0181**
			(0.00596)	(0.00596)	(0.00713)
BlackXGPS			0.0194**	0.0185*	0.0147
			(0.00983)	(0.00982)	(0.00961)
GPSXPost			0.00622	0.00616	0.00777
			(0.00707)	(0.00708)	(0.00703)
BlackXPostXGPS			-0.0328***	-0.0313***	-0.0309***
			(0.0114)	(0.0113)	(0.0111)
Year FE	Yes	Yes	No	Yes	Yes
Month-of-Year FE	Yes	Yes	No	Yes	Yes
Controls	Yes	Yes	No	No	Yes
Data Span	2010 - 2020	2015 - 2020	2015 - 2020	2015 - 2020	2015 - 2020
Mean	0.0890	0.0788	0.0648	0.0648	0.0643
Obs.	39264	22613	36419	36419	36300

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reports coefficients and standard errors from five separate regressions estimating Equations 1 (Columns (1) and (2)) and different specifications of Equation 3 (Columns (3)-(5)). The sample is described in the notes to Table 1. The sample is further restricted to screened-in referrals, and referrals made before October 2020 to allow for a 3 month follow up for each referral. In Columns (2) - (5), the sample is further restricted to referrals made in or after January 2015. The outcome variable in each regression is an indicator equal to one for referrals which are associated with any child who is removed within 3 months of the referral date, and zero otherwise. Controls include allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol exposure.

Table 5: Difference-in-Differences by Race

	(1) Case Opened (Black)	(2) Case Opened (White)	(3) Removal (Black)	(4) Removal (White)
Post	0.0620***	0.0172	0.0256**	0.0198**
	(0.0171)	(0.0162)	(0.0113)	(0.00922)
GPS	0.162***	0.127***	0.0326***	0.0127*
	(0.0147)	(0.0147)	(0.00958)	(0.00758)
GPSXPost	-0.0156	0.0349***	-0.0190**	0.00602
	(0.0129)	(0.0125)	(0.00875)	(0.00703)
Year FE	Yes	Yes	Yes	Yes
Month-of-Year FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Data Span	2015 - 2020	2015 - 2020	2015 - 2020	2015 - 2020
Mean	0.258	0.219	0.0753	0.0515
Obs.	20438	17487	19509	16791

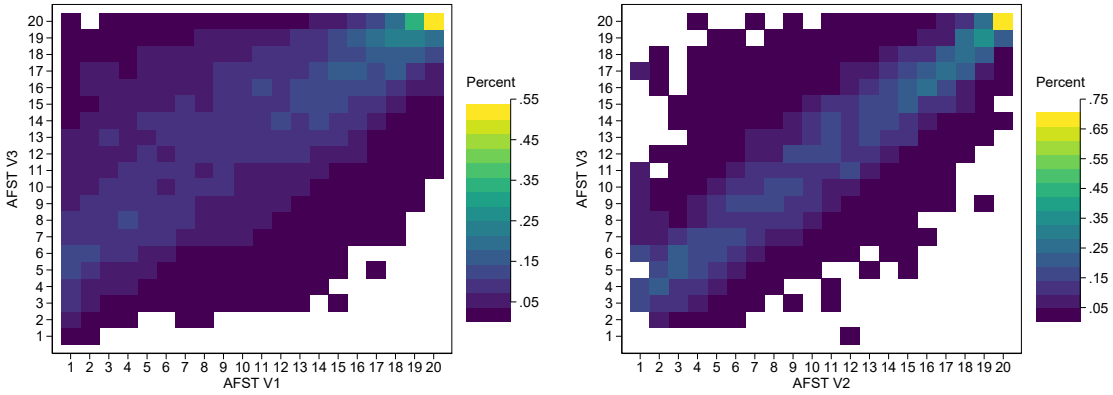
Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reports coefficients and standard errors from four separate regressions estimating Equation 4. The sample is described in the notes to Table 1. The sample is further restricted to screened-in referrals made in or after January 2015. In Columns (3) and (4) the sample is restricted to referrals made before October 2020 to allow for a 3 month follow up for each referral. In Columns (1) and (3) the sample is restricted to referrals involving Black children, and in Columns (2) and (4) the sample is restricted to referrals involving white children. The outcome variable in Columns (1) and (2) is an indicator equal to one for referrals which result in a case being opened, and zero otherwise. The outcome variable in Columns (3) and (4) is an indicator equal to one for referrals which are associated with any child who is removed within 3 months of the referral date, and zero otherwise. Controls include allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol exposure.

