Unwarranted Disparity in High-Stakes Decisions: Race Measurement and Policy Responses*

E. Jason Baron† Joseph J. Doyle, Jr.‡ Natalia Emanuel§
Peter Hull¶ Joseph Ryan‖

February 2024

Abstract

Studies of racial discrimination often condition on endogenous measures of race or on earlier decisions that might themselves be affected by discrimination. We develop quasi-experimental tools for estimating the impact of racial misclassification on measures of unwarranted disparity, and for designing policy responses to unwarranted disparity that account for discrimination in earlier decisions. We apply these tools to the setting of child protective services (CPS), where previous work has found that Black children are placed into foster care at higher rates than white children with identical potential for future maltreatment. CPS investigators regularly misclassify both Black and white children relative to their self-reported race, and this misclassification obscures around 24% of unwarranted disparity in foster care placement decisions. Policies that use algorithmic recommendations to eliminate total unwarranted disparity in placement rates are also meaningfully affected by earlier discrimination in CPS call screening.

*We thank Randall Akee, Larry Katz, and other participants in the Summer 2023 CRIW Pre-Conference for helpful comments.

†Duke University and NBER. E-mail: jason.baron@duke.edu.
‡MIT and NBER. Email: jjdoyle@mit.edu.
§Federal Reserve Bank of New York. Email: natalia@nataliaemanuel.com.
¶Brown University and NBER. Email: peter_hull@brown.edu.
‖University of Michigan. Email: joryan@umich.edu.

1 Introduction

There are widespread concerns about racial discrimination in many high-stakes settings, from criminal justice to employment, lending, healthcare, and child protection. Quantifying discrimination in such settings can be challenging however, since observed racial disparities may reflect omitted variables bias (OVB) as well as discrimination. For example, in the child protection setting, widely documented racial disparities in foster care placement rates might signal the kind of discrimination that both the United Nations and the American Bar Association have recently called on U.S. policymakers to address.[1] But these racial disparities might also capture underlying differences in the need for intervention—specifically, differences in the risk of future abuse or neglect among Black and white children referred to child protective services (CPS). Accounting for such unobservables is essential for measuring unwarranted racial disparities and for forming appropriate policy responses.

A recent literature develops and applies quasi-experimental tools for addressing this OVB challenge in various contexts like pretrial release (Arnold, Dobbie, and Hull, 2022) and child protection (Baron et al., 2024, henceforth BDEHR). In contrast to conventional experimental studies of direct discrimination (e.g., audit or correspondence studies), which address OVB by conditioning on all observable non-race characteristics, these quasi-experimental studies condition on a latent measure of an individual's need for intervention. BDEHR, for example, use quasi-random caseworker assignment to estimate unwarranted disparities (UDs) in foster care placement rates that condition on a child's potential for future at-home maltreatment. Such an approach yields broader measures of discrimination, which can account for indirect drivers that conventional studies typically condition on. Indeed, BDEHR show how UDs arise and perpetuate across the multi-stage CPS system from initial call screeners and subsequent investigators, and how analyses that condition on the call screening phase miss a meaningful share of discrimination. These kinds of analyses follow a growing interest in understanding how popular direct discrimination measures can obscure various forms of indirect or "systemic" discrimination (e.g., Bohren, Hull, and Imas, 2022).

This study develops further quasi-experimental tools to quantify the role of indirect drivers of discrimination that conventional studies often condition on. As in BDEHR, we apply these tools to the CPS setting. The foster care setting is particularly important as CPS involvement is both widespread and disparate in the U.S. By the time they reach the age of 18, 37% of U.S. children become the subject of formal CPS investigations due to alleged abuse or neglect.

[1] See Kelly, J (2022), *UN Committee Suggests the US Change or Repeal Major Child Welfare Policies*, The Imprint. Accessed at: https://imprintnews.org/ (5/6/2023); White, S and Persson, S (2022), *Racial Discrimination in Child Welfare Is a Human Rights Violation–Let's Talk About It That Way*, The American Bar Association. Accessed at: https://www.americanbar.org/ (5/6/2023).

Additionally, 5% of children are placed in foster care at some point during their childhood as a result of these investigations. Contact with CPS is significantly more prevalent among Black children, with over half being investigated and up to 9% entering foster care by age 18 (Yi, Edwards, and Wildeman, 2020; Kim et al., 2017). Given the large documented impacts of foster care placement on later-in-life outcomes like earnings, educational attainment, and contact with the criminal justice system (Bald et al., 2022), there is significant interest in these racial disparities and the extent to which they may reflect discrimination.

