ALGORITHMS, HUMANS AND RACIAL DISPARITIES IN CHILD PROTECTIVE SERVICES: 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

September 21, 2022

Abstract

We ask whether providing decision-makers with a machine learning tool 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 abuse or neglect to investigate. While the AFST does not dictate investigation decisions, referrals with the highest risk scores are "defaulted" to be screened in. Among this group of referrals, we find that the AFST reduced disparities in investigation rates. Using a triple difference strategy, we also find that the introduction of the AFST significantly reduced disparities in 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, 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 conclusion 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 (often referred to by the term "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, child protective services. However, as their prevalence grows, so too do concerns that algorithms may entrench or exacerbate existing disparities in system interactions, and in particular disparities across racial lines.

Several high-profile media reports have drawn attention to the potential for algorithms to discriminate.[1] 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)).[2]

Clearly, algorithms have the potential to be biased, in particular if they are trained on data from a biased system. However, human bias is also well-established, and has been shown to cause racial disparities in many of our society's institutions, including at every stage of the criminal justice system.[3] An important policy question, then, is whether the introduction of a predictive risk model increases or decreases disparities, relative to the relevant counterfactual. Previous work addressing this question is limited, and has found mixed results.[4]

This paper will be the first to analyze the effects of an algorithm on racial disparities in the context of child

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

[2] Obermeyer et al. (2019) study an algorithm which is widely used in healthcare systems to identify patients who have a high risk of serious health issues and is used by providers to identify patients to prioritize for preventive service. The authors find that, for Black patients and White patients with the same algorithm-predicted risk, Black patients are significantly more sick (when using directly observed and objective measures of morbidity) than White patients. These disparities mean that White patients who have lower "objectively" measured risk of poor health will be prioritized for prevention over Black patients. The authors identify the source of this bias in the objective function of the algorithm, which uses medical expenditures as a proxy for health needs. However, while health care costs and health needs are highly correlated, racial disparities in access to care mean that for the same level of underlying health need, medical expenditures are higher for White patients than for Black patients. The authors do not argue that this problem justifies getting rid of the predictive risk model entirely, but rather advocate for a deeper understanding of the biases baked into such models by the data it is fed. In fact, by simply changing the outcome variable to a combination of predicted costs and predicted health, bias was reduced in the algorithm by 84%. In the criminal justice system, predictive risk models are used to help inform decisions from setting pre-trial bail to parole. Arnold, Dobbie, and Hull (2021) propose a method to measure algorithmic discrimination (defined as differential treatment of equally qualified individuals), and apply that method to the context of pre-trial release decisions in New York City. They find that the algorithm does discriminate against Black defendants, and in particular suggests releasing White defendants at an eight percentage point higher rate than Black defendants with the same potential for pretrial misconduct.

[3] Police officers are less lenient towards Black drivers (Goncalves and Mello 2021), and are more likely to search drivers whose race differs from their own (Antonovics and Knight 2009). Rehavi and Starr (2014) find that, holding criminal history and arrest offense constant, Black defendants are more likely to be charged of a serious crime. Arnold, Dobbie, and Yang (2018) find that marginal White defendants who are released pre-trial are more likely to commit pre-trial misconduct than marginally-released Black defendants. The authors take this disparity in outcomes as evidence that judges are biased against Black defendants - if there was no bias we would expect outcomes for these marginal defendants to be equal across races. Abrams, Bertrand, and Mullainathan (2012) find heterogeneity in how randomly-assigned judges differentially incarcerate Black vs. White defendants. They take this as evidence of bias, noting that in the absence of bias we may expect randomly-assigned judges to differentially incarcerate Black and White defendants, but we would not expect that difference to vary across judges.

[4] Stevenson and Doleac (2021) study the adoption of algorithmic risk assessments on judge decisions in felony sentencing in Virginia, and 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.

protective services, an institution which interacts with approximately one third of U.S. children by the time they reach 18 (Kim et al. 2017). It will add 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.

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. In particular, this algorithm aims to help intake workers screen referrals alleging that a child is being maltreated. Allegheny County receives hundreds of these referrals every week. Prior to the introduction of the algorithm, call screening staff used their professional judgement to decide which referrals to "screen in" for an investigation. The predictive risk model was introduced with the goals of decreasing both false negatives (undetected cases of serious or chronic maltreatment) and false positives (investigation of non-existent maltreatment). The algorithm uses data about the referred families from the County's data warehouse to predict the risk that the child will be removed from their home if screened in (a proxy for serious abuse or neglect), and shows the human decision-maker a risk score ranging from 1 to 20. For referrals with the highest risk scores (above 17), the AFST implements a "high-risk protocol," defaulting to a screen-in decision. However,the call-screening staff remain the ultimate decision-maker, and may choose to override the protocol.

Racial disparities are well-documented in child protection systems across the U.S.[5] From reporting through screening, substantiation rates, placement in out-of-home care and length of stay in foster care, Black children are over-represented relative to White children at every stage of the process.[6] Overall, 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). While the existence of racial disparities is well-established, the causes for those disparities are not fully understood. In particular, 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.

The existence of racial disparities in the child protection system is an ongoing source of concern for the community. Reduction of those disparities is often cited as a policy goal by both policymakers and researchers. (See, for example, Gateway (2021) and Thomas and Halbert (2021).) This concern 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).

Several studies suggest that racial bias affects which families are referred for abuse or neglect. This body of

[5] We define "disparity" simply as observed differences across groups, regardless of the cause of those differences.
[6] See, for example, Administration on Children and Families (2021).

literature has focused on physicians' evaluations of potentially abused children, and found 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)).[7] Bartholet (2009) critiques this literature, noting that physician reporting studies often fail to account for predictive variables which correlate with race and are observable to physician decision-makers, but unobservable to researchers. Such omitted variable bias might provide an alternative explanation for disparities in reporting rates. Note also that these studies were all conducted over twenty years ago, and may not reflect the rates of bias in today's context.

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 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.[8] 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.[9] Importantly, these results do not preclude the possibility that racial bias could also affect disparities. Due to the ambiguity in the sources of racial disparities with child welfare systems, it is not a priori obvious whether and to what extent even an entirely unbiased algorithm would reduce those disparities.

In this paper, we ask an empirical rather than normative question: did the introduction of an algorithm at the "front-door" of the child protection system affect racial disparities in screening decisions and downstream outcomes? In particular, how did the implementation of the AFST affect: (1) the differential probability of investigating a Black vs. White child referred for maltreatment (2) the differential probability of opening a child welfare case for Black vs. White children; and (3) the differential probability of being removed from home for Black vs. White children? Note that this paper does not speak to the welfare effects of reducing disparities, or of the algorithm overall.

To answer these questions, we use a triple differences strategy, comparing referrals involving Black vs. White children, in the Pre-AFST vs. Post-AFST periods, for categories of referrals which were "treated" vs. not "treated" by

[7] Lane et al. (2002) study how race affects physicians' response when a child is hospitalized for fractures. They find that doctors are more likely to search for additional injuries by ordering a skeletal survey, and to report concerns to child protective services when the child is Black or Hispanic. These findings are consistent with either over-reporting of minority families, or under-reporting of White families. The authors conclude that "it is quite possible that cases of abuse were overlooked in White children because no [skeletal survey] was performed." Jenny et al. (1999) find that physicians are more likely to miss cases of abusive head trauma in White children than in minority children. In this particular example, since abuse is later confirmed, White children at high risk of severe abuse are likely under-served. Finally, Hampton and Newberger (1985) similarly found that hospitals were more likely to report suspected cases of abuse in minority families, than in White families.

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

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

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 group of referrals makes up our control group. This design enables us to control for any trends across time which might differentially affect Black vs. White families. We also study heterogeneity across algorithm scores, with a particular focus on effects for referrals which are likely to fall under the high-risk protocol.

Among all referrals, we find that the AFST had no significant effect on disparities in screening decisions. However, for referrals with the highest risk scores (19-20), we find that the AFST significantly reduced disparities in screening decisions by 9.6 percentage points, or 98% of the pre-existing gap, suggesting that the high-risk protocol in place for this group of referrals plays an important role. Turning to downstream outcomes, we find that among screened-in referrals, AFST reduced the disparity in case opening rates by 5.5 percentage points, (87% of the pre-existing gap) and reduced the disparity in 3-month removal rates by 2.9 percentage points (76% of the pre-existing gap).

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, PA is home to 1.2 million people, and includes the city of Pittsburgh within its boundaries. The Office of Children, Youth and Families (CYF) in Allegheny County is responsible for investigating allegations of child neglect and abuse. In Pennsylvania, child welfare is a State-supervised, county-administered system. All referrals of abuse or neglect are made through a State-administered web portal or by phone to ChildLine, a State-run 24-hour hotline service. The State of Pennsylvania categorizes each referral under either Child Protective Services (CPS) or General Protective Services (GPS), based on the type of allegation. 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.[10] After this classification is made by state staff, all referrals are sent to the County for further review and possible investigation. All CPS referrals are required to be investigated, while GPS referrals may be investigated at the discretion of the County.

[10]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.

2.1 Referral Process

Allegations of maltreatment are brought to the attention of the County by mandated reporters and community members, who make referrals to ChildLine.[11] Referrals that include allegations of abuse (as opposed to neglect) fall under the legal definition of CPS and must always be investigated. For the remainder of referrals (GPS), hotline staff (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.[12] 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. Based on their findings, the investigator and their supervisor then decide whether or not to open a case for services. An opened case might result in continued monitoring by a CYF worker, suggested or mandated participation in services, or, often as a last resort, a court-ordered removal of the child from the home. See Figure I for a depiction of how maltreatment referrals move through the child protection system in Allegheny County. Each decision point is defined and described in more detail below.

The screening decision determines whether the family will be subject to an investigation. Note that even if a referral has one victim child and multiple other children, it is the County's practice to consider all children residing in the same household as the alleged victim as "at risk" and assess each of them for risk and safety concerns.

For screened-in referrals, State law requires that the investigation is concluded within 60 days of receipt of the report. The investigation results in a case opening decision. In general, opening a case indicates that the family requires ongoing services or involvement from CYF social workers to ensure the safety of the child. Services might be ordered by the Court, be mandated under the threat of removal or Court activity or (less commonly) be truly voluntary. In some situations, a worker may open a case only to to ensure a family is able to access services. In other situations, opening a case may be an alternative to a foster care placement.

The placement decision involves a court order to remove a child from their home due to imminent and unresolved concerns for their 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

[11] See Table III for more details on reporters in our sample.

[12] 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.

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 CYF involvement, as well as involvement in a range of other County systems.[13] However, while these data were available and the County expected information to be systematically reviewed, there was little guidance for screeners on exactly how those data should be incorporated into their screening decision.[14] 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.[15] 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. 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 the screeners.[16] 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). Note, 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 II). 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 II).[17] These referrals are recommended to be screened out, but no supervisor approval is required to screen in these referrals.[18]

For referrals which do not meet the criteria for high- or low-risk protocols, the screener observes the numeric score,

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

[14] 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.

