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

Katherine Rittenhouse†
University of Texas at Austin

Emily Putnam-Hornstein
University of North Carolina at Chapel Hill

Rhema Vaithianathan
Auckland University of Technology

April 1, 2024

Abstract

Equity concerns are a primary constraint in the adoption of algorithms. We ask how providing decision-makers with a predictive risk model affects racial disparities, relative to the relevant counterfactual of human decision-making. Our context is the implementation of the Allegheny Family Screening Tool (AFST), a tool that aims to help child protection workers decide which allegations of maltreatment to screen in for investigation. Using difference-in-differences and triple difference designs, we find that the AFST reduced disparities in screening decisions and home removal rates for investigated referrals involving Black vs. white children. An analysis of false positive and false negative rates suggests that this reduction in disparities was achieved while improving welfare among both Black and white children.

JEL Codes: J13, J15, J18

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

†Contact e-mail: katherine.rittenhouse@austin.utexas.edu.

1 Introduction

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

We address this question in the context of the child protection system, an institution which interacts with approximately one third of U.S. children by the time they reach 18 (Kim et al. 2017). The child protection system is comprised of several high-stakes decision points, including whether to investigate reports of alleged maltreatment, and whether to remove children from abusive or neglectful homes. This is also a setting with large racial disparities – Black children are 88 percent more likely than white children to be investigated for maltreatment, and over twice as likely to enter foster care by the time they reach 18 (Kim et al. 2017; Wildeman and Emanuel 2014). These observed disparities may be driven by differences in the rates of maltreatment across race, and/or by biased decision-making. Recent work suggests that gaps are at least in part a product of discrimination: Baron et al. (2024) find in their context that "calls involving Black children are

[1] Discussing lessons learned from the movement towards algorithms in the justice system, Ludwig and Mullainathan (2021) note that: "the optimism for machine learning in criminal justice did not last long... Reports emerged of algorithms that were themselves discriminatory, producing racially disparate outcomes at a high enough rate that the phrase 'algorithmic bias' has entered the lexicon."

[2] See, for example, Goncalves and Mello (2021), Antonovics and Knight (2009), Rehavi and Starr (2014), Arnold, Dobbie, and Yang (2018), Abrams, Bertrand, and Mullainathan (2012), and Baron et al. (2024).

55% more likely to result in foster care placement than calls involving white children with the same potential for future maltreatment in the home," with up to 19% of this unwarranted disparity attributable to call screeners.

Predictive risk models have been introduced as a tool to improve quality and consistency of decision-making in child protection systems, but the use of these tools has been stalled by concerns about the impact of algorithms on racial disparities.[3] Improving racial equity is a top priority for child protection agencies across the U.S.[4] The United Nations has expressed "[concern] at the disproportionate number of children belonging to racial and ethnic minorities who are removed from their families and placed in foster care," and the Biden administration has noted the role of "systemic racism and economic barriers" in creating these disparities (*Concluding observations on the combined tenth to twelfth reports of the United States of America* 2022; Biden 2021). However, there is no consensus around how to improve racial equity while maintaining standards of child safety and protection.

We study the implementation of the Allegheny Family Screening Tool (AFST), the first automated predictive risk model used to aid decision-makers in the child protection system. This algorithm aims to help workers decide whether to investigate ("screen in") referrals alleging that a child is being maltreated.[5] The AFST uses information about the referred families from linked administrative data to predict the risk that the child will be removed from their home if screened in, and shows the human decision-maker a risk score ranging from 1 to 20. For referrals with the highest risk scores, the AFST activates a high-risk protocol, defaulting the referral to a screen-in recommendation. In all cases, the call-screening supervisors remain the ultimate decision-maker, and may choose to override the default. Screened-in referrals proceed to an investigation, con-

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

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

[5]Historically, around half of all calls are screened out.

ducted by a different caseworker. In severe cases where the caseworker determines that a child's safety is at risk, investigations may result in a child being removed from their home and placed in foster care. We ask how the implementation of the AFST affected: (1) the differential probability of screening in referrals involving Black vs. white children, and (2) the differential probability of home removal for referrals involving Black vs. white children.

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

We first establish the extent of racial disparities in screening rates in our setting prior to the implementation of the AFST. Referrals involving Black children are 11 percentage points (25%) more likely than referrals involving white children to be screened in for an investigation. Of course, disparities may be warranted if the screen-in rate reflects differences in the riskiness of referred Black vs. white children. Controlling for the (unobserved) underlying risk score, referrals involving Black children are still 6 percentage points more likely to be screened in for investigation. Finally, following the approach of Baron et al. (2024), we estimate unwarranted disparities in screening rates. Referrals involving Black children are 8.8-10.5 percentage points more likely to be screened in than white children with the same potential for future maltreatment.

To study the effects of the AFST on disparities in screening rates we employ a difference-in-differences design, comparing referrals involving Black vs. white children, made before vs. after the AFST was implemented. We find that the AFST reduced unconditional disparities in screening rates by 31%, and disparities conditional on algorithm score by 48%. Effects are concentrated

among high-risk referrals, which are likely to be defaulted to a screen in decision as per the high-risk protocol.

Next, we study the effects of the AFST on disparities in foster care. For this analysis, we use two empirical strategies. First, we again use a difference-in-differences design comparing referrals involving Black vs. white children, before vs. after the AFST was implemented. Second, we use a triple difference design, comparing across referrals which were "treated" vs. not "treated" by the AFST. Specifically, under Pennsylvania law, referrals which include certain allegations are automatically screened in for investigation, and as such were not affected by the AFST. This design enables us to control for any trends across time which might differentially affect Black vs. white families. One drawback of the triple difference design is that the control-group referrals are expunged from the data prior to 2015, giving us a shorter timeframe to study. Both designs yield similar results, showing that the AFST reduced disparities in home removals within three months by approximately 1.7 percentage points (48% of the pre-existing gap). Again, effects are concentrated among high-risk referrals.

Next we discuss potential mechanisms, and provide evidence of welfare improvements among both Black and white children. We show that variation in screening disparities across workers is lower after the implementation of the AFST, suggesting that the algorithm may have improved consistency in decision-making. In theory, improvements in equity have ambiguous welfare effects. We first show that *unwarranted* disparities were reduced after the AFST, in particular for high-risk referrals subject to the high-risk protocol. We investigate the potential welfare implications by studying the rates of false negatives and false positives, defined using downstream removals. False positives and false negative rates drop for both Black and white children after the implementation of the AFST. The reduction in false negatives is driven by high-risk referrals, while the reduction in false positives is driven by low- and medium-risk referrals.

We 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 dis-

parities. Several high-profile media reports have drawn attention to the potential for algorithms to discriminate.[6] Academic research has in many cases validated these concerns, documenting and exploring the consequences of algorithmic bias in a variety of contexts, including healthcare (Obermeyer et al. 2019) and the criminal justice system (Arnold, Dobbie, and Hull 2021). For policy-makers, a relevant question is whether algorithms *worsen* disparities, relative to the human-only systems they replace. Previous work addressing this question is limited, and has found mixed results (Stevenson and Doleac 2021; Albright 2019; Howell et al. 2021).

In highly related work, Grimon and Mills (2022) provide child protection system workers with randomized access to an algorithmic risk score. They find that providing access to the score reduced child injury hospitalizations. The pattern of their findings suggest that providing access to the algorithm score allows screeners to focus on other salient aspects of the allegations. Grimon and Mills (2022) also show that access to the tool reduced racial disparities in CPS contact, in particular among low-risk children. Our institutional context differs along several key dimensions. Their setting has a relatively small sample of Black families, with Black children making up four percent of CPS-involved children. We study disparities in a more diverse setting, where over 50% of referrals involve a Black child. Moreover, the high-risk protocol and associated default investigations in our setting allow us to speak to the effects of screening recommendations, as opposed to purely informational predicted risk scores.[7]

Prior studies in the computer science literature have studied the design and deployment of the AFST, as well as its possible implications for racial equity. Chouldechova et al. (2018) describe the development, validation, fairness auditing and deployment of the AFST. De-Arteaga, Fogliato, and Chouldechova (2020) study the effect of the AFST on call-screeners' decisions, finding that humans updated their behavior to align more closely with the risk score. They also study a period when a technical glitch led to some incorrect scores shown to call screeners, and find that screeners

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

[7] Albright (2024) shows that algorithmic recommendations may have importantly different effects from algorithmic predictions.

were less likely to adhere to the algorithm's recommendation in this case.[8] Finally, Cheng et al. (2022) compare screening decisions under the AFST to a theoretical setting where the AFST is used without human input. However, this exercise is purely theoretical, as the AFST was not designed to be used without human supervision. In contrast, our paper studies the effects of the AFST as it was used in practice, relative to the prior decision-making process.