7 Appendix Figures and Tables

Figure A1: Correlation Across AFST Versions

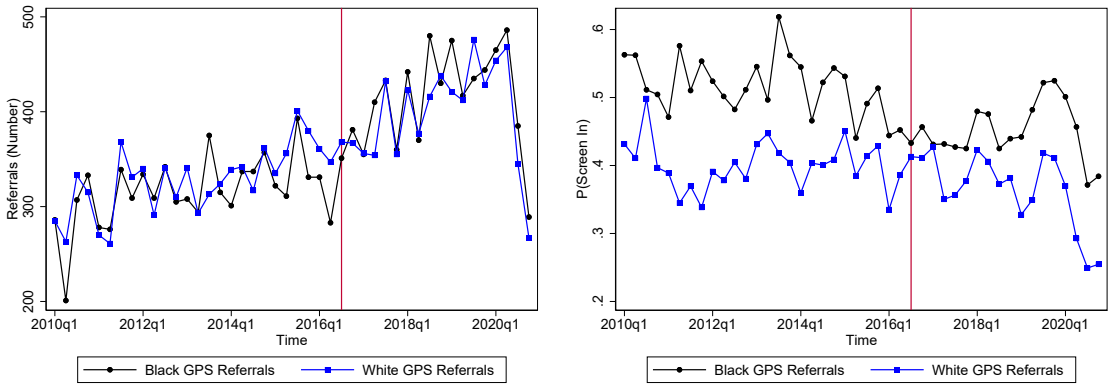


(a) V1-V3 Correlation

(b) V2-V3 Correlation

For each of AFST V1 (panel a) and AFST V2 (panel b), this figure shows the percent of each discrete score 1-20 which maps to each of AFST V3 score 1-20.

Figure A2: Time trends: Referrals and Screening decisions

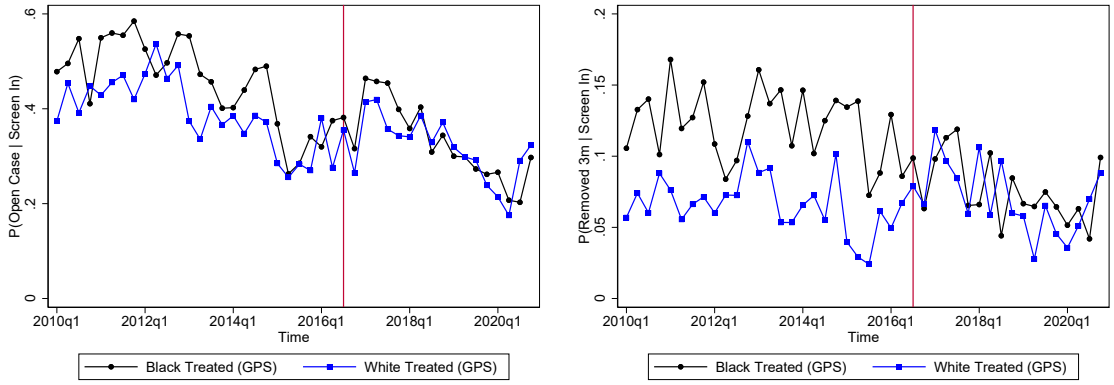


(a) Number of referrals

(b) Screening

This figure presents monthly numbers of referrals (in panel a) and quarterly average screen-in rates (in panel b) separately by race of referral (as defined in the text) for GPS referrals. See the notes to Table 1 for a description of the analysis sample.