[15] 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.

[16] 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.

[17] Note, low-risk protocols made up only 4% of referrals at the time the latest protocol was implemented (Vaithianathan et al. 2017). Across the country approximately 55% of referrals are screened in for investigation (Administration on Children and Families 2021).

[18] 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.

and there is no screening decision recommendation (see Figure II).

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. Note, 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.[19] 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.[20]

Allegheny County Department of Human Services (DHS) developed the AFST with the purpose of using existing data to improve the quality and consistency of screening decisions.[21] 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 DHS 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.[22] 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 by DHS to conduct an independent impact evaluation of the original AFST, and studied effects on accuracy, workload, disparities, and consistency.

3 Data

We obtained de-identified administrative data from Allegheny County on the universe of child maltreatment referrals (both CPS and GPS) referred via ChildLine between 2015 and 2020. Every referral links to unique IDs for each person

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

[20] 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.

[21] 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.

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

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 test the robustness of our results to different time frames. 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 2% of child-referrals and are excluded from our analysis sample.

For each referral, we observe both the score generated by the most recent version of the model ("comparable score") as well as a score from the version of the model in use at the time ("seen score"). Going forward, we primarily focus on the score from the most recent iteration of the model in order to enhance comparability across years. For all referrals prior to July 2019, this comparable score was retroactively calculated, and can be thought of as the score that the screener *would have* seen, had the most recent version of the algorithm been deployed.[23] The "seen score" was generated by the AFST version deployed at the time of the referral, and as such is the score that was seen by the screener at the time of the screening decision. Note, for all referrals made after July 2019, the comparable score is exactly equal to the seen score. For all referrals made prior to the implementation of the AFST, there is no seen score, but only the retroactively calculated comparable score. Unless otherwise noted, we use the comparable score in our figures and analyses.

Figure IV shows the distribution of scores at the referral level. Note that the distribution is skewed to the right, which reflects that referrals are assigned the maximum score for all associated children. There is significant overlap in the distributions of CPS and GPS scores, although GPS scores tend to be higher, likely reflecting the more chronic system involvement associated with allegations of neglect as opposed to abuse.

[23]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.

Figure III shows the predictive power of the algorithm. In particular, the x axes show the seen score for each of the three algorithm versions, and the y axes show the share of GPS referrals with that score which are associated with a child removal within two years.[24]

Overall, there are more than 60,000 referrals in our sample.[25] Table I presents summary statistics separately for GPS (all and screened-in only) and CPS referrals. Note, while only 13% of the population in Allegheny County is Black, 50% of GPS and 44% of CPS 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. Table II shows the shares of CPS and GPS referrals which include each allegation category in our data, and Table III shows the shares of CPS and GPS referrals coming from each reporter category.

4 Empirical Framework

4.1 Difference-in-differences

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

Our primary outcomes of interest are: (1) whether a referral is screened in for investigation, (2) whether a screened-in referral results in a case opening, and (3) whether or not any child on a screened-in referral is removed within three months. In robustness tests we also study effects over different time frames.

We estimate effects separately for GPS (treated) and CPS (control) referrals. Recall, CPS referrals are legally required to be screened in for investigation, both before and after the AFST was implemented. GPS referrals, on the other hand, allow human discretion in determining which referrals to investigate.[26] As such, any changes in outcomes for CPS referrals are unlikely to be driven by the policy change, and may indicate other changes over time. Note, we cannot estimate such "placebo" difference-in-difference models for screening decisions, as all CPS referrals are investigated in both time periods.

We begin by estimating the following regression equation:

[24] The AFST has also been shown to be predictive of more "ground-truth" and universally measured outcomes such as maltreatment-related hospitalization - see Vaithianathan et al. (2020).

[25] Note, we do not include referrals for families which, at the time of referral, have an active case with CYF. We also exclude referrals made by truancy courts.

[26] See Section 2 for a more detailed discussion of the difference between these two categories, and Tables I and II for a breakdown of summary statistics by referral category.

$$y_{it} = \beta_0 Black_{it} + \beta_1 Post_{it} + \beta_2 BlackxPost_{it} + \beta_4 X_{it} + \gamma_m + \gamma_y + \epsilon \quad (1)$$

Where y_{it} is one of the following: (1) an indicator equal to one if referral i, made in month t, is screened in for an investigation; (2) an indicator equal to one if referral i, made in month t, is associated with a case opening; (3) an indicator equal to one if any child associated with referral i, made in month t, is removed from their home within three months. As for the independent variables, $Black_{it}$ 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; $Post_{it}$ 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.[27] We include these variables to control for any differential trends across race and time. In particular, substance exposure might differ across both race and time, as the opioid crisis has affected primarily White communities.

The identification assumption for this model requires that there are no differential time trends in screening, removals or case opening rates for referrals involving Black vs. White children. This assumption may be violated, for example, if there are other efforts to reduce disparities within CYF, or if Black and White families are differently affected by local conditions. As such, we also estimate a triple differences model, using CPS referrals as the control group, to estimate effects for the downstream outcomes of case opening and removal.

4.2 Triple differences

The triple differences identification assumption is weaker than that for difference-in-differences, and requires common trends in racial disparities for CPS and GPS referrals. We plot trends separately by race and referral type for several outcomes: referral volume and screen-in rate (Figure V), conditional case opening and removal rate (Figure VI), and unconditional case opening and removal rate (Figure VII). There are several points to note about these figures. Trends across time are noisy, due in part to small sample sizes in any given month. In Figure VI, it appears that disparities in downstream outcomes among screened-in GPS referrals experience a sharp reduction in the months prior to the introduction of the AFST. This is a violation of the assumption required for difference-in-differences. We address this in two ways. First, we introduce CPS referrals as a control group. To the best of our knowledge, there are no policy

[27] Allegation categories are listed in Table II. Reporter categories are listed in Table III. 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.

changes (other than the introduction of the AFST) which would affect only GPS referrals in this time period.[28] We also study effects on downstream outcomes unconditional on screen in. Figure VII shows pre-trends for this group.

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 VIII. Figures VIIIc and VIIId show the evolution of racial disparities in case openings and removals, before and after algorithm implementation. Note, in Figures VIIIa and VIIIc, 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 VIIIb and VIIId, the pattern is similar for removal rates.

The triple difference model is estimated with the following equation:

$$\begin{aligned} y_{it} = \beta_0 Black_{it} + \beta_1 GPS_{it} + \beta_2 Post_{it} + \beta_3 BlackXGPS_{it} + \beta_4 BlackXPost_{it} \\ + \beta_5 GPSXPost_{it} + \beta_6 BlackXPostXGPS_{it} + \beta_7 X_{it} + \gamma_m + \gamma_y + \epsilon \end{aligned} \tag{2}$$

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.

4.3 Heterogeneity by Risk Score

It is not *a priori* clear whether effects should be similar across the range of algorithm scores. In particular, 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 XI. Figure XIa shows the pre-AFST screen-in rates by comparable score; Figure XIb shows the post-AFST screen-in rates by comparable score; Figure XId shows the post-AFST screen-in rates by the seen score. Note that the change to the racial gap in screen-in rates seems particularly stark for high-risk referrals. In particular, 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 $\in \{1, 10\}$, estimating the equation:

[28] County officials did not know of any policy changes that might have an impact on GPS referrals in the pre-AFST period.

$$y_{it} = \sum_{j\neq 6}^{10} \beta_{j+2} Sj_{it} + \sum^{10} \beta_{j+11} Sj_{it} xBl_i + \sum^{10} \beta_{j+10} Sj_{it} xPost_t + \sum^{10} \beta_{j+9} Sj_{it} xBl_i xPost_t + \beta_{29} X_{it} + \gamma_m + \gamma_y + \epsilon \quad (3)$$

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. We also estimate Equation 3 where y_{it} is an indicator for case opening, conditional on screen in, and where y_{it} is an indicator for removal within three months, conditional on screen in. This allows us to assess heterogeneity in downstream effects for GPS referrals.

5 Results and Discussion

5.1 Screening Decision

In Table IV we report results from estimating Equation 1, or the effects of the algorithm on disparities in screen-in rates. Column (1) reports results excluding fixed effects and controls, Column (2) adds year and month-of-year fixed effects, column (3) adds referral-level controls, and column (4) adds an additional control for the underlying algorithm score. First, note that referrals involving Black children are more likely to be screened in than referrals involving White children, even when controlling for referral-level characteristics and underlying risk scores. There is no statistically significant effect of the algorithm on this disparity. That is, the coefficient on *BlackXPost* is statistically insignificant.

Motivated by the patterns in screening disparities observed in Figure XI, we next study how effects on screening decisions vary across algorithm scores. We report the coefficients on $BlackX PostxScore_i$ from estimating Equation 3 in Table V. Column (1) reports from a regression which excludes fixed effects and controls, Column (2) adds month-of-year and year fixed effects, and Column (3) reports results from our preferred specification, which includes fixed effects and referral-level controls. Note that the effect is largest in magnitude for the referrals with the highest risk scores. This is also the only bin for which the effect is statistically significant. This implies that the implementation of the algorithm reduced disparities in screen-in rates for the highest-risk referrals. The coefficients reported in Column (3), along with coefficients on $BlackxScore_i$, $PostxScore_i$ and $Score_i$ are shown graphically along with their 95% confidence intervals in Figure XII. Figure XIIa shows a positive relationship between algorithm score bin and screen-in rate. Figure XIIb shows that referrals involving Black children are more likely to be screened in at almost every risk score. This disparity is most notable at the lower and upper ends of the risk distribution. In particular, referrals

in the highest risk bin are 9.8 percentage points more likely to be screened in if they involve Black children. Figure XIIc suggests that the algorithm may have reduced screen-in rates for referrals with risk scores between 6 and 14 (bins 3-7), and increased screen-in rates for referrals with the highest risk scores. Finally, Figure XIId graphically presents the coefficients on $BlackXPostxScore_i$, reported in Table V. The effect on screening disparities is statistically indistinguishable from zero, except for the referrals within the highest risk bin. Specifically, this coefficient implies that the algorithm reduced the gap in screening rates across race by 9.6 percentage points, or 97% percent of the pre-existing disparity in this score bin.[29]

Note, 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).[30] Recall, although referrals in this category may still be screened out, this decision requires a supervisor's override. This result suggests that the protocol plays an important role in changing screening outcomes. [31]

5.2 Downstream outcomes

We next turn to the effects of the algorithm on the downstream outcomes of case opening and removal.