The rest of the paper proceeds as follows. In Section 2 we describe the institutional context of the Allegheny County child protection system, as well as the Allegheny Family Screening Tool. Section 3 describes our data and Section 4 our empirical strategies. Section 5 presents and discusses results, Section 6 considers mechanisms and welfare implications, and Section 7 concludes.

2 Allegheny County

Allegheny County, Pennsylvania is home to 1.2 million people, and includes the city of Pittsburgh within its boundaries. While 13.5% of the population in Allegheny county is Black or African American, over 50% of referrals to the child protection system involve a Black child. The Office of Children, Youth and Families in Allegheny County is responsible for investigating allegations of child neglect and abuse. In Pennsylvania, the child protection system is State-supervised, and county-administered.

2.1 Referral Process

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

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

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

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

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

Other than the subset of referrals which are automatically screened in for investigation, a family's involvement with the child protection system is determined by the screener's decision. Prior

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

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

to August 2016, screeners relied solely on professional judgement to recommend which of the GPS referrals to screen in. In making this recommendation, screeners could use both information from the referral itself (e.g., reporter, allegations, age of children), as well as data on each child and adult included in the referral from the the linked Allegheny Data Warehouse (e.g., a child's history of foster care placements, adult arrest records). The Data Warehouse provides individual-level information on previous child protection system involvement, as well as involvement in a range of other County systems.[11] However, while these data were available and the County expected information to be systematically reviewed, there was little guidance for screeners on exactly how those data should be incorporated into their screening decision.[12] There was also no way for the County to confirm whether or not a screener had reviewed data to inform their decision. Call screeners' recommendations are reviewed and approved by a supervisor. Note, after making their recommendation, call screeners would not learn about any outcomes for investigated or screened-out families.[13] As such, there was little opportunity for improvement in decision making.

2.2 Referral Process and the Allegheny Family Screening Tool

In August 2016, Allegheny County implemented the AFST, a predictive risk model to help screeners decide which referrals to screen in for investigation. Note, the AFST score is generated for both CPS and GPS referrals, but is only included in the decision-making process for GPS referrals. CPS referrals are screened in 100% of the time both before and after AFST implementation. Further details on the design and implementation of the algorithm are included in subsection 2.3 below. After reviewing the allegations, 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

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

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

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

risk) and 20 (highest risk). Figure 2a shows an example of what the screener would see. The score is generated using only information that the screener has access to, but may not have time to review in detail and may not know how to incorporate into their assessment of safety and risk.

The use of the AFST is informed by two default decision protocols. For referrals with a score greater than 17 and at least one child aged 16 or under, the screener sees a "High Risk Protocol" notification, with no numeric score (see Figure 2b). These referrals are defaulted to be screened in for investigation, and require explicit supervisor approval to be screened out. For referrals with a score less than 11 and no children under the age of 12, the screener sees a "Low Risk Protocol" notification, again with no numeric score (see Figure 2c). These referrals are recommended to be screened out, but no supervisor approval is required to screen in these referrals.[14] Low-risk protocols initially made up only 4% of referrals (Vaithianathan et al. 2017), and as such have a limited possible impact on overall screening rates. For referrals which do not meet the criteria for high- or low-risk protocols, the screener observes the numeric score and makes their own decision; i.e., there is no default screening decision.

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

2.3 Allegheny Family Screening Tool

For each child associated with a referral, the AFST predicts the risk that, if screened in for investigation, that child will experience a court-ordered removal from their home within two years. The

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

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 the child protection system, past and present involvement with the courts, jail and other County systems, as well as information from the child's birth record, are all used to generate a risk score.[15] A risk score is generated for each child living in the home of the alleged victim, but the screener only observes the maximum of these scores.[16]

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

In August 2016 Allegheny County deployed the AFST. Since then, they have updated the predictive risk model and the screening tool several times. In November 2018, the original model was updated with weights estimated using LASSO.[18] In January 2019, the LASSO model was updated in response to a change in the data that were available to the model. Goldhaber-Fiebert and Prince (2019) were contracted to conduct an independent impact evaluation of the original AFST, and studied effects on accuracy, workload, disparities, and consistency.[19]

[15] A full list of the features used in the latest version of the algorithm can be found in Vaithianathan et al. (2017). Race is not included as a predictor.

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

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

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

[19] This evaluation used interrupted time-series methods, and a follow-up period of 17 months. Although limited in their power to detect statistically significant effects, their analysis suggests that the AFST increased the accuracy of screen-in decisions, in particular for referrals involving white children.

3 Data

We obtained de-identified administrative data from Allegheny County on the universe of child maltreatment referrals (both CPS and GPS) made between 2015 and 2020. For GPS referrals, we also obtain data on referrals made between 2010 and 2015.[20] We restrict our analysis to referrals made before October 2019, in order to observe at least a six-month follow-up for each referral before the child protection system was severely impacted by COVID-19 in March 2020. Every referral links to unique IDs for each person living in the household, including the alleged victim, other children, parents and alleged perpetrators. For each alleged victim, the referral lists one or more allegations of abuse or neglect. For each individual, we observe demographic information including race, gender, and age. We also observe outcomes within the child protection system, including screening decision and foster care placements. We observe a unique ID for each call screener, but do not observe any demographic or other information about the workers.

Screening decisions 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 approximately within this time. We also define race at the referral level, classifying a referral as Black if at least one child on that referral is identified as Black or African American, and classifying a referral as white if there are no Black children and at least one white child on the referral. Referrals with no children identified as either Black or white make up approximately 6% of referrals from 2010 through 2020 and are excluded from our analysis sample.

Table 1 presents summary statistics separately for GPS (all and screened-in only) and CPS

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

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

For each referral, we observe one or more algorithm-generated scores. First, we observe the score as calculated by the algorithm in use at the time of referral. In addition, we observe retroactively-calculated scores generated by AFST V3, the most recent version of the model as of this writing, for all referrals made after January 2013. Going forward, we primarily focus on the V3 score in order to allow for comparability across years. For referrals made between January 2013 and July 2019, this comparable score was retroactively calculated, and can be thought of as the score that the screener *would have* seen, had this version of the algorithm been deployed.[21] Scores are strongly correlated across AFST versions. Figure A1 shows a heatmap of the correlation between retroactively-calculated V3 scores and each of V1 and V2 scores.

4 Empirical Framework

4.1 Screening

To test whether the implementation of the algorithm changed disparities in screening rates we use a difference-in-differences-like approach, comparing GPS referrals involving Black children to those involving white children before and after the algorithm was implemented. While both groups were

[21] 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.

"treated" by the AFST, this approach tests the hypothesis that they were differentially affected.

We begin by estimating the following regression equation:

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

Where y_{it} is an indicator equal to one if referral i, made in month t, is screened in for an investigation; $Black_{it}$ is in indicator equal to one if there are any Black children listed on referral i, and zero if there are no Black children, and at least one white child, listed on referral i; and $Post_{it}$ is in indicator variable equal to one if referral i in month t was made after the implementation of the AFST, and zero otherwise. We include fixed effects for Year (γ_y) and Month-of-Year (γ_m), to control for variation across time and seasonality. Finally, X_{it} is a vector of referral-level controls, which includes allegation and reporter category indicators, indicators for child age groups, the number of children associated with the referral, and an indicator for any drug or alcohol concerns.[22] We include these variables to control for any differential trends across race and time. For example, substance exposure might differ across both race and time, as the opioid crisis has affected primarily white communities.

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

Finally, effects may differ across the range of algorithm scores. Given the low-risk and high-

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

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 classify scores as "low risk" (1-12), "medium risk" (13-17) and "high risk" (18-20).[23] This classification aligns with the risk score requirements of the most recent low- and high-risk protocols. We plot screening decisions for low, middle and high risk scores, before and after the AFST deployment and separately for referrals involving Black vs. white children, in Figure 3. We show this comparison first using the entire post period in Figure 3b, and then defining the post period as only the time when AFST V3 was in place (i.e. after July 2019) in Figure 3c. Note that the change to the racial gap in screen-in rates seems particularly stark for high-risk referrals. 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 risk bin $S^i, i \in \{1, 3\}$, estimating the equation:

$$\begin{aligned} y_{it} = &\sum_{j\neq 2}^{3} \beta_j S_{it}^j + \sum^{3} \beta_{j+2} S_{it}^j \times Black_i + \sum^{3} \beta_{j+5} S_{it}^j \times Post_t \\ &+ \sum^{3} \beta_{j+8} S_{it}^j \times Black_i \times Post_t + \beta_{11} X_{it} + \gamma_m + \gamma_y + \epsilon \end{aligned} \tag{2}$$

Where y_{it} is an indicator equal to one if a referral is screened in for investigation, and zero otherwise. This approach allows us to test how the algorithm affected disparities in screen-in rates differently by comparable risk score.