We conduct two main analyses in this setting. First, we consider the role of race or ethnicity misclassification in the administrative datasets often underlying discrimination studies. In CPS systems, investigators typically record a child's race after conducting interviews with the child and family. Researchers like BDEHR use this investigator-coded race to study discrimination, often discarding certain racial codes from their analysis (e.g., dropping individuals who are not classified as either Black or white). A concern with these sorts of studies is that conditioning on such "endogenous" racial classification decisions might obscure the level of true discrimination (Luh, 2022; Finlay, Luh, and Mueller-Smith, 2024). One could imagine, for example, investigators either consciously or subconsciously coding riskier Black children as another race or as multiracial—causing them to be discarded from analyses like BDEHR's and potentially affecting their measures of unwarranted disparity.

We study the impact of racial misclassification on unwarranted disparity measures by linking the BDEHR administrative data to public school records, which contain a measure of self-reported race. Relative to this measure, we find that CPS investigators regularly misclassify the race of investigated children: around 9% of self-reported white children and 11% of self-reported Black children have a different race in the CPS administrative data. The majority of such misclassified children are coded as multiracial or as "other" race in the CPS data; interestingly, these misclassified children appear to be less at risk of future maltreatment. To estimate the impact of this misclassification on conventional UD measures, we then develop an extension of the BDEHR methodology that again leverages quasi-random investigator assignment. We find that racial misclassification obscures around 24% of unwarranted racial disparity in foster care placement decisions. Thus, while the majority of UD is revealed by investigator-coded race, conventional discrimination analyses that condition on this measure meaningfully *understate* the true level of discrimination in investigator decisions.

Our second analysis considers potential policy responses to the finding of unwarranted racial disparity, and how these responses might be shaped by systemic biases in multi-stage decisions. Specifically, we study the scope for mitigating discrimination in foster care placement decisions by incorporating algorithmic guidance at the investigation stage—both with and without

an awareness of previous discrimination in CPS call screening. Algorithmic predictions of underlying risk are increasingly found in CPS settings, often with the explicit goal of reducing racial disparities (Samant et al., 2021). Here, building on Arnold, Dobbie, and Hull (2021), we extend the quasi-experimental BDEHR approach to design algorithmic recommendations that eliminate UD at the investigation phase and compare this policy to algorithmic recommendations that eliminate UD in the multi-phase CPS system overall.

This analysis shows that unwarranted racial disparity at the CPS investigation phase can be eliminated by an algorithmic policy that reduces the predicted risk threshold for foster care placement among white children by around 4%. However this policy leaves significant UD in the CPS system as a whole because of the unwarranted disparity in initial call screening. An alternative policy that lowers the white risk threshold by an additional 2% addresses this initial discrimination and eliminates unwarranted disparity among all incoming calls. Thus, as in our first analysis, we find that policy responses to unwarranted racial disparity are meaningfully affected by factors that conventional analyses typically hold fixed.

This paper connects to three distinct literatures. First, we add to a literature studying ways in which racial misclassification or miscoding may affect studies of racial inequity (Conrick et al., 2023; Finlay, Luh, and Mueller-Smith, 2024; Luh, 2022; Rose, 2023; Smith et al., 2010).[2] We study this question by combining a measure of self-reported race with administrative data on decision-maker-coded race, and by leveraging the quasi-random assignment of decision-makers to address OVB. Importantly, this analysis is not premised on the idea that self-reported race is the "ground truth" or even that race is a static or objective concept (Agadjanian, 2022; DeFina and Hannon, 2016; Telles and Paschel, 2014; Roth, 2012). Instead, our approach shows how researchers can study the effect of different racial classifications on measures of unwarranted disparity that take one or more racial classifications as given.

Second, we add to a literature studying how predictive algorithms might reduce inequity in high-stakes decision-making (Agan et al., 2023; Arnold, Dobbie, and Hull, 2021; Grimon and Mills, 2022; Kleinberg et al., 2018b; Rittenhouse, Putnam-Hornstein, and Vaithianathan, 2022). We focus on the question of designing such algorithms in multi-phase systems like CPS, where an appropriate algorithmic policy response might need to account for discrimination at both phases. Our analysis is most closely related to the proposal in Rambachan et al. (2021), by setting race-specific thresholds on algorithmic predictions of maltreatment potential to eliminate unwarranted disparity.