Results from estimating Equations 1 and 2 on the conditional likelihood of having a case opened are reported in Table VI. Before turning to the triple difference, Columns (1) and (2) report results from estimating Equation 1 separately on the GPS (treatment) and CPS (control) groups, respectively. In each of these regressions, we include year- and month-of-year fixed effects, as well as referral-level controls. In the treatment group (Column 1), screened-in referrals involving Black children are on average 6.25 percentage points more likely to have a case opened (18 percent of the sample mean). After the implementation of the algorithm, case opening rates increased overall by 8.60 percentage points, consistent with an increase in the average severity/accuracy of screened-in referrals. In the control group (Column 2), in which all referrals are screened in, referrals involving Black children are 2.67 percentage points more likely to have a case opened (31 percent of the sample mean). There is no statistically significant change in case opening rates in the Post period, consistent with the fact that the AFST only affected screening decisions for GPS referrals. Note that the coefficient on $BlackXPost$ is negative and significant for GPS (treated) referrals, and small and indistinguishable from zero for CPS (control) referrals. This suggests that any results in the triple difference are driven by changes in the treated, rather than control group. In particular, the difference-in-difference result suggests

[29] In unreported results from this regression, the coefficient on BlackxScore10 is .0983 (significant at the 1% level).

[30] Note that since we use the comparable score, the highest risk decile does not correspond exactly with the high-risk protocol. That is, there may be referrals with a comparable score of 19-20, but a seen score (from an earlier algorithm version) below 17. However, we cannot use the seen score in analyses comparing pre- to post- periods, as there is no seen score before the algorithm was implemented.

[31] One might wonder why we do not see similar 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).

that the implementation of the algorithm reduced disparities in case opening rates by 5.41 percentage points, or 87% of the pre-existing Black-White gap.

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, Column (5) reports results from a regression which adds referral-level controls (our preferred specification), and Column (6) replicates Column (5) for the entire sample of referrals (i.e., both screened-in and screened-out referrals). Our coefficient of interest, on the triple interaction $BlackXPostXGPS$, 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 5.48 percentage points. Note that this accounts for 87% of the pre-existing difference between GPS referrals involving Black and White children (6.27 percentage points).[32] In Column (6), the effect is somewhat attenuated, which can in part be explained by the lower average case opening rate for this sample.

Results from estimating Equation 2 on the conditional likelihood of removal within three months are reported in Table VII. Again, Columns (1) and (2) report results from difference-in-differences estimations for each of the treatment and control group. Reassuringly, the coefficient on $BlackXPost$ is again close to zero and insignificant in the control group. In the triple difference specification (Columns 3-5), our coefficient of interest (on the triple interaction $BlackXPostXGPS$), 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 2.87 percentage points, or 76% of the pre-existing difference (4.21 percentage points).[33] This result is statistically significant at the 1% level. In Column (6), effects are again attenuated, due to the lower average removal rates for this sample.

Finally, given the heterogeneity in screening decision effects by risk score, we estimate Equation 3, 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. The coefficients on $Score_i$, $BlackxScore_i$, $PostxScore_i$ and $BlackxPostxScore_i$ are shown in Figures XIII (case openings) and XIV (placements). In both Figures XIII (a) and XIV (a), note that conditional case openings and 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 the idea that both case openings and removals are indicators of severe or chronic problems (i.e., associated with higher risk). In Figure XIII (b), note that screened-in Black children are generally more likely to have a case opened across the range of risk scores. However, as shown in Figure XIV (b), the race differential in likelihood of removal is generally increasing in risk score. Both conditional

[32] 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 $BlackXGPS$. 0.0312 + 0.0315 = 0.0627.

[33] 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 $BlackXGPS$. 0.0222 + 0.0158 = 0.038.

case openings and removals are higher across the risk distribution in the Post period (consistent with an increase in accuracy of screening decisions). However, this result should be interpreted with caution, as CPS referrals also experienced an increase in removal rate in the Post period (see, e.g., Table VII). Finally, for case openings, the effect on disparities is relatively evenly distributed across the risk score distribution, as shown in Figure XIII (d). This result is somewhat surprising, given that effects on screening disparities were concentrated in the highest risk decile. One interpretation is that there may be more subtle, within-risk bin changes in the composition of screened-in referrals. For removals, the effect on disparities is increasing with risk, as shown in Figure XIV (d), mirroring the pattern of pre-existing racial disparities.

5.2.1 Dynamic Effects

We next test the effects of the algorithm implementation on case openings and removals over a longer time frame. In Table VIII we report results from estimating Equation 2 on case openings within 2 months (Column 1) through 24 months (Column 5) of a given referral. For comparability across columns, we use the same sample of calls, from January 2015 through December 2018, so that we can observe placements at least 24 months after each referral. Consider our triple difference coefficient of interest, on *BlackXPostxGPS*. Relative to our main results, effects are noisier but stable and directionally consistent.

In Table IX, we report the results from estimating equation 2 on removals within 2 months (Column 1) through 24 months (Column 5) of referrals. Effects are stable over time, although attenuated in the very long run (24 months). This result suggests that the algorithm may be affecting the timing of removals, and is consistent with two possible explanations: (1) The algorithm is pushing *forward* the removals of White children, (2) The algorithm is pushing *back* the removals of Black children. To test these hypotheses, we run a difference-in-differences regression separately for referrals involving White children, and referrals involving Black children. In particular, we estimate the equation:

$$y_{it} = \beta_0 GPS_i + \beta_1 Post_t + \beta_2 GPSxPost_{it} + \beta_3 X_{it} + \gamma_m + \gamma_y + \epsilon \tag{4}$$

Where y_{it} represents removals within a given time frame, conditional on screen in, and all other variables are as defined in Equation 2. Under the assumption of common trends for GPS and CPS referrals involving Black children, and common trends for GPS and CPS referrals involving White children, β_2 can be interpreted as the causal effect of the algorithm on removals. The results are shown in Tables X (referrals involving White children) and XI (referrals involving Black children). Note that the effect of the algorithm is consistently positive for referrals involving White children, and negative for referrals involving Black children. These results are consistent with both explanations above.

That is, removals for White children increase, while removals for Black children decrease.

5.2.2 Re-referrals and Score Progression

A common concern with predictive risk models is that bias in the data will create a biased algorithm. In our setting, one might be concerned that differential rates of referrals will differentially affect observed algorithm scores. In particular, since past involvement with the child protection system (including past referrals, investigations and removals) are included as risk factors in the algorithm, community bias in the decision to refer a child could in theory artificially increase scores for Black children at a faster rate than for White children.

To address this concern, we first directly ask whether re-referrals differ by race, conditional on risk score. Figure IX plots average number of times a child on a referral is re-referred within 12 months, separately by referral risk score and by race. Note that there is no consistent pattern across algorithm score in re-referral disparities by race.

Next, we ask whether average scores are changing over time at different rates for referrals involving Black children vs. referrals involving White children. To answer this question, we plot algorithm scores over time for the cohort of children who are first seen on referrals in 2015. Figure X shows the average scores for Black children vs. White children first seen in 2015, over time. Visually it appears that, if anything, scores for White children first referred in 2015 are increasing at a faster rate than scores for Black children.[34]

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 disparities, relative to human decision makers.

Note, we cannot and do not speak to either optimal investigation and removal rates, or the welfare consequences of reducing disparities in these outcomes. Policy-makers and communities are concerned about unwarranted disparities in screening decisions, case openings and removals in and of itself, and are actively working to reduce those disparities. Our work shows that predictive risk models may be a useful tool for reaching this particular policy goal.

[34] A chi-squared test rejects the null hypothesis that the two slopes are equal.

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.

Administration on Children, Youth, and Children's Bureau Families. 2021. *Child Maltreatment 2019.* Technical report. U.S. Department of Health Human Services, Administration for Children and Families.

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.

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.*

Bartholet, Elizabeth. 2009. "The racial disproportionality movement in child welfare: False facts and dangerous directions." *Ariz. L. Rev.* 51:871.

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.

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

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.

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.

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.

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, Alexandra Chouldechova, Diana Benavides-Prado, and Rachel Berger. 2020. "Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals: evidence from the Allegheny Family Screening Tool." *JAMA pediatrics* 174 (11): e202770–e202770.

Vaithianathan, 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.

Tables

Table I: Comparative summary statistics - GPS vs CPS

	All GPS	***Screened-in GPS***	***All CPS***
	Mean	Mean	Mean
Black	0.498	0.543	0.444
Screened In	0.428	1.000	0.797
Case Opened	0.154	0.358	0.072
Case Opened 3m	0.176	0.356	0.086
Case Opened 12m	0.265	0.445	0.157
Placed 3m	0.045	0.080	0.030
Placed 12m	0.094	0.150	0.065
Any Child (infant)	0.147	0.233	0.158
Any Child (1-5)	0.435	0.480	0.413
Any Child (6-12)	0.589	0.598	0.572
Any Child (13-17)	0.380	0.356	0.414
Number of Children	2.174	2.344	2.188
N	44343	18983	14099

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. 2015 and Dec. 2019 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Table II: Allegations - GPS vs. CPS

	Mean (GPS)	Mean (CPS)	Diff
Caregiver Behavioral Health	0.061	0.017	0.044***
Caregiver Substance Abuse	0.222	0.041	0.181***
Causing Death of Child	0.003	0.005	-0.002***
Child Behaviors	0.068	0.023	0.045***
Domestic Violence	0.072	0.029	0.043***
Exposure to Risk	0.160	0.031	0.129***
Failure to Protect	0.037	0.014	0.023***
Imminent Risk	0.019	0.013	0.006***
Inadequate Physical Care	0.182	0.036	0.145***
Medical Neglect	0.042	0.017	0.025***
Mental Health	0.079	0.030	0.049***
Mental Injuries	0.015	0.024	-0.009***
Neglect	0.098	0.015	0.083***
No/Inadequate Home	0.102	0.014	0.087***
None	0.061	0.001	0.060***
Other	0.029	0.004	0.025***
Other Referral Source	0.005	0.004	0.002*
Parent/Child Conflict	0.040	0.016	0.024***
Physical Altercaction	0.006	0.018	-0.012***
Physical Maltreatment	0.159	0.735	-0.576***
Sexual Abuse or Exploitation	0.031	0.175	-0.145***
Sexual Contact Between Children	0.045	0.006	0.039***
Truancy	0.042	0.006	0.036***
Unknown	0.007	0.004	0.003***
Unwilling or Unable To Provide Care	0.093	0.015	0.078***
Youth Substance Abuse	0.014	0.004	0.010***
N	61678		

This table reports the share of referrals involving certain allegations for GPS vs. CPS referrals, as well as t-test results. The sample is all referrals made between Jan. 2015 and Dec. 2019 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Table III: Reporters - GPS vs. CPS

	Mean (GPS)	Mean (CPS)	Diff
Agency	0.309	0.282	0.026***
Anonymous	0.100	0.029	0.071***
Community	0.053	0.025	0.028***
Family	0.147	0.068	0.079***
Law	0.093	0.093	-0.000
Medical	0.063	0.130	-0.067***
School	0.098	0.149	-0.051***
Self	0.005	0.005	0.000
Therapist	0.132	0.220	-0.088***
N	61678		

This table reports the share of referrals from different reporter categories for CPS vs. GPS referrals, as well as t-test results. The sample is all referrals made between Jan. 2015 and Dec. 2019 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Table IV: Results: Screen In