4.2 Removals

We next turn to the downstream outcome of home removal. For this analysis we conduct two empirical exercises, with different strengths and weaknesses. We start with a difference-in-differences approach. We estimate Equation 1, where y_{it} is an indicator equal to one if any child associated with referral i, made in month t, is removed from their home within three months. The identifica-

[23] We use the retroactively-generated V3 score for this classification, which allows us to compare like with like.

tion strategy requires the parallel trends assumption that removals evolve over time similarly for referrals involving Black vs. white children. We plot 3-month removals over time in Figure A3(a), and test this assumption directly using an event-study specification of Equation 1.

We also estimate a triple differences model, using CPS referrals as the control group. For this analysis we use a sub-sample of referrals made between 2015 and 2020, as CPS referrals made before 2015 are expunged in our data. The triple difference model is estimated with the following equation:

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

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

The identification assumption required for this triple differences specification is weaker than that for difference-in-differences, and requires common trends in racial disparities for CPS and GPS referrals. Removal trends are plotted for CPS referrals in Figure A3(b).

5 Results and Discussion

5.1 Screening

In Table 2 we report results from estimating Equation 1, or the effects of the algorithm on disparities in screen-in rates. Column (1) reports results from a difference-in-differences specification excluding fixed effects and controls, Column (2) adds year and month-of-year fixed effects, Column (3) adds referral-level controls, Column (4) replicates Column 3 for the time period in which we have AFST V3 scores, and Column (5) adds an additional fixed effect for the underlying AFST V3 score. First, note that referrals involving Black children are more likely to be screened in than referrals involving white children. There is a 10.9 percentage point gap in screen-in rates before controlling for referral-level characteristics, which is reduced to 6.4 percentage points after controlling for referral-level characteristics and underlying risk scores. The algorithm reduces the unconditional gap by 3.3 percentage points (Column 1), or 31%. It reduces the conditional gap by 3.1 percentage points (Column 5), or 48%.

Figure A4a presents coefficients from a quarterly event study version of Equation 1, where $Black_i$ is interacted with quarterly indicator variables. While the estimated coefficients are noisy, there appears to be a shift in average screen-in disparities around the time of the AFST deployment.[24] A version of the event study aggregated to the annual level presents a less-noisy story, suggesting screening disparities dropped significantly the year the AFST was introduced, and rebounded over time (see Figure A4b).[25]

Motivated by the patterns in screening disparities observed in Figure 3, we next study how effects on screening decisions vary across algorithm scores. Figure 4 shows coefficients and 95% confidence intervals from estimating Equation 2. Figure 4a shows a positive relationship between

[24]The omitted period (2016q2) has disparities that are lower than average for the pre-period, giving the appearance of significant effects in the pre-period, and null effects in the post period. However, we do not observe a consistent trend in the pre-period.

[25]Note, in the annual event study, the excluded year is 2015 and the first treated year (2016) includes several untreated months (January - July).

algorithm score bin and screen-in rate. Figure 4b shows that referrals involving Black children are more likely to be screened in at every risk score. Figure 4c suggests that the algorithm primarily increased screen-in rates for referrals with the highest risk scores, consistent with the intended effect of the high-risk protocol. Finally, Figure 4d graphically presents the coefficients on $Black \times Post \times Score^i$. The effect on screening disparities is negative for every score bin, with varying magnitudes. For referrals in the highest risk bin, the algorithm reduced the gap in screening rates across race by 6.0 percentage points, or 65% percent of the pre-existing disparity in this score bin.[26] An annual event study restricted to the high-risk bin is presented in Figure A4c suggests that these reductions in disparities were sustained over time.[27]

Referrals within this high-risk bin are most often defaulted to be screened in through the high risk protocol under AFST (see Section 2.2).[28] 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.

5.2 Removals

We next turn to the effects of the algorithm on home removals. We start by estimating a difference-in-differences model, using only the GPS referrals. Results from estimating Equation 1, where y_{it} is an indicator for home removal within three months of referral are reported in Column (1) of Table 4. Referrals involving Black children are on average 3.7 percentage points more likely to involve a child who is removed from their home within 3 months (72 percent of the sample mean). The coefficient on $Black \times Post$ suggests that the implementation of the algorithm reduced disparities in removal rates by 1.9 percentage points, or 51% of the pre-existing Black-white gap.

[26] In unreported results from this regression, the coefficient on $Black \times Post \times Score^3$ is -0.0601 and the coefficient on $Black \times Score^3$ is 0.0919 (each significant at the 1% level).

[27] Note that this event study is restricted to years in which we have retroactively-calculated AFST scores (2013 on).

[28] Since we use the comparable AFST V3 score, the high-risk bin does not correspond exactly with the high-risk protocol. That is, there may be referrals with an AFST V3 score of 18-20, but a screener-observed score (from an earlier algorithm version) below 17.

Figure A6a presents coefficients from a quarterly event study versions of Equation 1. Again, there appears to be a shift in removal disparities around the time of AFST implementation. An annual version of the event study presents a less-noisy picture (Figure A6b), showing a sharp drop in disparities beginning in 2016.

Next, we incorporate CPS referrals as a control group in a triple differences specification. For this analysis, we must restrict our sample to referrals made in or after 2015. For easier comparison across specifications, we also run a difference-in-differences model using this shorter time span. Results from estimating Equation 1 on this restricted sample are reported in Column (2) of Table 4, and are statistically indistinguishable from the results in Column (1).

Results from estimating Equation 3, where y_{it} is an indicator for removal within three months of referral, are reported in Columns (3) through (5) of Table 4. Column (3) reports results from a basic triple differences specification, Column (4) reports results from a regression which adds year and month-of-year fixed effects, and Column (5) reports results from a regression which adds referral-level controls (our preferred specification). Our coefficient of interest, on the triple interaction $Black \times Post \times GPS$, is significant at the 10% level, and stable across specifications. In our preferred specification (reported in Column 5), the coefficient implies that the introduction of the algorithm reduced the racial disparity in three-month removal rates by 1.7 percentage points, or 48% of the pre-existing difference.[29]

To study where in the risk distribution effects are concentrated we estimate Equation 2, setting the outcome variable equal to an indicator for removal within three months of referral. For this analysis we restrict our sample to referrals made in or after 2013, as we have comparable scores for these dates. The coefficients on $Score_i$, $Black \times Score_i$, $Post \times Score_i$ and $Black \times Post \times Score_i$ are shown in Figure 7. In Figure 7b, the race differential in likelihood of removal is most stark for the highest-risk referrals. The effect on disparities is also concentrated in the highest risk bin,

[29]To calculate the pre-existing difference in 3-month removal rates for GPS referrals involving Black vs. white children, we sum the coefficients on $Black$ and $Black \times GPS$. 0.0264 + 0.00893 = 0.03533.

as shown in Figure 7d. An annual event study specification restricted to high-risk referrals is presented in Figure A6c, and shows a sharp and sustained drop in disparities beginning in 2016.

6 Mechanisms and Welfare

One goal of using predictive risk models in the child welfare setting is to improve consistency across decision-makers. We explore how the AFST affected individual call screeners, studying racial disparities by worker before and after the AFST is implemented. We first limit our sample to call screeners who handle at least 20 calls in each of the pre and the post period. For each of these twenty-nine screeners, we calculate the raw gap in screening rate by race.

Figure 6 plots the distribution of disparities across screeners, before and after the AFST is implemented. In the pre period, there are several workers with large gaps in screening rates by race. In the post period, there is less variation in disparities across screeners. This pattern is consistent with the AFST increasing consistency across workers, and reining in screeners whose disparities are furthest from the mean.[30]

One may be concerned that a reduction in observed disparities is harmful if the disparities were in fact warranted. To address this concern, we adopt a measure of unwarranted disparities proposed in Baron et al. (2024). We estimate unwarranted disparities in screening rates before and after AFST implementation.[31] We define unwarranted disparity as differences in screening rates across race conditional on potential for future maltreatment, where future maltreatment is proxied for by foster care placement within 6 months if left at home.[32] Given that screeners do not appear to be randomly assigned in our setting, we estimate bounds for unwarranted disparities,

[30] An F-test comparing variance in disparities across the two distributions results in an f-statistic of 4.7, strongly rejecting the null hypothesis that variance is equal across the two distributions.

[31] We calculate unwarranted disparities following the methodology of Baron et al. (2024)

[32] This is a different definition than Baron et al. (2024), who use investigation within six months as a proxy for subsequent maltreatment. We chose this outcome as it is not directly influenced by the AFST. Results are similar when using future investigations.

again following the approach in Baron et al. (2024).[33]

Prior to the implementation of the AFST, referrals involving Black children were 8.8-10.4% more likely to be screened in than referrals involving white children with the same potential for future maltreatment. After the implementation of the AFST this rate dropped to 6.2-7.2%. Figure A5 shows the ranges of unwarranted disparities in the pre- and post-periods, for each risk bin. Note that pre-AFST unwarranted disparities are largest in the high-risk bin, and reductions in unwarranted disparities after the implementation of the AFST are largest in this risk bin as well, consistent with the effects on raw screening disparities. This analysis again suggests that the high-risk protocol, which defaults referrals above a certain risk score to an investigation, plays an important part in minimizing the role of human bias in decision-making.