Finally, as noted above, we add to a recent literature using quasi-experimental variation to

[2] As Luh (2022) notes, these studies relate to a broader literature examining impacts of evaluations on the strategic behavior of evaluated agents (Dee et al., 2019; Dinerstein and Opper, 2022; Jacob and Levitt, 2003).

estimate various notions of bias and discrimination in high-stakes decisions, such as pretrial release (e.g., Arnold, Dobbie, and Yang (2018), Hull (2021), Arnold, Dobbie, and Hull (2022), Rambachan (2022), and Canay, Mogstad, and Mountjoy (2022)), traffic stops (e.g., Goncalves and Mello (2021) and Feigenberg and Miller (2022)) and lending (e.g., Dobbie et al. (2021)). The quasi-experimental tools we develop may also prove useful in these and other settings, where race is often coded by decision-makers and decisions are often made in multiple phases.

The remainder of this paper is organized as follows. Section 2 summarizes the CPS context and the earlier methods and findings of BDEHR. Section 3 discusses racial classification by CPS investigators, documents misclassification relative to self-reported race, and studies the impact of such misclassification on measures of unwarranted disparity. Section 4 explores algorithmic counterfactuals designed to eliminate unwarranted disparity both at the investigation phase and in the CPS system overall. Section 5 concludes.

2 Background

This section first outlines the Michigan CPS setting, the context of this study and BDEHR. We then review the methods that BDHER use to estimate unwarranted racial disparity, as well as their main findings. This review sets the stage for our extended analysis of the role of racial misclassification and algorithmic policy responses.

2.1 The CPS Setting

Our analysis uses administrative data from the Michigan CPS system. The CPS process in Michigan, as in most states, typically begins when someone contacts the state's centralized child maltreatment hotline with an allegation of child abuse or neglect. While reports can be submitted by anyone, the most common sources are mandated reporters such as teachers, doctors, and police officers (Benson, Fitzpatrick, and Bondurant, 2022). Incoming calls enter a queue, with the hotline system routing each call to the available screener who has been waiting the longest since their last call. After approximately a 15-minute call, the screener determines whether to "screen-in" the call, triggering a formal CPS investigation, or to "screen-out" the call which officially concludes CPS involvement.

Screened-in cases are assigned to a regional office based on the location of the alleged child victim. Cases are promptly assigned to investigators through a rotational system: incoming cases are assigned to the investigator at the top of the queue, and that investigator then moves to the end of the queue. These investigators examine the allegations and provide a recommendation regarding whether the child should be removed from their home and placed

in foster care. Investigations can take up to 30 days and children who are removed from home spend around 17 months in foster care on average.[3]

Two features of the CPS setting are worth highlighting for our analysis. First, CPS guidelines set a clear mandate for the high-stakes decision of removing a child from home. Investigators are instructed to place a child in foster care only if the child is in "imminent risk" of maltreatment in the home and to otherwise keep the child with their family.[4] This mandate yields a natural measure of unwarranted racial disparity: disparities in foster care placement rates among white and Black children with the same potential future at-home maltreatment. In practice, we follow BDEHR and much of the child welfare literature (e.g. Antle et al., 2009; Putnam-Hornstein, Prindle, and Hammond, 2021) by proxying for future maltreatment potential with a child's potential for CPS re-investigation within six months of the focal investigation.[5]

Second, call screeners and investigators are both quasi-randomly assigned in the Michigan CPS system. This assignment in turn yields quasi-experimental variation in screen-in rates and placement decisions which can be used to estimate disparate impact, as discussed next.[6]

2.2 BDEHR Methods

BDEHR estimate unwarranted racial disparities as differences in foster care placement rates by race, conditional on subsequent at-home maltreatment potential. This measure aligns with economic notions of discrimination, the legal theory of disparate impact, and notions of algorithmic discrimination from the computer science literature (Arnold, Dobbie, and Hull, 2022). Importantly, unwarranted disparity can arise from both "direct" discrimination on the basis of race itself (including canonical sources of racial bias and statistical discrimination) as well as "indirect" discrimination through other characteristics that are correlated with race.