	(1) DD	(2) +FE	(3) +Controls	(4) +Risk Score
Post	-0.0118*	0.00422	-0.000263	0.00322
	(0.00699)	(0.0129)	(0.0116)	(0.0116)
Black	0.0779***	0.0767***	0.0610***	0.0355***
	(0.00890)	(0.00890)	(0.00817)	(0.00822)
BlackXPost	0.00754	0.00871	-0.00538	-0.0106
	(0.0102)	(0.0102)	(0.00925)	(0.00926)
Score				0.0198***
				(0.000539)
Month of Intake	No	Yes	Yes	Yes
Yr of Intake	No	Yes	Yes	Yes
Controls	No	No	Yes	Yes
Mean	0.425	0.425	0.425	0.430
Obs.	52533	52533	52533	51378

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 **??**. The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which are screened in, 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 V: Results: Screen In by Score

	(1) Base	(2) + FE	(3) + Controls
BlackxPostxScore2	-0.0280	-0.0297	-0.0473
	(0.0625)	(0.0625)	(0.0576)
BlackxPostxScore3	-0.0363	-0.0362	-0.0406
	(0.0378)	(0.0379)	(0.0356)
BlackxPostxScore4	-0.00420	-0.00438	-0.0184
	(0.0332)	(0.0332)	(0.0311)
BlackxPostxScore5	0.0213	0.0203	0.0265
	(0.0299)	(0.0299)	(0.0274)
BlackxPostxScore6	0.0294	0.0316	0.00361
	(0.0279)	(0.0280)	(0.0252)
BlackxPostxScore7	-0.0135	-0.0111	-0.0253
	(0.0266)	(0.0266)	(0.0238)
BlackxPostxScore8	-0.00737	-0.00650	-0.0267
	(0.0258)	(0.0258)	(0.0234)
BlackxPostxScore9	0.0351	0.0369	0.0136
	(0.0261)	(0.0261)	(0.0240)
BlackxPostxScore10	-0.108***	-0.107***	-0.0958***
	(0.0334)	(0.0334)	(0.0314)
Month of Intake	No	Yes	Yes
Yr of Intake	No	Yes	Yes
Controls	No	No	Yes
Mean	0.431	0.431	0.431
Obs.	51177	51177	51177

Standard errors in parentheses

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

This table reports coefficients and standard errors from estimating three specifications of equation 3. The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which are screened in, 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. Unreported coefficients are available upon request.

Table VI: Results: Open Case | Screen In

	(1) DD GPS	(2) DD Control	(3) DDD Base	(4) +FE	(5) +Controls	(6) Unconditional
Post	0.0860***	0.0138	-0.0113	0.0414***	0.0329**	0.0146*
	(0.0194)	(0.0140)	(0.00732)	(0.0132)	(0.0131)	(0.00827)
Black	0.0625***	0.0267***	0.0401***	0.0413***	0.0312***	0.0271***
	(0.0130)	(0.0100)	(0.0103)	(0.0103)	(0.0101)	(0.00872)
BlackXPost	-0.0541***	0.00180	0.00217	0.00194	-0.000980	-0.00198
	(0.0147)	(0.0114)	(0.0117)	(0.0118)	(0.0115)	(0.00976)
GPS			0.235***	0.235***	0.131***	-0.0303***
			(0.0112)	(0.0112)	(0.0135)	(0.00675)
BlackXGPS			0.0303*	0.0286*	0.0315*	0.0216**
			(0.0166)	(0.0166)	(0.0163)	(0.0107)
GPSXPost			0.0313**	0.0331**	0.0398***	0.0221***
			(0.0130)	(0.0130)	(0.0129)	(0.00708)
BlackXPostXGPS			-0.0591***	-0.0557***	-0.0548***	-0.0256**
			(0.0190)	(0.0190)	(0.0187)	(0.0121)
Month of Intake	Yes	Yes	No	Yes	Yes	Yes
Yr of Intake	Yes	Yes	No	Yes	Yes	Yes
Controls	Yes	Yes	No	No	Yes	Yes
Mean	0.341	0.0853	0.249	0.249	0.248	0.127
Obs.	22329	12877	35258	35258	35206	68789

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)) different specifications of equation 2 (Columns (3)-(5)). The sample for Columns (1) - (5) is all screened-in referrals made between Jan. 2015 and Sept. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. In Columns (1) and (2) the sample is restricted further to, respectively, only GPS and only CPS referrals. The sample for Column (6) is expanded to include screened-out referrals. 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 VII: Results: Placed within 3 months | Screen In

	(1) DD GPS	(2) DD Control	(3) DDD Base	(4) +FE	(5) +Controls	(6) Unconditional
Post	0.0338***	0.0241***	0.00666	0.0307***	0.0267***	0.0168***
	(0.0112)	(0.00891)	(0.00410)	(0.00776)	(0.00771)	(0.00495)
Black	0.0381***	0.0179***	0.0210***	0.0211***	0.0222***	0.0195***
	(0.00747)	(0.00594)	(0.00597)	(0.00597)	(0.00597)	(0.00532)
BlackXPost	-0.0296***	-0.0000442	-0.000617	-0.000646	-0.00130	0.00158
	(0.00850)	(0.00707)	(0.00713)	(0.00714)	(0.00710)	(0.00619)
GPS			0.0417***	0.0417***	0.0191***	-0.00570
			(0.00583)	(0.00583)	(0.00709)	(0.00379)
BlackXGPS			0.0221**	0.0215**	0.0158*	0.0126*
			(0.00960)	(0.00959)	(0.00949)	(0.00653)
GPSXPost			0.00159	0.00155	0.00478	0.000404
			(0.00698)	(0.00699)	(0.00700)	(0.00399)
BlackXPostXGPS			-0.0318***	-0.0308***	-0.0287***	-0.0177**
			(0.0112)	(0.0112)	(0.0111)	(0.00755)
Month of Intake	Yes	Yes	No	Yes	Yes	Yes
Yr of Intake	Yes	Yes	No	Yes	Yes	Yes
Controls	Yes	Yes	No	No	Yes	Yes
Mean	0.0776	0.0342	0.0618	0.0618	0.0616	0.0393
Obs.	21331	12436	33819	33819	33767	66074

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)) different specifications of equation 2 (Columns (3)-(5)). The sample for Columns (1) - (5) is all screened-in referrals made between Jan. 2015 and Sept. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. In Columns (1) and (2) the sample is restricted further to, respectively, only GPS and only CPS referrals. The sample for Column (6) is expanded to include screened-out referrals. The outcome variable in each regression is an indicator equal to one for referrals which include any child who is placed outside their home within three months of the screening decision, 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 VIII: Results: Open Case within x months | Screen In

	(1) 2 months	(2) 3 months	(3) 6 months	(4) 12 months	(5) 24 months
GPS	0.118***	0.122***	0.137***	0.142***	0.163***
	(0.0148)	(0.0150)	(0.0157)	(0.0166)	(0.0174)
Post	0.00343	0.0107	0.0216	0.0212	0.0219
	(0.0139)	(0.0142)	(0.0150)	(0.0163)	(0.0175)
Black	0.0417***	0.0536***	0.0751***	0.104***	0.150***
	(0.0106)	(0.0111)	(0.0122)	(0.0137)	(0.0152)
BlackXGPS	0.0128	0.000770	-0.00724	-0.0204	-0.0633***
	(0.0166)	(0.0170)	(0.0180)	(0.0192)	(0.0203)
BlackXPost	-0.00716	-0.0124	-0.0148	-0.0202	-0.0252
	(0.0130)	(0.0137)	(0.0151)	(0.0171)	(0.0189)
GPSXPost	0.0790***	0.0749***	0.0686***	0.0678***	0.0505***
	(0.0142)	(0.0145)	(0.0152)	(0.0162)	(0.0171)
BlackXPostXGPS	-0.0418**	-0.0347	-0.0345	-0.0339	-0.0205
	(0.0207)	(0.0213)	(0.0225)	(0.0240)	(0.0253)
Month of Intake	Yes	Yes	Yes	Yes	Yes
Yr of Intake	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Mean	0.258	0.273	0.308	0.361	0.432
Obs.	23669	23669	23669	23669	23669

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 equation 2. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2018 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which include any child for whom a case is opened within 2 months (Column (1)) to 24 months (Column (5)) of the screening decision, 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 IX: Results: Placed within x months | Screen In

	(1) 2 months	(2) 3 months	(3) 6 months	(4) 12 months	(5) 24 months
GPS	0.0166**	0.0147*	0.0244***	0.0249**	0.0435***
	(0.00687)	(0.00758)	(0.00923)	(0.0110)	(0.0127)
Post	0.0184**	0.0233***	0.0219**	-0.00361	0.00683
	(0.00731)	(0.00809)	(0.00963)	(0.0115)	(0.0131)
Black	0.0153***	0.0218***	0.0401***	0.0618***	0.105***
	(0.00529)	(0.00600)	(0.00756)	(0.00959)	(0.0116)
BlackXGPS	0.0146*	0.0161*	0.0140	0.00511	-0.0284*
	(0.00864)	(0.00951)	(0.0114)	(0.0137)	(0.0160)
BlackXPost	0.00340	0.00463	0.00131	-0.00558	-0.0271*
	(0.00696)	(0.00785)	(0.00963)	(0.0119)	(0.0141)
GPSXPost	0.00528	0.0117	0.0149	0.0226**	0.0107
	(0.00708)	(0.00775)	(0.00913)	(0.0108)	(0.0125)
BlackXPostXGPS	-0.0305***	-0.0344***	-0.0400***	-0.0358**	-0.0132
	(0.0110)	(0.0122)	(0.0144)	(0.0170)	(0.0196)
Month of Intake	Yes	Yes	Yes	Yes	Yes
Yr of Intake	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Mean	0.0526	0.0656	0.0920	0.128	0.169
Obs.	23669	23669	23669	23669	23669

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 equation 2. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2018 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which include any child who is placed outside their home within 2 months (Column (1)) to 24 months (Column (5)) of the screening decision, 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 X: Results: Placed within x months | SI (White)

	(1) 2 months	(2) 3 months	(3) 6 months	(4) 12 months	(5) 24 months
GPS	0.0149*	0.0126	0.0212**	0.0181	0.0403***
	(0.00761)	(0.00839)	(0.0102)	(0.0120)	(0.0142)
Post	0.00857	0.0141	0.00712	-0.0185	-0.0162
	(0.00886)	(0.00973)	(0.0114)	(0.0137)	(0.0155)
GPSXPost	0.00386	0.00928	0.0126	0.0191*	0.00618
	(0.00710)	(0.00777)	(0.00912)	(0.0108)	(0.0124)
Month of Intake	Yes	Yes	Yes	Yes	Yes
Yr of Intake	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Mean	0.0430	0.0523	0.0721	0.0993	0.131
Obs.	11338	11338	11338	11338	11338