Still, the welfare effects of reductions in disparities are *a priori* ambiguous. A reduction in unwarranted disparities implies more consistent, but not necessarily more accurate decision-making. If equity is achieved through a reduction in screen-ins for children who would benefit from an investigation, or an increase in screen-ins for children who do not benefit from an investigation, children may be harmed even by a reduction in unwarranted disparities. We do not know the costs and benefits of investigation, or how those values differ across race and other observable characteristics. Instead, we turn to proxy measures in order to classify screening decisions as "false negatives" (referrals which were screened out, but would have led to benefits for the children if screened in) and "false positives" (referrals which were screened in unnecessarily).

We define a referral as a false negative if it is screened out and any child from the referral is placed in foster care within six months. This implies that the child was left at home in circumstances that led to severe maltreatment. Similarly, we define a referral as a false positive if it was screened in and no child from the referral is placed in foster care within six months. This measure is more difficult to interpret, as it may be the case that children benefit from an investigation even if they are not placed in foster care. Earlier intervention may even help to reduce the need for home

[33]Details of this calculation are included in the notes to Figure A5.

removal. A reduction in false positives as defined here would suggest that fewer non-severe cases of maltreatment are investigated, while an increase in false positives would indicate that a wider net is being cast. Table 3 presents false negative and positive rates, before and after the implementation of the AFST and separately for referrals involving Black vs. white children. False negative rates fall by 22% for referrals involving white children, and by 11% for referrals involving Black children. False positive rates fall by 4% for both referrals involving white and Black children. We are also able to study false positive rates for our control group of CPS referrals. Among this group, false positive rates actually rose slightly for both referrals involving Black and white children. This is reassuring in that the patterns observed among the treated GPS referrals cannot be explained by overall trends in reporting or removals that would affect both GPS and CPS referrals. Unfortunately, we are not able to conduct a similar placebo analysis for false negative rates, as no CPS referrals are screened out.

We further investigate these patterns by risk bin in Figure 5. Panels (a) and (b) show false negative rates for referrals involving Black and white children, respectively. Among low- and medium-risk referrals, false negative rates are similar before and after the AFST for both referrals involving Black and white children. However, false negative rates are reduced among high-risk referrals for both race groups. This aligns with the purpose of the high-risk protocol, which defaults referrals above a certain score to be screened in for investigation. Indeed, it seems that the algorithm is catching cases of severe maltreatment that were previously missed. Panels (c) and (d) show false positive rates for referrals involving Black and white children, respectively. The patterns are similar across race. For low- and medium-risk referrals, false positives decline in the post-AFST period. This suggests that within these risk bins, the AFST is improving targeting of investigations towards those who are at risk of foster care. However, false positive rates increase for those in the highest risk bin. Again, this is consistent with the high-risk protocol increasing screen-ins above a given score. The observed increase may represent a decline in welfare among this group if investigations with no subsequent foster care are costly to children and households.

However, it may be the case that investigations among this high-risk group are providing other benefits or interventions that reduce the need for foster care.

Welfare implications depend on the relative costs of false negatives and false positives among low, medium and high-risk referrals. In order for the observed patterns to be consistent with a reduction in welfare, the cost of the increase in false positives among high-risk referrals would need to outweigh the benefit of reductions in false positives among low- and medium-risk referrals, combined with the benefit of reductions in false negatives among high-risk referrals. This is unlikely to be the case. First, false positives as defined may be less costly for higher-risk referrals. Children on high-risk referrals may see benefits even from investigations that do not result in home removal, if they are provided with in-home services.[34] Even if these children do not benefit from an investigation which does not result in foster care, it is not clear why they would be more harmed by an investigation than lower-risk children. Second, false negatives as defined are likely to be highly costly, as they require that a child be screened out and left in a home where they experience subsequent maltreatment severe enough to warrant foster care placement. Given that observed patterns are similar across race, this analysis suggests that the AFST reduced disparities while improving welfare in each group.

Together, these patterns suggest that the AFST reduced the role of human bias in decision-making, particularly among referrals subject to the high-risk protocol. Crucially for understanding welfare effects, children on these high-risk referrals are exactly those who are most likely to benefit from intervention.

7 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

[34] In ongoing work, Lacey et al (2024) study the impacts of investigations for children at the high-risk protocol threshold.

exacerbate existing biases, or even create new ones. We show that predictive risk models can also serve to reduce racial disparities, relative to human decision makers.

Using a difference-in-differences and a triple difference design, we study the effects of the Allegheny Family Screening Tool on racial disparities in screening decisions and home removals. Relative to the prior decision-making approach, the implementation of the AFST reduced disparities in each of these outcomes. An analysis of false positive and false negative rates in screening decisions before and after AFST implementation suggests that the reduction in disparities was achieved while improving welfare both among Black and white children.

Policy-makers and communities are concerned about racial disparities in the child protection system, and are actively working to improve racial equity. Our work shows that predictive risk models may be a useful tool for reaching this specific policy goal.

References

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

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

———. 2024. "The Hidden Effects of Algorithmic Recommendations." *Working Paper.*

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

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

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

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

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

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

Baron, E Jason, Joseph J Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph P Ryan. 2024. "Discrimination in Multi-Phase Systems: Evidence from Child Protection."

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

Biden, Joseph R. 2021. *A Proclamation on National Foster Care Month, 2021.* https://www.whitehouse.gov/briefing-room/presidential-actions/2021/04/30/a-proclamation-on-national-foster-care-month-2021/.

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

Chohlas-Wood, Alex. 2020. *Understanding risk assessment instruments in criminal justice.* Technical report. Brookings Institution.

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

Concluding observations on the combined tenth to twelfth reports of the United States of America. 2022. Technical report. United Nations Committee on the Elimination of Racial Discrimination.

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

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

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

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

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

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

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

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.

Ludwig, Jens, and Sendhil Mullainathan. 2021. "Fragile algorithms and fallible decision-makers: lessons from the justice system." *Journal of Economic Perspectives* 35 (4): 71–96.

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.

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.

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.

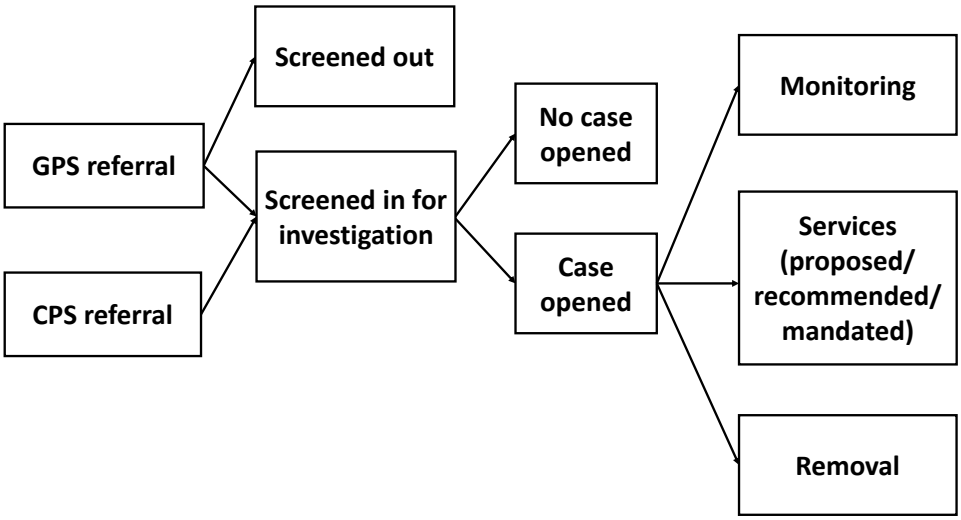
Vaithianathan, Rhema, Emily Kulick, Emily Putnam-Hornstein, and Diana Benavides Prado. 2019. *Allegheny Family Screening Tool: Methodology, Version 2.* Technical report. Centre for Social Data Analytics.