To formalize these UD measures, consider a population of cases indexed by i which involve

[3]Regardless of the placement decision, investigators can formally open a CPS case and recommend "targeted services" to support the family in cases that have been substantiated. These services are typically preventative referrals and range from substance abuse to parenting classes, though parents are not typically compelled to use them. Anecdotally, takeup of such service recommendations is very low; BDEHR show that there is correspondingly little scope for these services causing exclusion restriction violations in their analysis.

[4]For example, the Michigan CPS *Children's Protective Services Policy Manual* instructs investigators to initiate foster care placement "in situations where the child is unsafe, or when there is resistance to, or failure to benefit from, CPS intervention and that resistance/failure is causing an imminent risk of harm to the child" (p.5). Similar instructions are given to CPS call screeners.

[5]See Section IV.A of BDEHR for a discussion of this proxy for future maltreatment potential. They also show that their primary findings are robust to a wide range of alternative proxies, including re-investigation over different horizons and other forms of subsequent contact with CPS.

[6]See Section II of BDEHR for a discussion of screener and investigator assignment processes. They also show assignment is empirically balanced on a wide range of observable characteristics.

either a white or Black child, $R_i \in \{b, w\}$. Each child has a potential for future maltreatment $Y_i^* \in \{0, 1\}$, with $Y_i^* = 1$ indicating that the child would experience subsequent maltreatment when not removed from home.

Each case is initially handled by a hotline call screener who decides whether or not to screen-in the case and advance it to investigation. Conditional on the call being screened-in, a CPS investigator then decides whether to place the child into foster care. For children who are either screened-out or screened-in but not removed from home, maltreatment potential is realized and observable via their subsequent maltreatment outcomes. Otherwise, Y_i^* is not realized and hence unobservable.

To develop the UD measures, first suppose that cases are handled by a single representative hotline screener and, if screened-in, a single representative investigator. Let $S_i \in \{0, 1\}$ indicate whether case i is screened-in by the screener and (if screened-in) $D_i \in \{0, 1\}$ indicate whether the investigator places the child in foster care. The product, $P_i = S_i D_i$, then indicates whether case i ultimately results in foster care placement.

We first consider a disparity in the screener's decision to screen-in a call or not among Black and white children without future maltreatment potential:

$$\Delta_0^S = E[S_i \mid R_i = b, Y_i^* = 0] - E[S_i \mid R_i = w, Y_i^* = 0], \tag{1}$$

and a corresponding disparity among children with future maltreatment potential:

$$\Delta_1^S = E[S_i \mid R_i = b, Y_i^* = 1] - E[S_i \mid R_i = w, Y_i^* = 1]. \tag{2}$$

A measure of overall screener UD is given by averaging these conditional disparities:

$$\Delta^S = \Delta_0^S(1 - \bar{\mu}) + \Delta_1^S \bar{\mu}, \tag{3}$$

with weights $\bar{\mu} = E[Y_i^*]$, which measures the average future maltreatment risk in the population of all calls. Here, Δ^S measures the expected level of UD when the screener encounters a representative pool of cases with unknown future maltreatment potential.

We next define UDs in the investigator's decision to place screened-in children in foster care:

$$\Delta_0^D = E[D_i \mid R_i = b, Y_i^* = 0, S_i = 1] - E[D_i \mid R_i = w, Y_i^* = 0, S_i = 1] \tag{4}$$

$$\Delta_1^D = E[D_i \mid R_i = b, Y_i^* = 1, S_i = 1] - E[D_i \mid R_i = w, Y_i^* = 1, S_i = 1]. \tag{5}$$

Here, Δ_0^D measures the investigator's placement rate disparity among screened-in cases

without future maltreatment potential and Δ_1^D measures the corresponding disparity among screened-in cases with future maltreatment potential. We again measure the overall UD in the investigator's decisions by averaging these two disparities: $\Delta^D = \Delta_0^D(1-\bar{\mu}^{S=1}) + \Delta_1^D\bar{\mu}^{S=1}$, where the $\bar{\mu}^{S=1} = E[Y_i^* \mid S_i = 1]$ weights now represent average maltreatment risk among the subpopulation of screened-in cases.

Finally, we define unwarranted disparities in placement rates among all cases referred to CPS:

$$\Delta_0^P = E[P_i \mid R_i = b, Y_i^* = 0] - E[P_i \mid R_i = w, Y_i^* = 0] \tag{6}$$

$$\Delta_1^P = E[P_i \mid R_i = b, Y_i^* = 1] - E[P_i \mid R_i = w, Y_i^* = 1], \tag{7}$$

with overall placement UD given by the average: $\Delta^P = \Delta_0^P(1-\bar{\mu}) + \Delta_1^P\bar{\mu}$. Intuitively, these measures capture unwarranted disparity from both screener and investigator decisions.