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 equation 4. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2018 to CYF which include at least one White child and no Black children, excluding active-family referrals, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which include any child who is placed outside their home within 2 months (Column (1)) to 24 months (Column (5)) of the screening decision, 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 XI: Results: Placed within x months | SI (Black)

	(1) 2 months	(2) 3 months	(3) 6 months	(4) 12 months	(5) 24 months
GPS	0.0292***	0.0280***	0.0345***	0.0279*	0.0120
	(0.00901)	(0.0100)	(0.0125)	(0.0148)	(0.0172)
Post	0.0294***	0.0352***	0.0347**	0.00201	-0.00178
	(0.0107)	(0.0121)	(0.0145)	(0.0169)	(0.0191)
GPSXPost	-0.0219**	-0.0182*	-0.0198*	-0.00610	0.00493
	(0.00854)	(0.00955)	(0.0113)	(0.0133)	(0.0152)
Month of Intake	Yes	Yes	Yes	Yes	Yes
Yr of Intake	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Mean	0.0615	0.0778	0.110	0.154	0.204
Obs.	12331	12331	12331	12331	12331

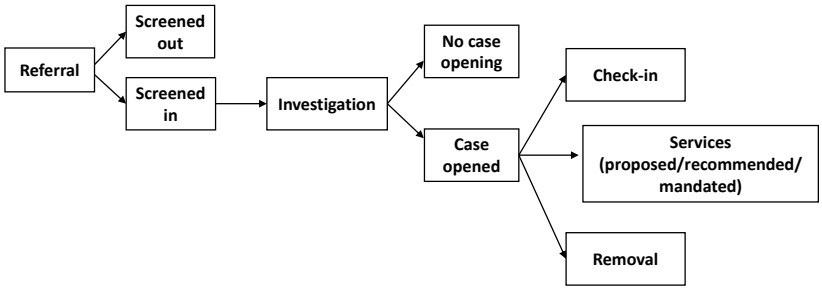
Standard errors in parentheses

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

This table reports coefficients and standard errors from five separate regressions estimating equation 4. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2018 to CYF which include at least one Black child, excluding active-family referrals, and referrals stemming from truancy courts. The outcome variable in each regression is an indicator equal to one for referrals which include any child who is placed outside their home within 2 months (Column (1)) to 24 months (Column (5)) of the screening decision, 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.

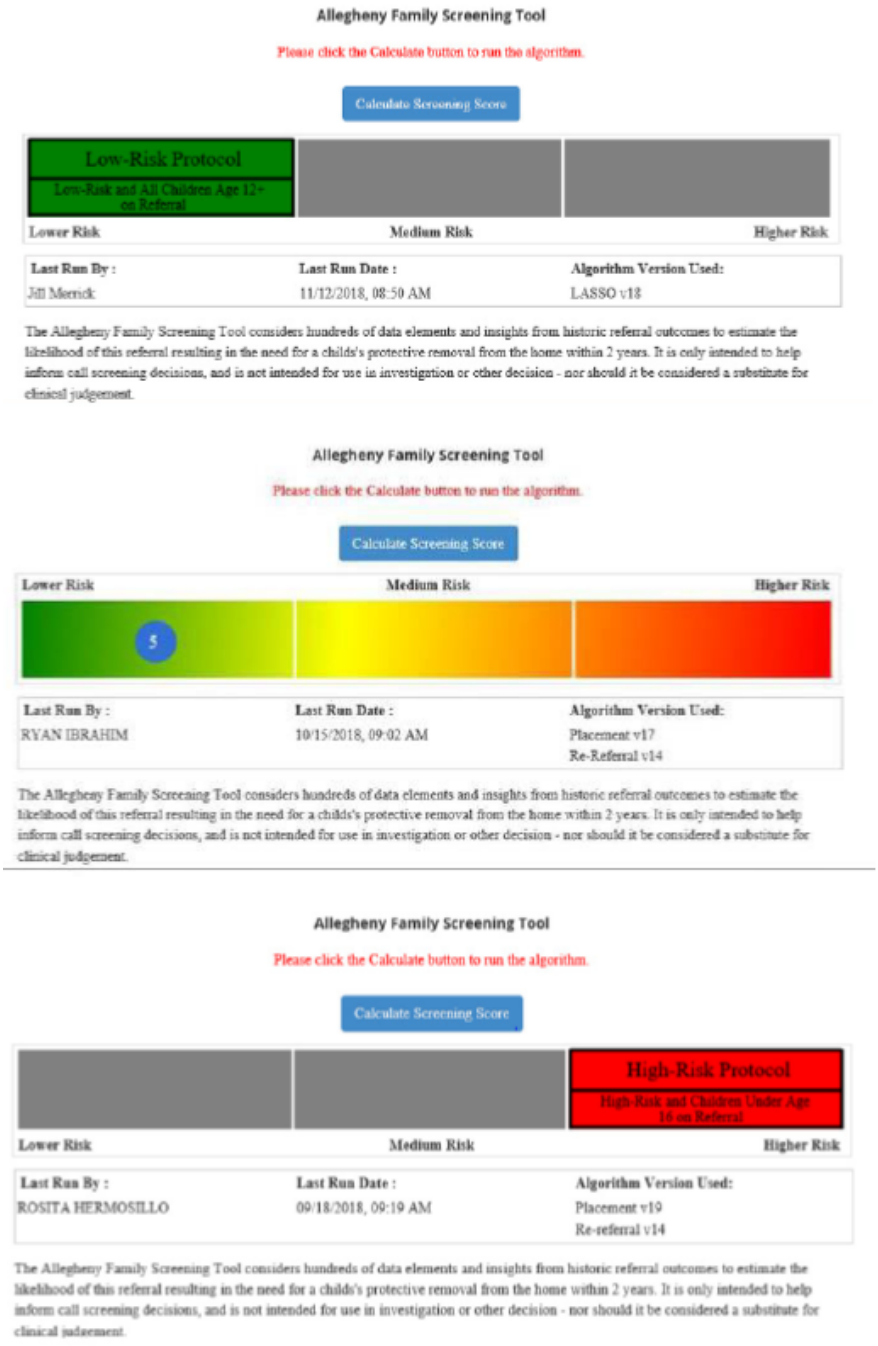
Figures

Figure I: Referral process



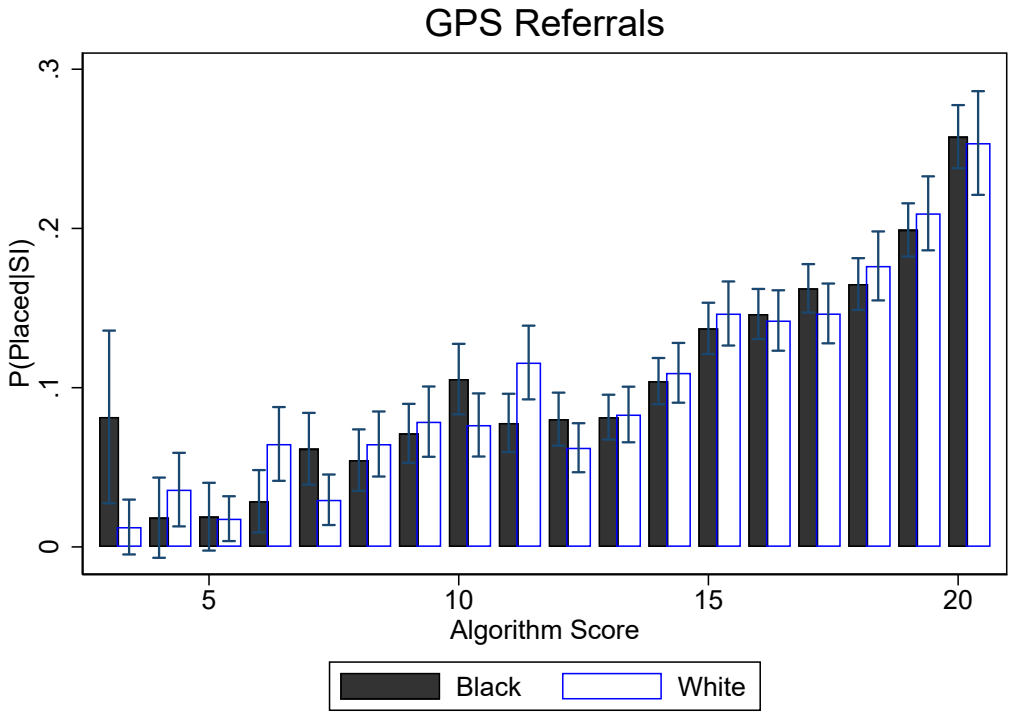
This figure presents the steps by which a referral to CYF moves through the system in Allegheny County, PA.

Figure II: AFST Screener View



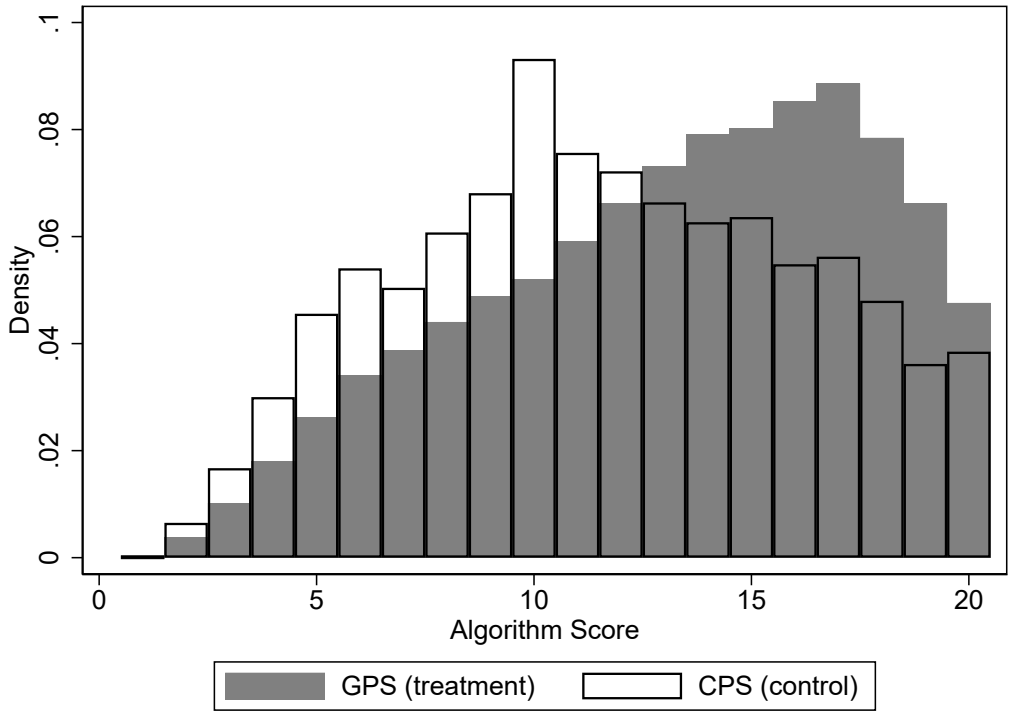
This figure presents the view that a screener has after running the algorithm, in three different scenarios. The top panel shows the screener's view if a low-risk protocol is implemented (i.e. all children are above a given age and all scores are below a given cutoff – the exact age and score cutoffs have changed over time). The middle panel shows the screener's view if neither the low-risk nor the high-risk protocol is implemented. The bottom panel 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).