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

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

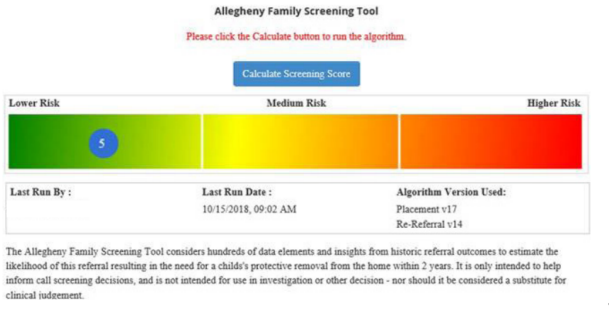
Figures

Figure 1: Referral process

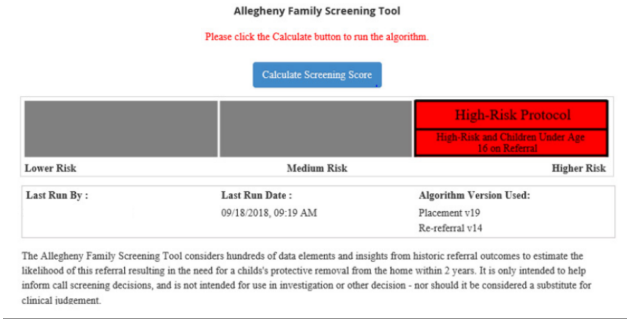


This figure presents the steps by which referrals to CYF move through the child protection system in Allegheny County, PA.

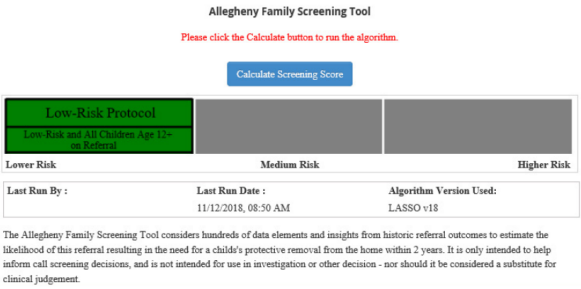
Figure 2: Screener View of AFST Output



(a) No protocol



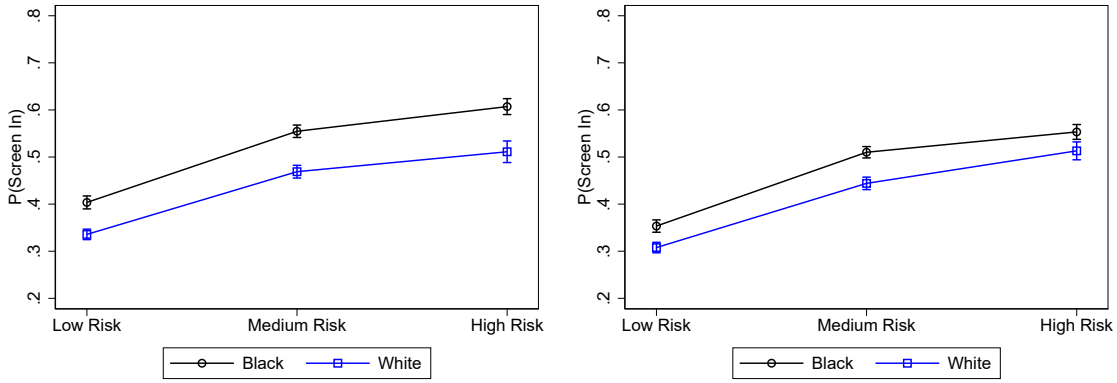
(b) High-risk protocol



(c) Low-risk protocol

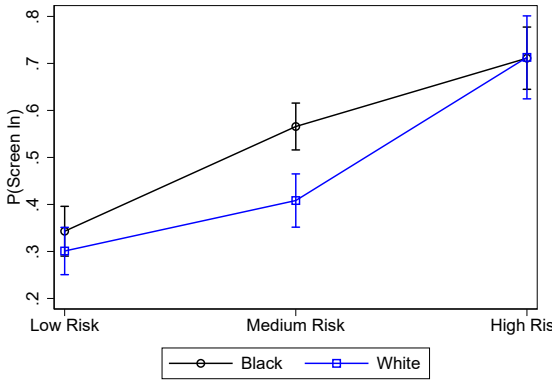
This figure presents the view that a screener has after running the algorithm, in three different scenarios. Panel (a) shows the output when neither the high-risk protocol nor the low-risk protocol is in place. Panel (b) shows the screener's view if the high-risk protocol is implemented (i.e. any child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