Figure A3: Time trends: Downstream outcomes

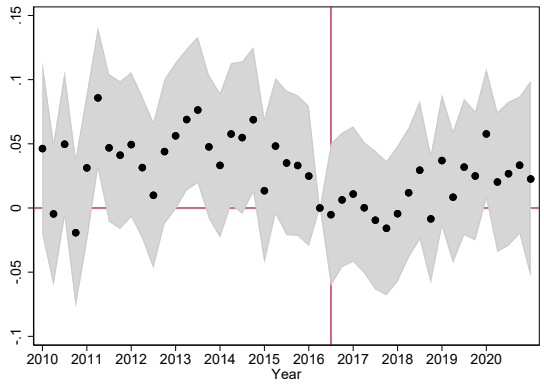


(a) Open Case | Screen In

(b) Removal | Screen In

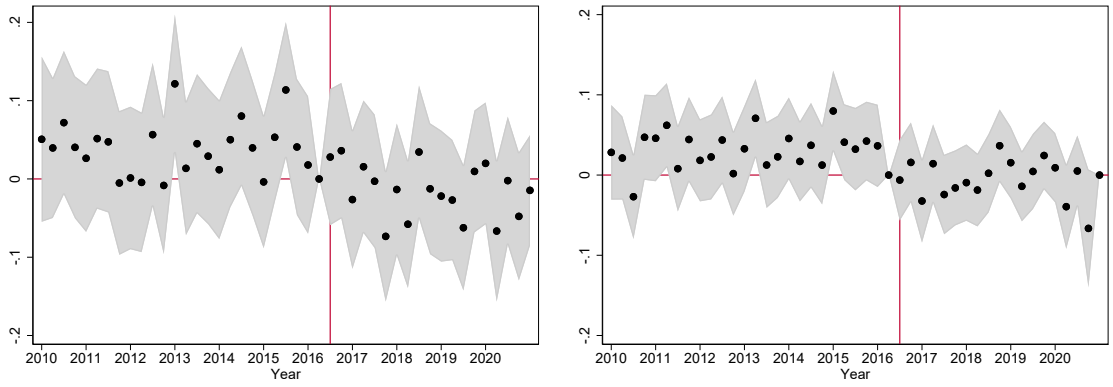
This figure presents monthly case opening (panel a) and removal (panel b) rates by race for screened-in GPS referrals. See the notes to Table 1 for a description of the analysis sample.

Figure A4: Event Study - Screening



This figure presents coefficients and 95% confidence intervals from estimating an event study version of Equation 1, where the outcome variable is an indicator equal to one if the referral was screened in and zero otherwise. Indicators for quarter relative to 2016Q3 are interacted with Pre_{it} and $Post_{it}$. 2016Q2 is the excluded quarter. See the notes to Table 1 for a description of the analysis sample.