A decomposition can be used to quantify the relative importance of screener and investigator UDs in shaping eventual UDs in placement rates. Specifically, since $P_i = S_i D_i$, we have:

$$\begin{aligned} \underbrace{\Delta_y^P}_{\text{Placement UD}} &= E[S_i \mid R_i = b, Y_i^* = y]E[D_i \mid S_i = 1, R_i = b, Y_i^* = y] \\ &\quad - E[S_i \mid R_i = w, Y_i^* = y]E[D_i \mid S_i = 1, R_i = w, Y_i^* = y] \\ &= \underbrace{\Delta_y^S\tilde{\omega}_y^S}_{\text{Screener component}} + \underbrace{\Delta_y^D\tilde{\omega}_y^D}_{\text{Investigator component}}, \end{aligned} \tag{8}$$

for $y \in \{0, 1\}$, where $\tilde{\omega}_y^S = E[D_i \mid R_i = b, Y_i^* = y, S_i = 1]$ and $\tilde{\omega}_y^D = E[S_i \mid R_i = w, Y_i^* = y]$.

Equation (8) decomposes eventual placement UD into two components: one involving UD in screener decisions to screen-in a call (Δ_y^S) and the other involving UD in investigator decisions to place screened-in children in foster care (Δ_y^D). These UDs are weighted by $\tilde{\omega}_y^S$—the placement rate of screened-in Black children with $Y_i^* = y$—and $\tilde{\omega}_y^D$—the screened-in rate of white children with $Y_i^* = y$—respectively.[7]

The fundamental challenge in bringing each of these UD measures to data is the fact that maltreatment potential is only selectively observed. Among children who are not placed into foster care, their maltreatment potential is directly revealed by their subsequent maltreatment outcomes. However, since future maltreatment in the home is unobserved among those who are placed in foster care, one cannot directly condition on it to estimate Equations (1)-(8).

[7]This equation is derived similarly to a Kitagawa–Oaxaca-Blinder (KOB) decomposition, by adding and subtracting $E[D_i \mid R_i = b, Y_i^* = y, S_i = 1]E[S_i = 1 \mid R_i = w, Y_i^* = y]$ to the first line of (8) and rearranging terms. As is the case with KOB decompositions, an alternate version of Equation (8) can be obtained by changing the "order" of the decomposition.

BDEHR show how this identification challenge can be overcome with quasi-experimental variation. Namely, they first show how the different UD measures and the decomposition can be written in terms of directly estimable moments and a set of key mean risk parameters which are obscured by the selective observability of Y_i^*. They then show how these key parameters (and therefore each of the Equations (1)-(8)) can be estimated by leveraging the assignment of cases to screeners and investigators, using an "identification at infinity" approach that builds on Arnold, Dobbie, and Hull (2021, 2022) (see Appendix A.1 for details). This strategy relies on the as-good-as-random assignment of screeners and investigations as well as exclusion restrictions which BDEHR test extensively in the Michigan CPS setting. The strategy benefits from the fact that foster care placement rates are low for both Black and white children, with some quasi-randomly assigned screeners and investigators revealing nearly the full distribution of Y_i^* by leaving nearly all children in the home.

2.3 BDEHR Findings

BDEHR present three main findings from this approach. First, they estimate significant UD in the decisions of CPS screeners and investigators. On average, hotline calls involving Black children are found to be screened-in at a 5 percentage point (8%) higher rate than calls involving white children, conditional on subsequent at-home maltreatment potential. Specifically, BDEHR estimate an average Δ^S across screeners of 0.050 (with a standard error of 0.001). Subsequent investigators inherit these initial disparities and amplify them; BDEHR find that investigators place screened-in Black children in foster care at a 1.7 percentage points (50%) higher rate than screened-in white children with identical maltreatment potential. That is, they estimate an average Δ^D across investigators of 0.017 (with a standard error of 0.002).

Second, BDEHR use the Equation (8) decomposition to link screener and investigator UDs and estimate significant overall UD in foster care placement decisions. Specifically, they find that hotline calls involving Black children are around 1.1 percentage points (55%) more likely to end up in foster care relative to calls involving white children with identical maltreatment potential (i.e., they estimate an average Δ^P of 0.011, with a standard error of 0.001). Screener decisions account for between 13% and 19% of this overall placement UD, with investigator decisions driving the remainder.