Figure 3: Screen In Disparities by Score Bin



(a) Pre-deployment

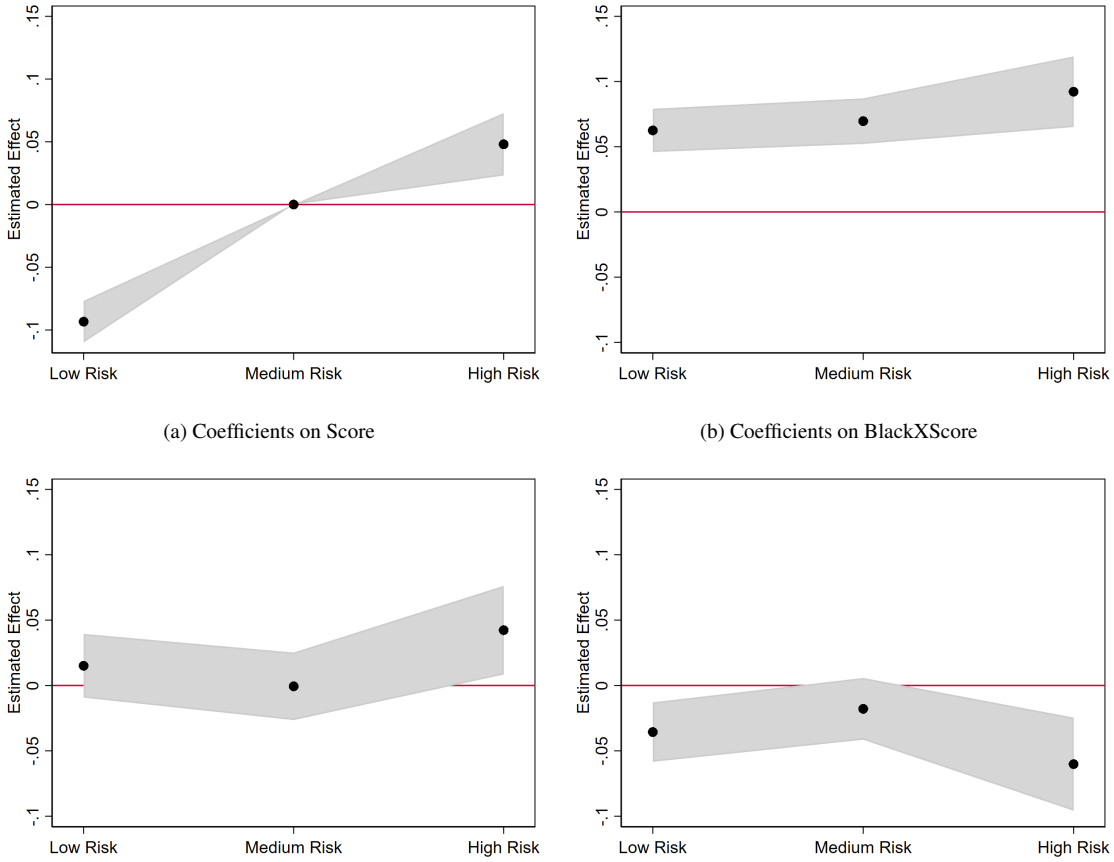
(b) Post-deployment



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

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

Figure 4: Screening by Score Bin



(a) Coefficients on Score

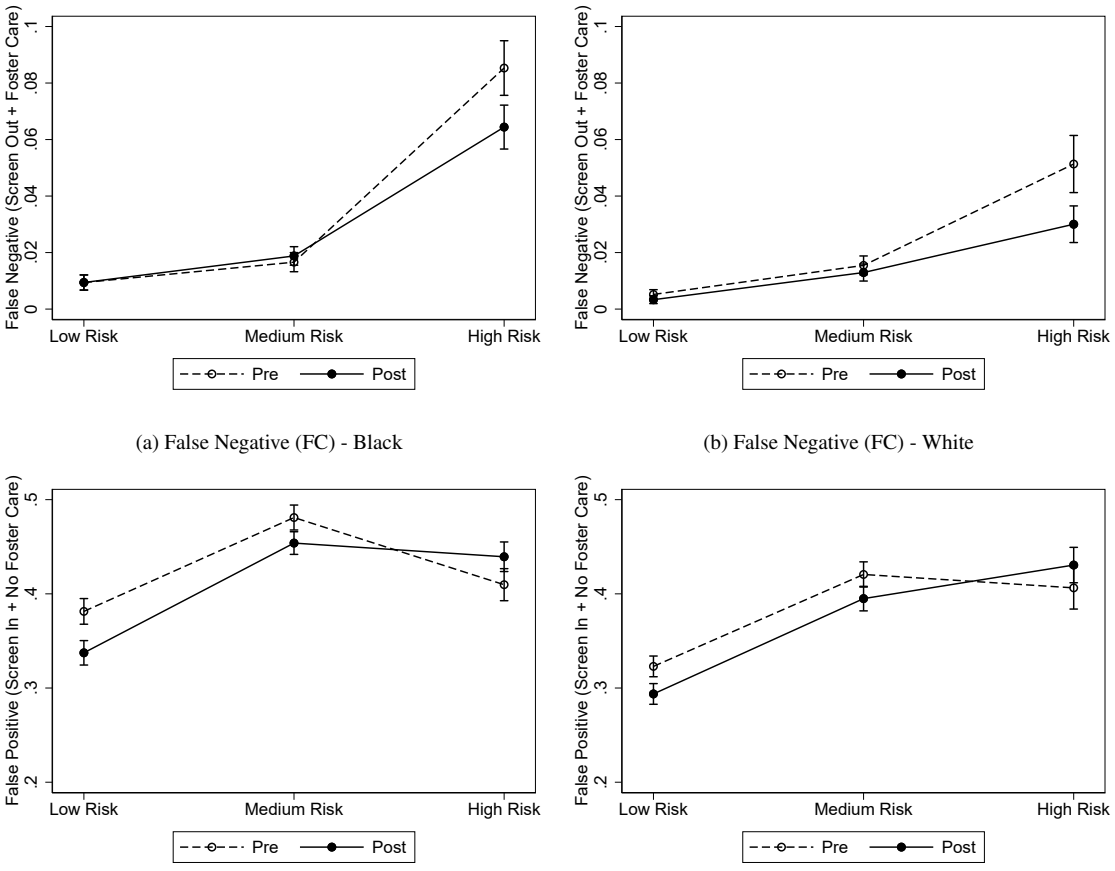
(b) Coefficients on BlackXScore

(c) Coefficients on PostXScore

(d) Coefficients on BlackXPostXScore

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

Figure 5: False Positives and False Negatives



(a) False Negative (FC) - Black

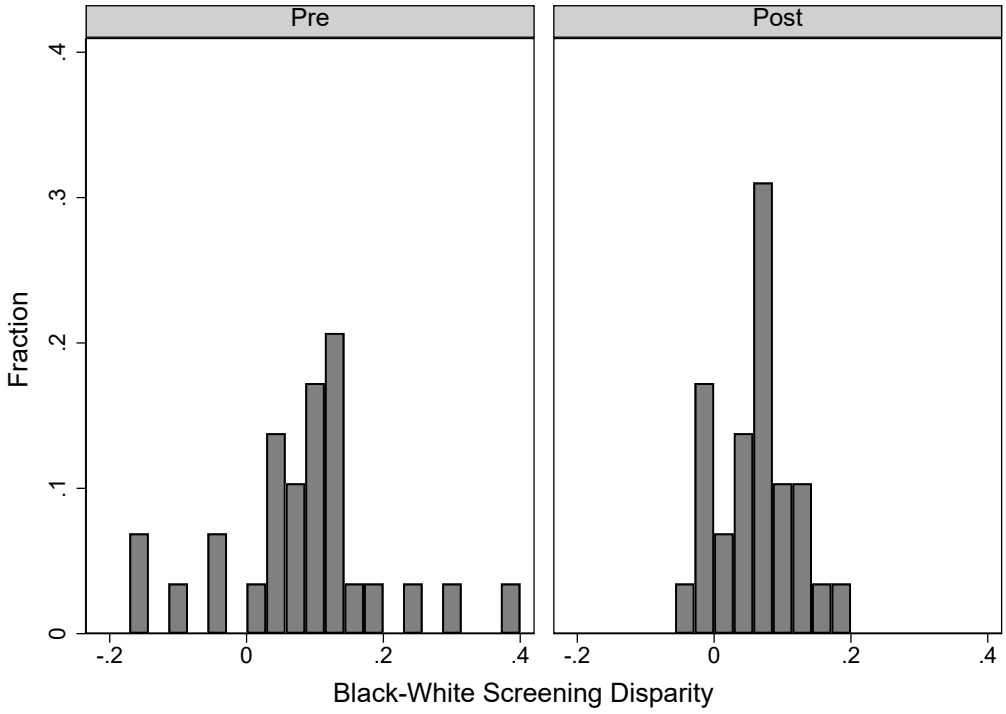
(b) False Negative (FC) - White

(c) False Positive (FC) - Black

(d) False Positive (FC) - White

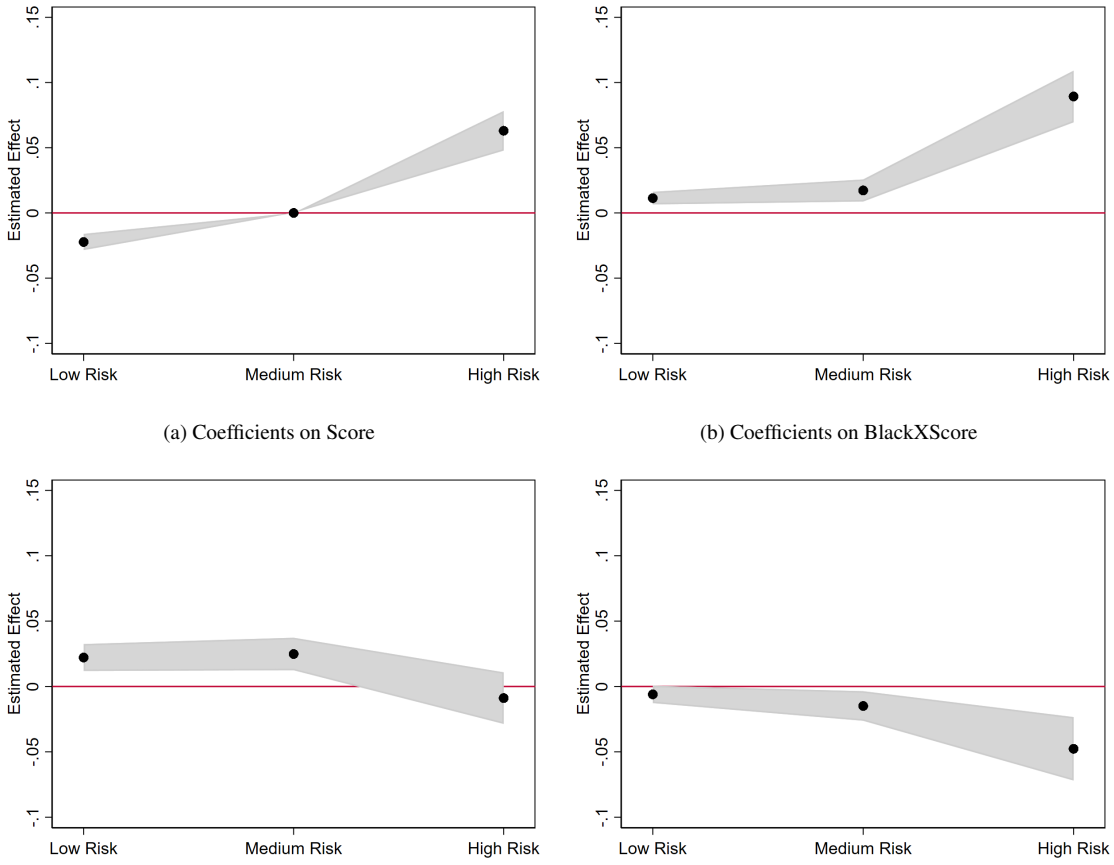
This figure presents rates of false positives and false negatives, as defined in section 6. In each subfigure, AFST risk bin is shown on the x-axis and false positive/negative rate is shown on the y-axis. Hollow circles and dashed lines represent the period prior to AFST implementation, while solid circles and solid lines represent the period after AFST implementation. Panels (a) and (c) present patterns for referrals involving Black children, while panels (b) and (d) present patterns for referrals involving white children. The time period is limited to referrals made between January 2013 and September 2019.

Figure 6: Distribution of Screening Gaps Across Workers



This figure presents the distribution of worker-specific racial disparities in screening rates, before and after AFST implementation. Worker-specific disparities are measured as the raw gap in screen-in rates, i.e., the average screen-in rate for referrals involving Black children minus the average screen-in rate for referrals involving white children. The sample is limited to twenty-nine screeners who handle at least 20 referrals in each of the pre and post period.