Figure III: Removal within 24 months



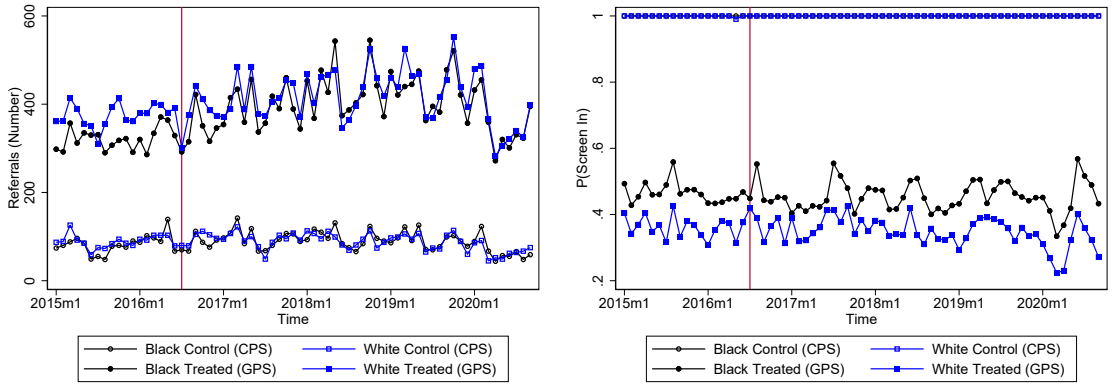
For each algorithm score from 3 to 20, and for Black vs. White children, this figure presents the likelihood that a referral with that score involves a child who will be removed and placed in foster care within 24 months. We use the score from the latest version of the algorithm, which was retroactively calculated for referrals prior to July 2019. Note also that we use the referral-level score, which is the maximum of all scores calculated for each child on the referral. The sample is all screened-in and screened-out GPS referrals made between Jan. 2015 and Dec. 2018 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. For each score/race, 95% confidence intervals are shown. Note that the observed correlation between scores and removals illustrated here is at the referral level for GPS, whilst the AFST is trained at the child level for all referrals. The accuracy is lower as a result (such as in Figure 4 , Vaithianathan et al (2019)) .

Figure IV: Score distribution (CPS v GPS)



This figure presents separately the share of GPS and CPS referrals which are scored at each risk score, 1-20. We use the score from the latest version of the algorithm, which was retroactively calculated for referrals prior to July 2019. Note also that we use the referral-level score, which is the maximum of all scores calculated for each child on the referral. The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Figure V: Time trends: Referrals and Screening decisions

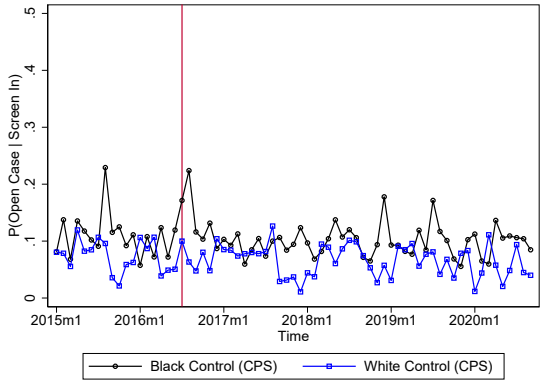


(a) Number of referrals

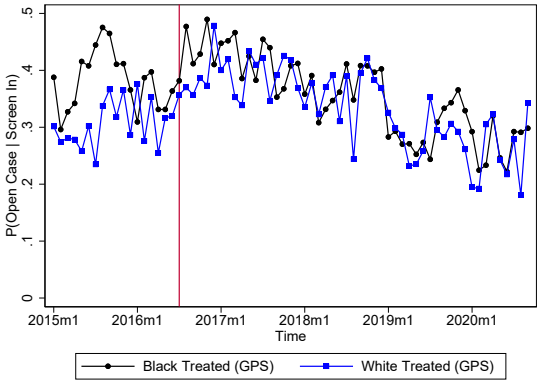
(b) Screening

This figure presents monthly numbers of referrals (in panel a) and monthly average screen-in rates (in panel b) separately by race of referral (as defined in the text) and CPS/GPS status. The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

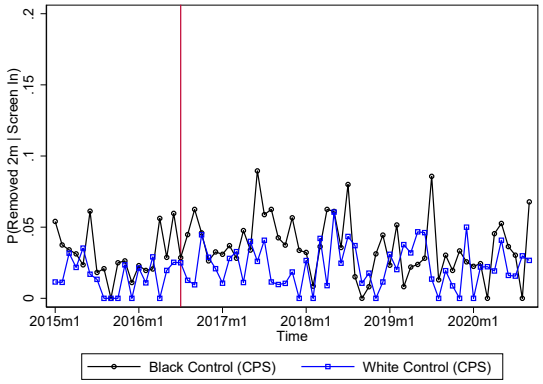
Figure VI: Time trends: Downstream outcomes (conditional on screen in)



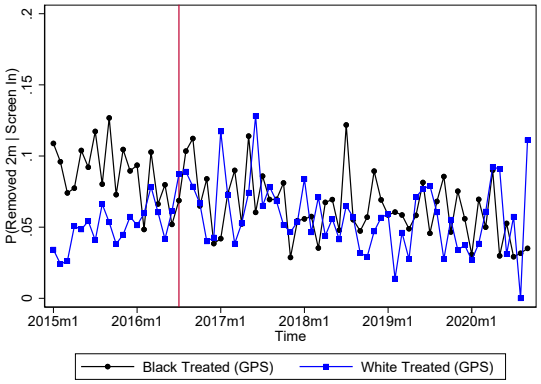
(a) Open Case | Screen In (CPS)



(b) Open Case | Screen In (GPS)



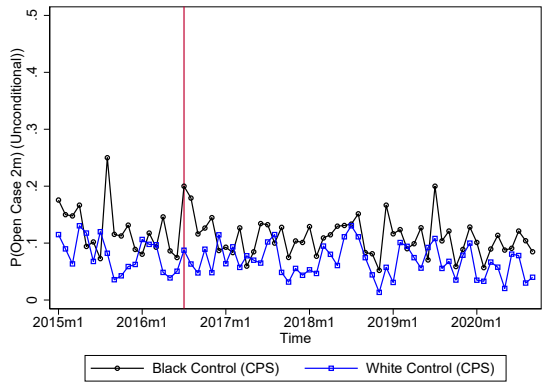
(c) Removal | Screen In (CPS)



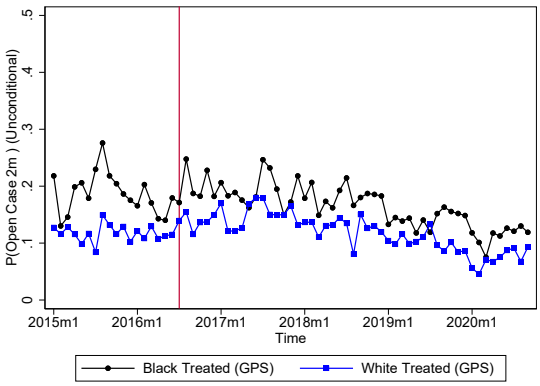
(d) Removal | Screen In (GPS)

This figure presents monthly case opening and removal rates by race for both CPS referrals (panels a and c) and GPS referrals (panels b and d). The sample is all screened-in referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

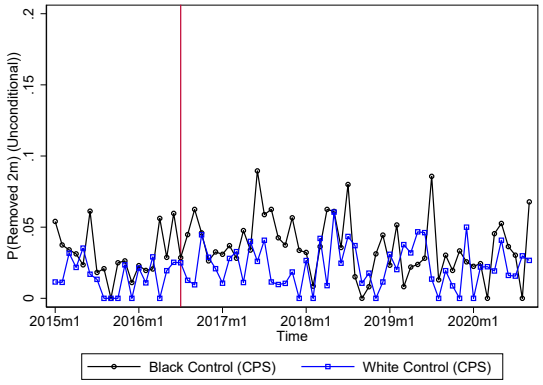
Figure VII: Time trends: Downstream outcomes (unconditional on screen in)



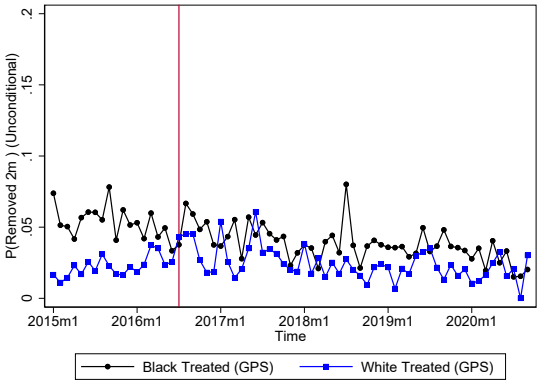
(a) Open Case (CPS)



(b) Open Case (GPS)



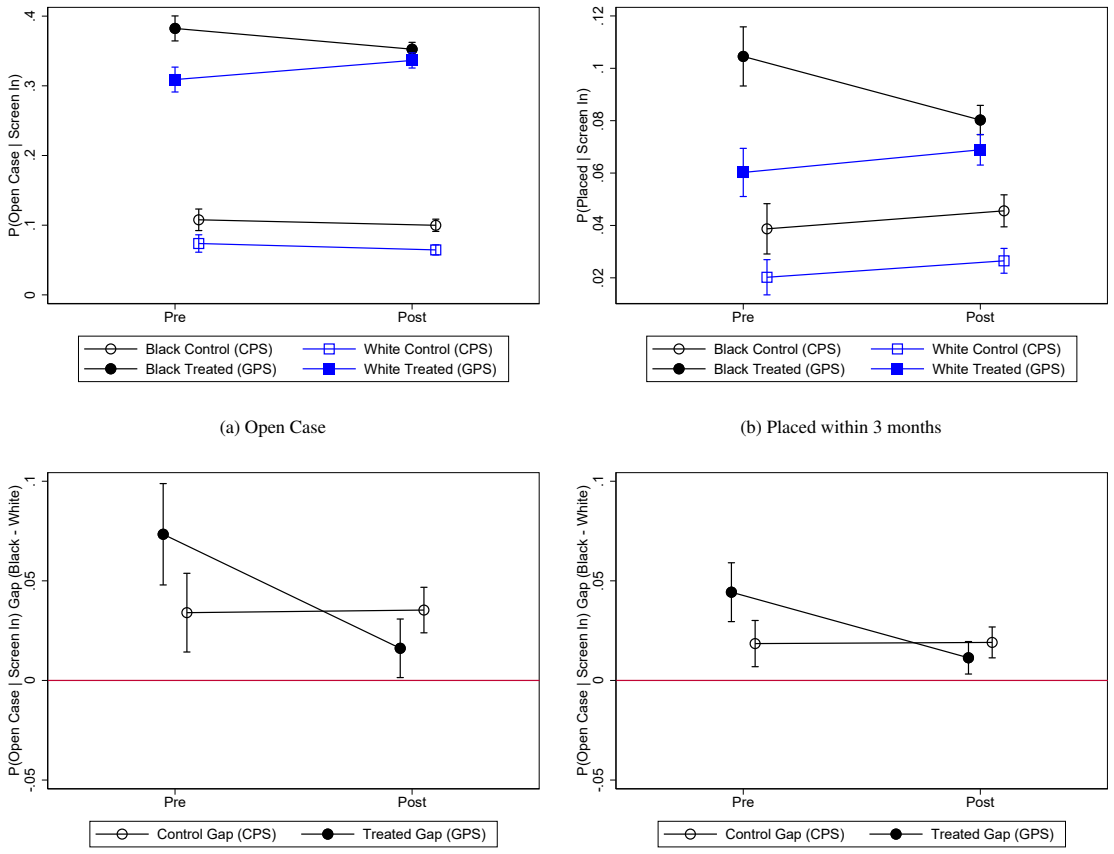
(c) Removal (CPS)