Third, BDEHR show that the placement disparity is concentrated among children with subsequent maltreatment potential (i.e., driven by Δ_1^P rather than Δ_0^P). This finding adds nuance to ongoing discussions surrounding the potential "over-placement" of Black children in foster care, as it suggests that a higher placement rate may offer relative protection to Black children. This is particularly the case in light of recent work in our context, which finds that

foster care placement leads to improvements in safety and other welfare-relevant outcomes for children at the margin of placement (Baron and Gross, 2022; Gross and Baron, 2022).

Taken together, the BDEHR analysis shows that there is significant unwarranted racial disparity in Michigan CPS foster care decisions, and that this UD comes primarily from CPS investigators decisions among cases with future maltreatment potential. We correspondingly focus our analysis on this subpopulation.

3 Race Measurement and Misclassification

This section first details the measurement of race and ethnicity in the administrative data used in BDEHR. We then examine potential racial misclassification in these records by linking the CPS data to public education records, where race is self-reported. Finally, we develop and apply an empirical framework for quantifying the role of racial misclassification in estimates of investigator UDs.

3.1 Race Measurement in CPS Data

The primary dataset in BDEHR comes from the Michigan Department of Health and Human Services (MDHHS). This dataset spans 2008 to 2019 and contains information on CPS hotline calls and subsequent investigations in Michigan. Overall, the data contain the details of each call, including the nature of the allegation (e.g., physical abuse versus neglect), the relationship of the alleged perpetrator to the child (e.g., a parent or uncle), the age and gender of the child, and the identity of the screener and (if screened-in) investigator assigned to the case.

The MDHHS dataset also contains the child's race both for screened-out and screened-in calls. While hotline screeners in Michigan do not directly request callers to provide information about the alleged victim's race, there are two primary methods by which hotline screeners assess the race of children. First, for children with prior interactions with CPS, the race recorded previously will be available to the call screener. For children without any prior CPS involvement, screeners can observe race through a state-wide database called MIBridges. This database contains detailed information—including self-reported race—of families who have received various state benefits. These benefits encompass programs such as Medicaid coverage, food assistance, cash assistance (including the Family Independence Program, Refugee Cash Assistance, and TANF), child development and care, and state emergency relief. If race information for the child is missing in the CPS system, the system will populate the race field with race information from MIBridges. Since the vast majority of children reported to the hotline have prior interactions with either CPS or MIBridges, hotline screeners can determine

the child's race in most cases.[8]

Among screeened-in calls, the coding of a child's race by screeners may be updated by investigators at any point in the investigation process. These updates, and any disagreements in the investigators' coding vs. the administrative MIBridges coding, are not observable in the MDHHS data, which reports a single racial code for each child.

A potential concern is that investigators could systematically reclassify white or Black children in a way that affects a researcher's estimates of unwarranted racial disparity in placement decisions. Such non-random reclassification might reflect that the beliefs of the investigator differ from self-reported race, or could be intentional—for example, if investigators deliberately code Black children in risky home environments as multiracial or "other" race in order to hide their discriminatory behavior. Evidence of such intentional misclassification has previously been documented in the criminal justice system (Luh, 2022; Finlay, Luh, and Mueller-Smith, 2024). Of course, misclassification might also be idiosyncratic or otherwise inconsequential for UD calculations. We next investigate this issue with additional data.

3.2 Racial Misclassification

Descriptive Statistics

To study the extent and impact of racial misclassification, we use an administrative link between children appearing in the CPS data from 2008 to 2017 and public school records from the state. These records include a variable for the child's race as self-reported by their family. Specifically, whenever a student first enrolls in a school, their parents (or custodial caregivers) complete a form from the school district which includes a question about the child's race and ethnicity. We use this question to identify children with a self-reported race of either white or Black.[9]

We focus this analysis on screened-in children, for two reasons. First, this allows us to compare self-reported race with race classified by investigators in the CPS system; as noted above, our measure of race for screened-out calls often comes from state databases where race was self-reported. Second, as noted in Section 2.3, BDEHR find that most of the unwarranted

[8]Race information is missing in fewer than 10% of all calls. The state believes that the majority of these instances involve Native American children, for whom the screener does not input further race details once their Native American status is established due to concerns related to jurisdiction.