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



(a) Coefficients on Score

(b) Coefficients on BlackXScore

(c) Coefficients on PostXScore

(d) Coefficients on BlackXPostXScore

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

Tables

Table 1: Summary Statistics

	All GPS	*Screened-in GPS*	*All CPS*
Panel A: Demographics			
Black	0.498	0.552	0.510
Any Infant	0.150	0.221	0.118
Any Child Age 1-5	0.454	0.476	0.443
Any Child Age 6-12	0.582	0.587	0.636
Any Child Age 13-17	0.388	0.391	0.438
Number of Children	2.267	2.408	2.394
Panel B: Referral Outcomes			
Screened In	0.444	1.000	0.997
Removed within 3 months	0.054	0.092	0.053
Panel C: Allegation Categories			
Abandonment	0.009	0.011	0.003
Caregiver Behavioral Issues	0.035	0.047	0.012
Caregiver Substance Abuse	0.215	0.287	0.035
Causing Death of Child	0.004	0.006	0.007
Child Behaviors	0.055	0.060	0.023
Domestic Violence	0.068	0.081	0.026
Exposure to Risk	0.096	0.105	0.022
Failure to Protect	0.058	0.055	0.015
Imminent Risk	0.035	0.040	0.015
Inadequate Physical Care	0.294	0.265	0.032
Medical Neglect	0.036	0.046	0.017
Mental Health	0.051	0.040	0.024
Mental Injury	0.028	0.033	0.020
Neglect	0.108	0.108	0.014
No/Inadequate Home	0.108	0.139	0.012
Parent/Child Conflict	0.053	0.046	0.018
Physical Altercation	0.015	0.015	0.023
Physical Maltreatment	0.091	0.047	0.643
Sexual Abuse or Exploitation	0.020	0.010	0.177
Sexual Contact Between Children	0.040	0.024	0.005
Truancy	0.042	0.068	0.006
Unwilling or Unable to Provide Care	0.072	0.079	0.013
Youth Substance Abuse	0.011	0.009	0.003
Panel D: Reporter Categories			
Agency	0.217	0.229	0.274
Anonymous	0.125	0.088	0.032
Community	0.050	0.042	0.022
Family	0.172	0.163	0.068
Law	0.120	0.187	0.086
Medical	0.093	0.106	0.158
School	0.134	0.126	0.159
Self	0.004	0.004	0.012
Therapist	0.078	0.047	0.186
N	79725	35416	13925

This table reports means of referral characteristics separately for all GPS, screened-in GPS, and all CPS referrals. Note, since 100% of CPS referrals are screened in, averages for screened-in CPS referrals are identical to those in the full CPS sample. The sample is all referrals made between Jan. 2010 and Sep. 2019 to CYF, excluding active-family referrals, referrals involving neither Black children nor white children, and referrals stemming from truancy courts.

Table 2: Screen In

	(1) DD	(2) +FE	(3) +Controls	(4) Restricted Sample	(5) +Risk Score
Post	-0.00652	0.0252**	0.00988	0.0137	0.0139
	(0.00506)	(0.0120)	(0.0109)	(0.0112)	(0.0112)
Black	0.109***	0.108***	0.0852***	0.0873***	0.0641***
	(0.00446)	(0.00446)	(0.00419)	(0.00558)	(0.00563)
BlackXPost	-0.0334***	-0.0331***	-0.0280***	-0.0317***	-0.0310***
	(0.00720)	(0.00720)	(0.00662)	(0.00753)	(0.00751)
Year FE	No	Yes	Yes	Yes	Yes
Month-of-Year FE	No	Yes	Yes	Yes	Yes
Score FE	No	No	No	No	Yes
Controls	No	No	Yes	Yes	Yes
Data Span	2010 - 2019	2010 - 2019	2010 - 2019	2013 - 2019	2013 - 2019
Mean	0.444	0.444	0.444	0.444	0.446
Obs.	79725	79725	79725	58273	57825

Standard errors in parentheses

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

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

Table 3: False Negatives and Positives

	Pre	Post
False Negative Rates - GPS		
White	.0150	.0117
Black	.0302	.0270
False Positive Rates - GPS		
White	.3703	.3558
Black	.4282	.4120
False Positive Rates - CPS		
White	.9423	.9570
Black	.8927	.9143

This table reports false positive and false negative rates, as defined in section 6. The first two panels report rates among GPS referrals separately for referrals involving Black vs. white children, before and after AFST implementation. The third panel reports false positive rates for CPS referrals, separately for referrals involving Black vs. white children, before and after AFST implementation.

Table 4: Results: Placed within 3 months

	(1) DD	(2) Restricted Sample	(3) DDD Base	(4) +FE	(5) +Controls
Post	0.0149***	0.0182***	0.00494	0.0224***	0.0194***
	(0.00527)	(0.00562)	(0.00463)	(0.00620)	(0.00610)
Black	0.0371***	0.0357***	0.0294***	0.0297***	0.0264***
	(0.00210)	(0.00383)	(0.00654)	(0.00653)	(0.00634)
BlackXPost	-0.0190***	-0.0179***	-0.00217	-0.00238	-0.000880
	(0.00317)	(0.00448)	(0.00796)	(0.00796)	(0.00777)
GPS			0.00434	0.00444	-0.00942**
			(0.00427)	(0.00426)	(0.00451)
BlackXGPS			0.00961	0.00896	0.00893
			(0.00758)	(0.00758)	(0.00739)
GPSXPost			-0.000242	-0.000276	0.000626
			(0.00531)	(0.00531)	(0.00525)
BlackXPostXGPS			-0.0164*	-0.0158*	-0.0170*
			(0.00916)	(0.00915)	(0.00896)
Year FE	Yes	Yes	No	Yes	Yes
Month-of-Year FE	Yes	Yes	No	Yes	Yes
Controls	Yes	Yes	No	No	Yes
Data Span	2010 - 2019	2015 - 2019	2015 - 2019	2015 - 2019	2015 - 2019
Mean	0.0543	0.0469	0.0463	0.0463	0.0459
Obs.	79725	43399	55262	55262	55134