(d) Removal (GPS)

This figure presents monthly case opening and removal rates by race for both CPS referrals (panels a and c) and GPS referrals (panels b and d). The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Figure VIII: Downstream Disparities



(a) Open Case

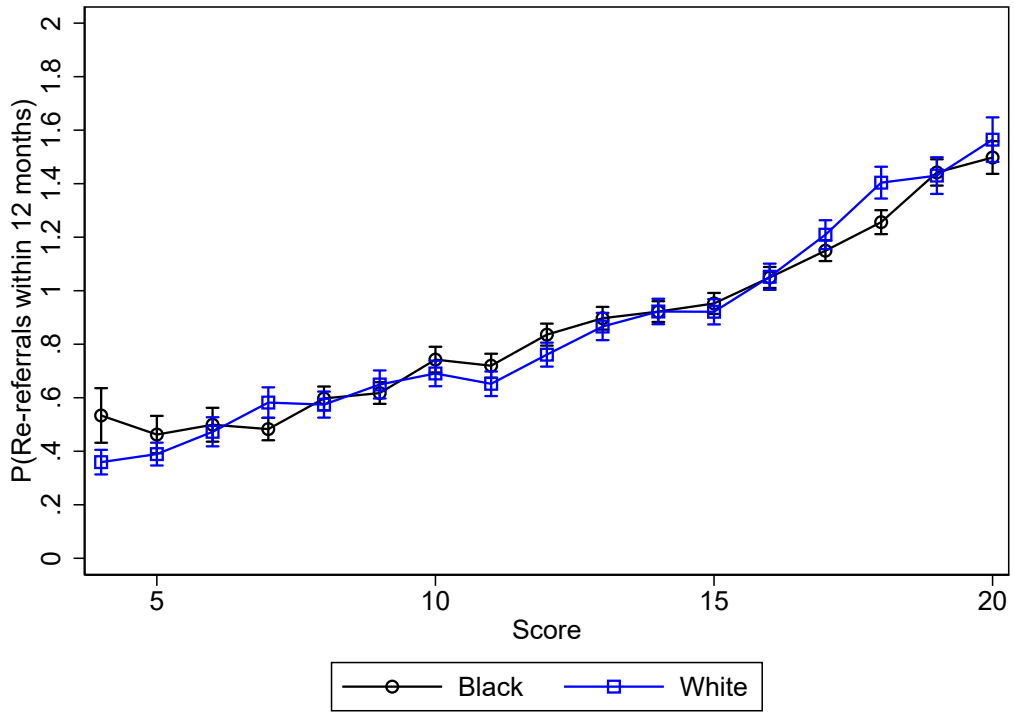
(b) Placed within 3 months

(c) Disparity - Open Case

(d) Disparity - Placed within 3 months

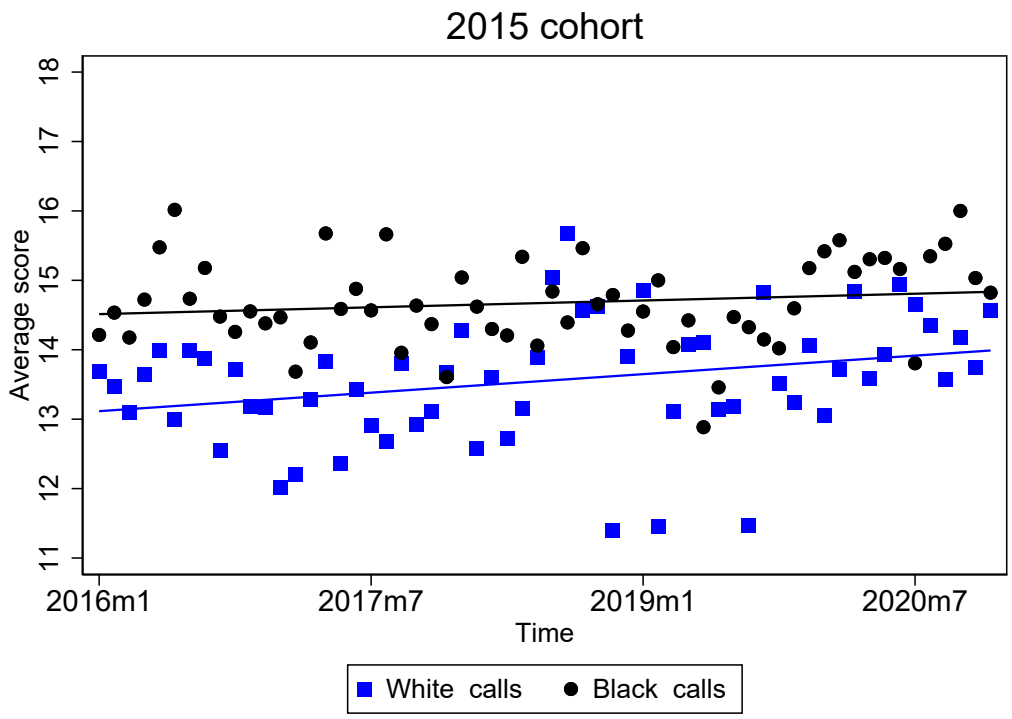
This figure presents our main outcome variables of interest, across time period, treatment/control group, and race of children on the referral. The sample is all screened-in referrals made between Jan. 2015 and Sept. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. 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 IX: Re-referrals



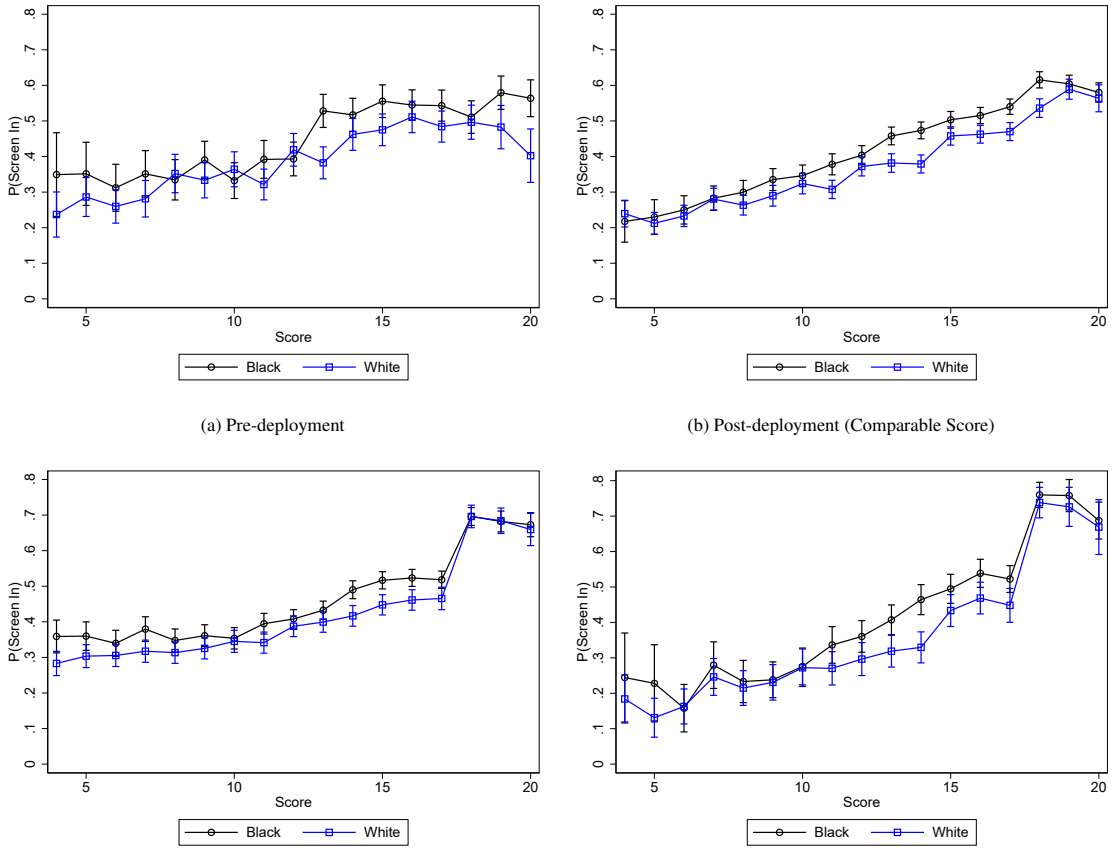
This figure reports average number of re-referrals within 12 months by algorithm score and child race. The sample is all children associated with screened-in and screened-out referrals made between Jan. 2015 and Dec. 2019 to CYF, excluding children involved in active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. For each score and race, 95% confidence intervals are shown.

Figure X: Score Progression



This figure presents the score progression for children who are first referred in 2015, and are re-referred at any point before Jan. 2021. The sample is all children associated with any referral in 2015, who are also involved in at least one subsequent referral. Active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts are excluded from the sample. Each dot represents the average score for children of a given race, who are referred in a given month. Note, we report referral-level scores, as calculated by the most recent version of the algorithm. The lines are best-fit lines through the data.