Figure A5: Event Study - Downstream Diff-in-Diff

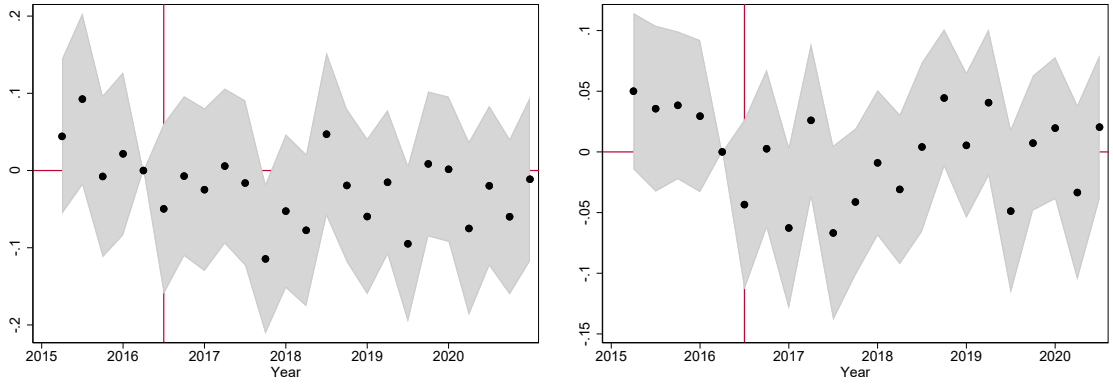


(a) Open Case | Screen In

(b) Removal within 3 months | Screen In

This figure presents coefficients and 95% confidence intervals from estimating an event study version of Equation 1, where the outcome variable is either an indicator equal to one if the screened-in referral had a case opened and zero otherwise (panel a) or an indicator equal to one if the screened-in referral had any child removed from their home within 3 months, and zero otherwise (panel b). Indicators for quarter relative to 2016Q3 are interacted with Pre_{it} and $Post_{it}$. 2016Q2 is the excluded quarter. See the notes to Table 1 for a description of the analysis sample. In panel (b) the sample is further restricted to referrals made before October 2020, to allow for a 3 month follow up for each referral.

Figure A6: Event Study - Downstream DDD



(a) Open Case | Screen In

(b) Removal within 3 months | Screen In

This figure presents coefficients and 95% confidence intervals from estimating an event study version of Equation 3, where the outcome variable is either an indicator equal to one if the screened-in referral had a case opened and zero otherwise (panel a) or an indicator equal to one if the screened-in referral had any child removed from their home within 3 months, and zero otherwise (panel b). Indicators for quarter relative to 2016Q3 are interacted with Pre_{it} and $Post_{it}$. 2016Q2 is the excluded quarter. See the notes to Table 1 for a description of the analysis sample. The sample is further restricted to screened-in referrals made in or after 2015. In panel (b) the sample is further restricted to referrals made before October 2020, to allow for a 3 month follow up for each referral.

Table A1: Results: Placed within 3 months (Unconditional)

	(1) DD	(2) Restricted Sample	(3) DDD Base	(4) +FE	(5) +Controls
Post	0.0151***	0.0175***	0.00299	0.0220***	0.0185***
	(0.00519)	(0.00545)	(0.00442)	(0.00599)	(0.00588)
Black	0.0372***	0.0359***	0.0294***	0.0296***	0.0263***
	(0.00210)	(0.00382)	(0.00654)	(0.00653)	(0.00633)
BlackXPost	-0.0200***	-0.0189***	-0.00441	-0.00455	-0.00282
	(0.00291)	(0.00430)	(0.00763)	(0.00763)	(0.00743)
GPS			0.00434	0.00443	-0.00834*
			(0.00427)	(0.00426)	(0.00444)
BlackXGPS			0.00961	0.00910	0.00932
			(0.00758)	(0.00758)	(0.00737)
GPSXPost			-0.000990	-0.000385	0.00112
			(0.00506)	(0.00507)	(0.00499)
BlackXPostXGPS			-0.0150*	-0.0145*	-0.0162*
			(0.00878)	(0.00878)	(0.00857)
Year FE	Yes	Yes	No	Yes	Yes
Month-of-Year FE	Yes	Yes	No	Yes	Yes
Controls	Yes	Yes	No	No	Yes
Data Span	2010 - 2020	2015 - 2020	2015 - 2020	2015 - 2020	2015 - 2020
Mean	0.0520	0.0444	0.0438	0.0438	0.0435
Obs.	88901	52575	66394	66394	66266

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

This table reports coefficients and standard errors from five separate regressions estimating Equations 1 (Columns (1) and (2)) and different specifications of Equation 3 (Columns (3)-(5)). The sample is described in the notes to Table 1. The sample is further restricted to referrals made before October 2020 to allow for a 3 month follow up for each referral. In Columns (2) - (5), the sample is further restricted to referrals made in or after January 2015. The outcome variable in each regression is an indicator equal to one for referrals which are associated with any child who is removed within 3 months of the referral date (regardless of whether the referral resulted in an investigation), and zero otherwise. Controls include allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol exposure.