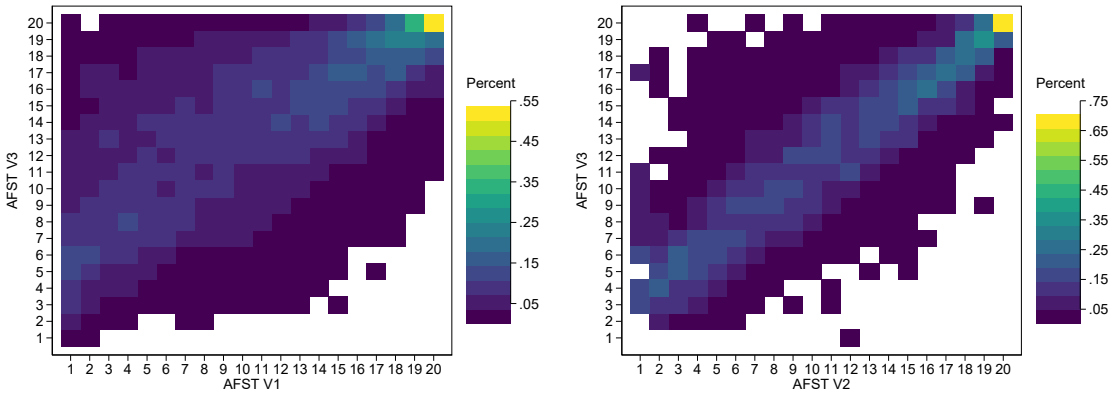
Standard errors in parentheses

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

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

Appendix Figures and Tables

Figure A1: Correlation Across AFST Versions

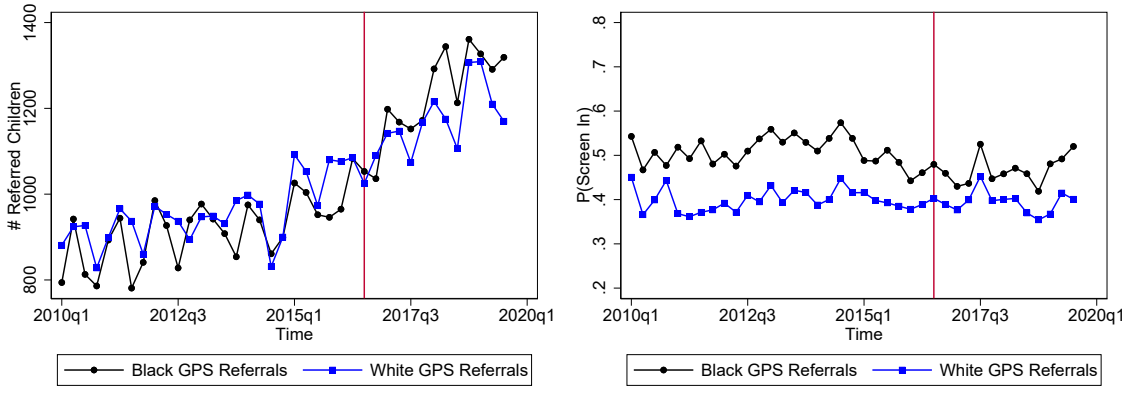


(a) V1-V3 Correlation

(b) V2-V3 Correlation

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

Figure A2: Time trends: Referrals and Screening decisions

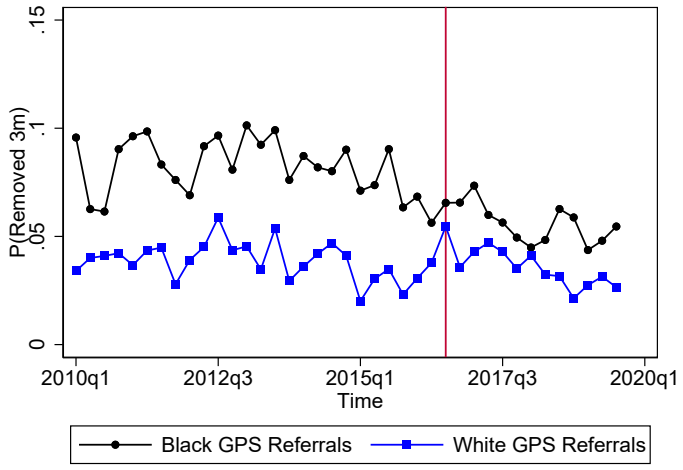


(a) Number of referrals

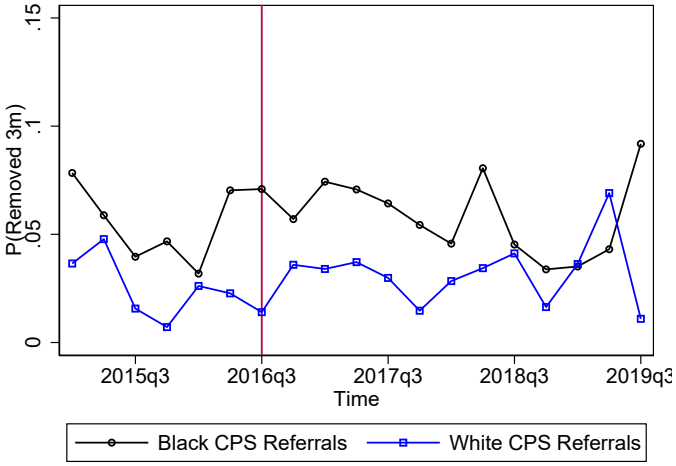
(b) Screening

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

Figure A3: Time trends: Removals



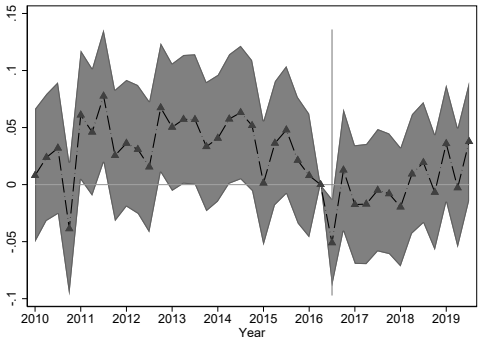
(a) Removal within 3 months - GPS



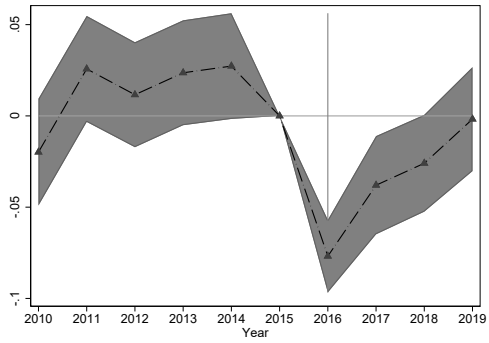
(b) Removal within 3 months - CPS

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

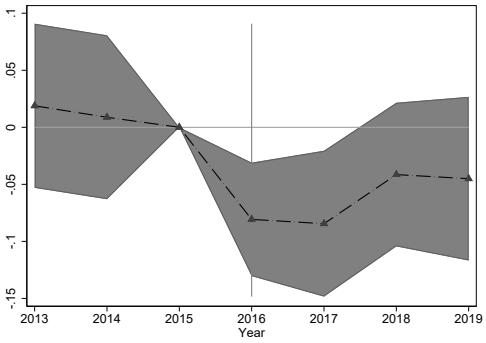
Figure A4: Event Study: Screening



(a) Quarterly all



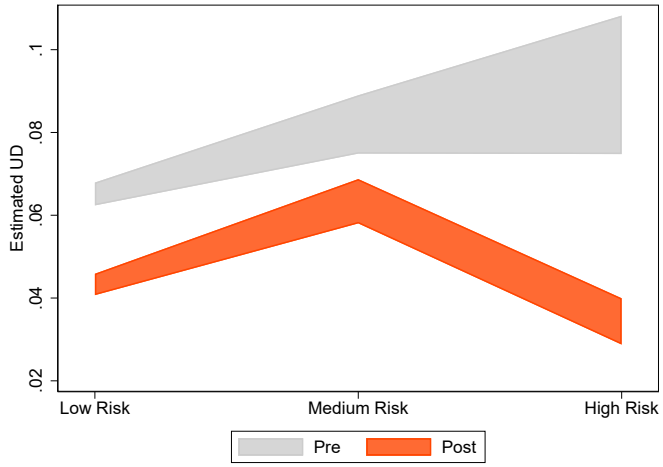
(b) Yearly all



(c) Yearly high risk

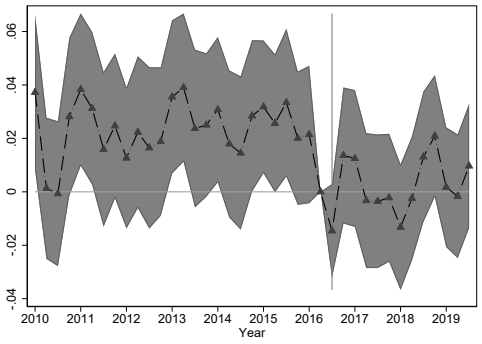
This figure presents coefficients and 95% confidence intervals from estimating an event study version of Equation 1, where the outcome variable is an indicator equal to one if the referral was screened in and zero otherwise. Panel (a) presents a quarterly event study, where 2016Q2 is the excluded quarter. Panels (b) and (c) present annual event studies, where the omitted year is 2015. Panel (c) restricts the sample to high-risk referrals (i.e. with AFST scores above 17). Retroactively scores are only available beginning in 2013.

Figure A5: Unwarranted Disparities

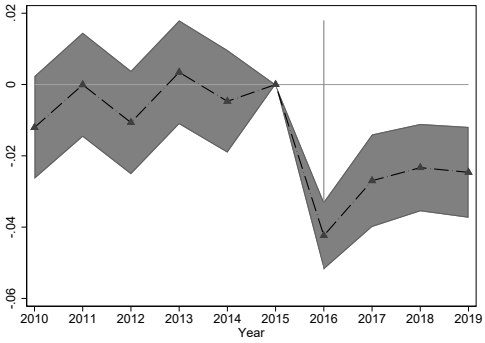


This figure presents unwarranted disparity bounds before and after the AFST, separately by referral risk score. Unwarranted disparities in screening decisions are estimated using methodology from Baron et al. (2024). More specifically, we define subsequent maltreatment as foster care placement within six months if left at home. That is, subsequent maltreatment potential is equal to zero if a referral is screened out and no child on that referral is placed in foster care within six months; zero if a referral is screened in and no child is placed in foster care within six months; one if the referral is screened out and a child is re-referred and placed within six months, and one if the referral is screened in, not immediately placed, re-referred and subsequently placed within six months. To estimate bounds on unwarranted disparities in our setting without random assignment of call screeners, we consider two hypothetical worlds: (1) where all children who are screened in and placed in foster care would have been subsequently maltreated if left at home, and (2) where all children who are screened in and placed in foster care would *not* have been subsequently maltreated if left at home. We use these values to obtain maximum and minimum levels of race-specific average subsequent maltreatment levels in the population. Using 10 equally-spaced values between the maximum and minimum values of race-specific population maltreatment, we calculate 100 estimates of unwarranted disparities in each risk bin, before vs. after AFST implementation, following equations (3), (10) and (11) from Baron et al. (2024). The figure shows the range of unwarranted disparities calculated in each time period and risk bin.

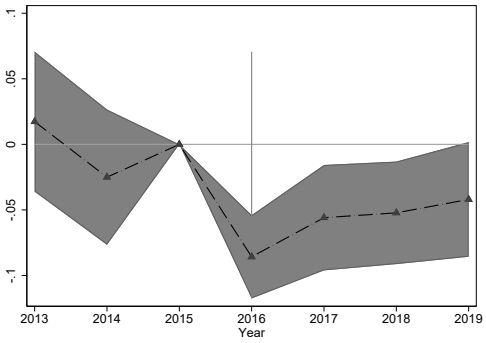
Figure A6: Event Study: Removals



(a) Quarterly



(b) Yearly



(c) High-risk only

This figure presents coefficients and 95% confidence intervals from estimating an event study version of Equation 1, where the outcome variable is an indicator equal to one if any child on the referral was removed within three months and zero otherwise. Panel (a) presents a quarterly event study, where 2016Q2 is the excluded quarter. Panels (b) and (c) present annual event studies, where the omitted year is 2015. Panel (c) restricts the sample to high-risk referrals (i.e. with AFST scores above 17). Retroactively scores are only available beginning in 2013.