Figure XI: Screen In Disparities



(a) Pre-deployment

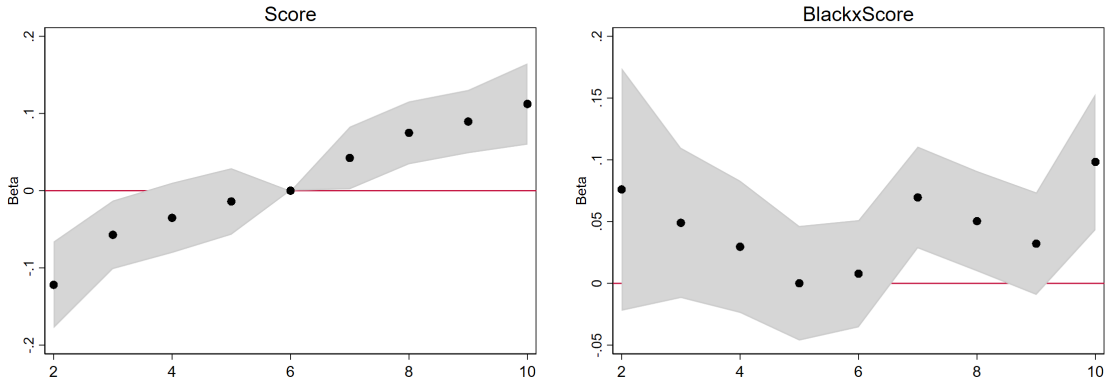
(b) Post-deployment (Comparable Score)

(c) Post-deployment (Observed Score)

(d) Post-deployment AFST3 (Observed Score = Comparable Score)

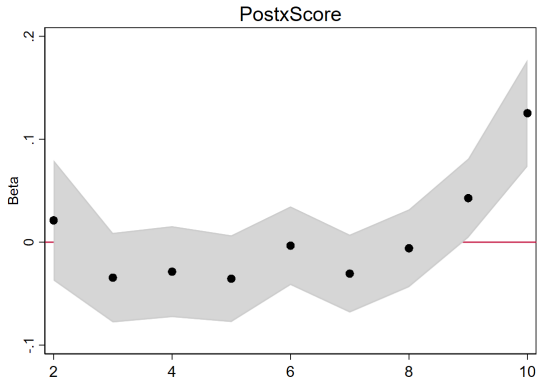
This figure reports average screen-in rates by algorithm score and race. Screen-in rate is defined as the share of referrals which are screened in for an investigation. The sample is all screened-in and screened-out referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts. For each score and race, 95% confidence intervals are shown. Each sub-figure presents results from a different time period, or definition of algorithm score. Panel (a) presents screen-in rates from the period before the algorithm was implemented, from Jan. 2015 through Jun. 2016. The algorithm score in this panel was retroactively calculated, using the latest version of the AFST (in place since July 2019). 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 reflects the "comparable score," i.e. the score which was calculated using the latest version of the AFST (in place since July 2019). For those referrals made after July 2019, the score as seen by the screeners is used. Panel (c) presents screen-in rates from the period after the algorithm was implemented, from July 2016 through Dec. 2020. The algorithm score in this panel reflects the "seen score," i.e. the score from the algorithm which was in place at the time of the referral. Panel (d) presents screen-in rates from the period after the last version of the algorithm was implemented, i.e. for July 2019 through Dec. 2020. The algorithm score in this panel reflects the "seen score" during this time period, which is exactly equal to the "comparable score," for this time period only.

Figure XII: Screening by Score

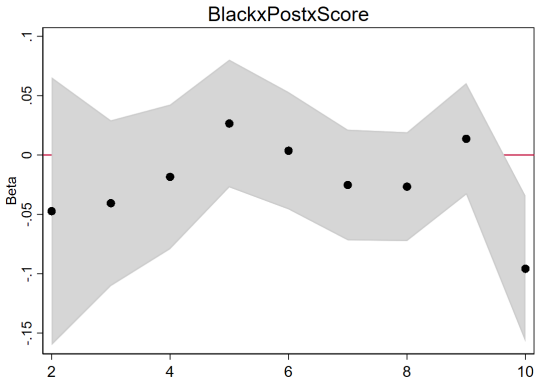


(a) Coefficients on Score

(b) Coefficients on BlackxScore



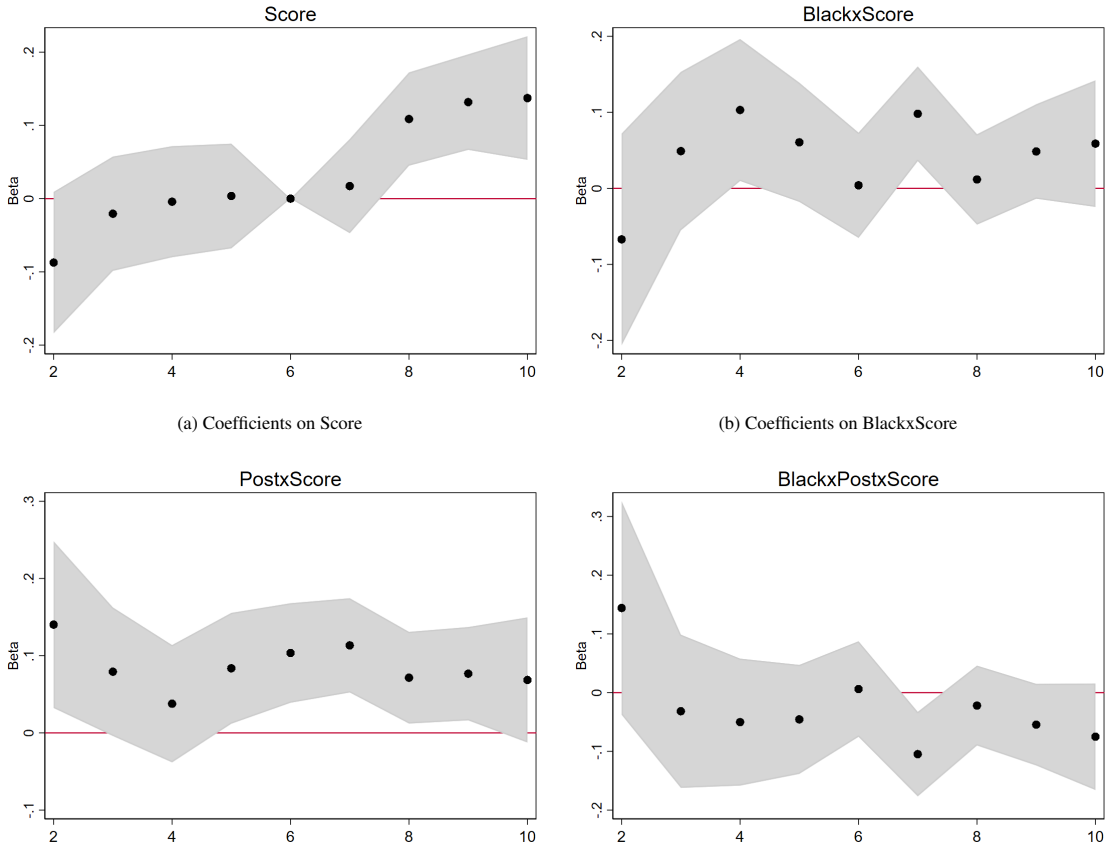
(c) Coefficients on PostxScore



(d) Coefficients on BlackxPostxScore

This figure graphically presents coefficients and 95% confidence intervals from estimating equation 3, where the outcome variable is equal to one if the referral is screened-in, and zero otherwise. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Figure XIII: Open Case by Score



(a) Coefficients on Score

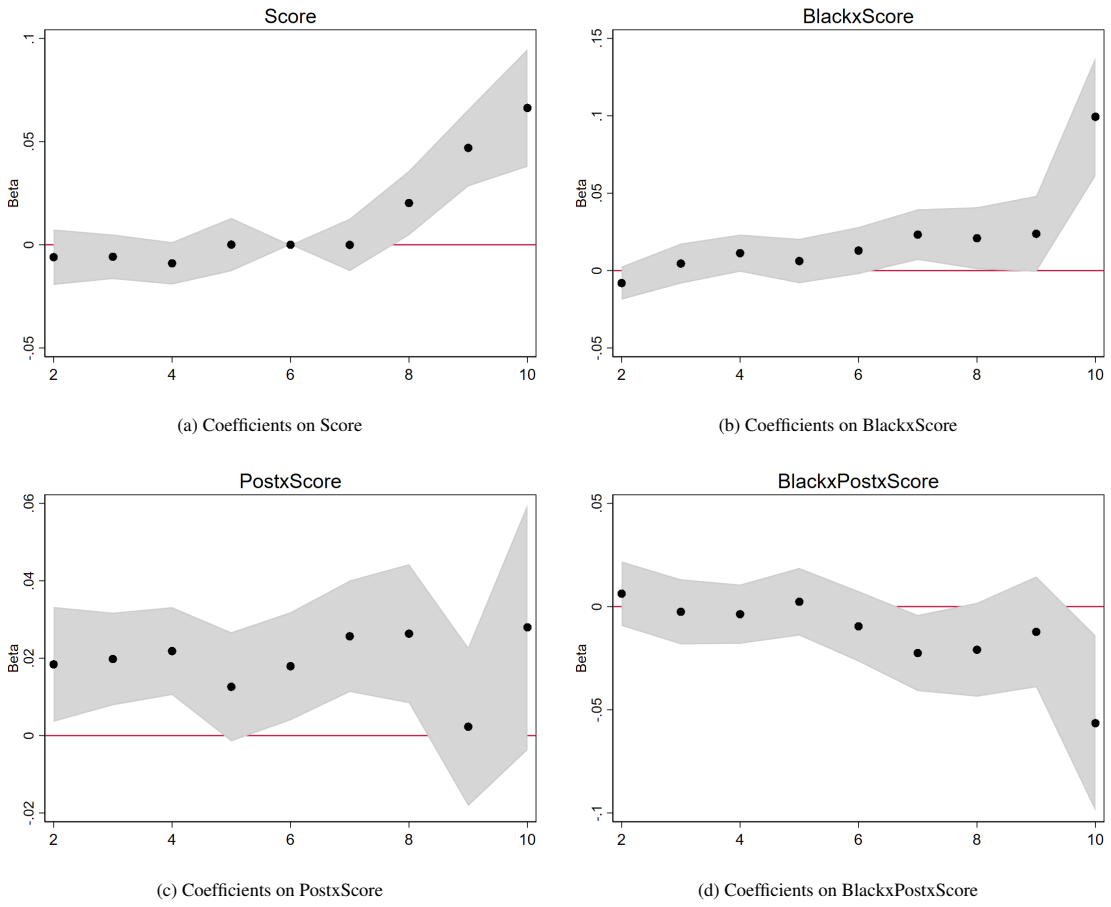
(b) Coefficients on BlackxScore

(c) Coefficients on PostxScore

(d) Coefficients on BlackxPostxScore

This figure graphically presents coefficients and 95% confidence intervals from estimating equation 3, where the outcome variable is equal to one if the referral results in a case opening, and zero otherwise. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.

Figure XIV: 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 3, where the outcome variable is equal to one if any child associated with the referral is placed within 3 months, and zero otherwise. The sample is all screened-in referrals made between Jan. 2015 and Dec. 2020 to CYF, excluding active-family referrals, referrals involving neither Black children nor White children, and referrals stemming from truancy courts.