[9]In contrast to our CPS data, which contains a single value for the child's race (e.g., only white, only Black, or multi-racial), the self-reported race in the public education data encompasses all the values identified by parents. For instance, if a parent records that a child is both white and Black, both variables are retained separately rather than summarizing the child's race as multiracial. To ensure comparability across the two datasets, we classify a child in the education data as white or Black only if the child's family specifically identifies them as white or Black without listing any other race.

racial disparity in foster care placement arises at the investigation stage.[10] We thus construct a dataset of Michigan CPS investigations from 2008 to 2017 following the main sample restrictions in BDEHR.[11] Using a crosswalk created for the analysis in Gross and Baron (2022) and Baron and Gross (2022), we are able to match around 96% of this sample to a public school record and thus to a measure of self-reported race.[12]

The resulting analysis sample consists of 220,832 investigations involving 178,372 children. Column 1 in Table 1 summarizes the sample. Female children make up 48% of the sample and the average child is roughly 7.2 years old. About 48% children in the sample experienced a formal CPS investigation prior to the focal one. Roughly 29% of investigations in the sample include a physical abuse allegation, and in 91% of investigations the alleged perpetrator is the parent or stepparent. 2.8% of children in the sample are placed in foster care and 16.4% of children experience another investigation for maltreatment in the home within six months if not placed in foster care.

Figure 1 illustrates the distributions of self-reported race from the education data and investigator-coded race from the CPS data. Overall, the proportion of white and Black children (reported inside the bars) is very similar across the two classifications. In both classifications, 59% of children are identified as white and 23% of children are categorized as Black. However, differences emerge in the datasets concerning the proportion of children coded as multiracial. In the self-reported classification, 13% of children are multiracial compared to 8.4% according to the CPS classification. The remaining children with "other" race consist of those whose race is categorized as Hispanic, Asian, American Indian, or have a missing race.

While most white and Black children are consistently classified across both datasets, many self-reported white and Black children are misclassified by CPS investigators. Specifically, 9% of the children whose parents self-identify them as white in the education data are not classified as white in the CPS data. Similarly, 11% of those self-identified as Black in the

[10] A related concern regarding the analysis in BDEHR is the potential for selection by race into initial allegations of maltreatment. This concern is especially relevant in light of evidence from other fields indicating that Black children may be more frequently reported compared to their white counterparts in similar situations (Lane et al., 2002). Addressing this issue is difficult, since it requires observing children who could have been reported to the hotline but were not. To explore this issue, Table 1 in BDEHR shows that the distribution of mandated and non-mandated reporters is similar across racial groups. Furthermore, their Table A16 shows that estimates of investigator unwarranted disparity are similar across different types of reporters.

[11] As in BDEHR, we drop (i) cases assigned to investigators with fewer than 200 investigations in our sample; (ii) children for whom we cannot observe outcomes for at least six months; (iii) repeat cases and cases involving sexual abuse; and (iv) cases with a missing child zip code.

[12] Because the datasets do not contain a common numeric identifier, the records were linked using a probabilistic algorithm based on the child's first name, last name, date of birth, and gender. As explained in Gross and Baron (2022), the match rate is not expected to be 100% since some investigated children may be homeschooled or may move out of the state after an investigation. See Online Appendix D in Gross and Baron (2022) for additional details regarding the match.

Table 1: Racial Misclassification Analysis: Summary Statistics

		Well-Classified by CPS		Misclassified by CPS	
	All Children	**White**	**Black**	**White**	**Black**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
		Panel A: Child characteristics			
Female	0.480	0.479	0.480	0.471	0.470
Age at investigation	7.230	7.371	6.953	8.248	7.673
Had a previous investigation	0.480	0.494	0.481	0.365	0.424
No. of previous investigations	1.115	1.171	1.030	0.971	1.059
		Panel B: Investigation characteristics			
Physical abuse allegation	0.287	0.280	0.302	0.293	0.296
Alleged parent perpetrator	0.913	0.913	0.908	0.927	0.912
Alleged other relative perpetrator	0.051	0.045	0.061	0.057	0.063
		Panel C: Treatment rates			
Foster care placement	0.028	0.025	0.037	0.026	0.032
		Panel D: Maltreatment outcome, if not placed			
Re-investigated within six months	0.164	0.175	0.143	0.144	0.138
Observations	220,832	118,583	46,382	9,993	4,631

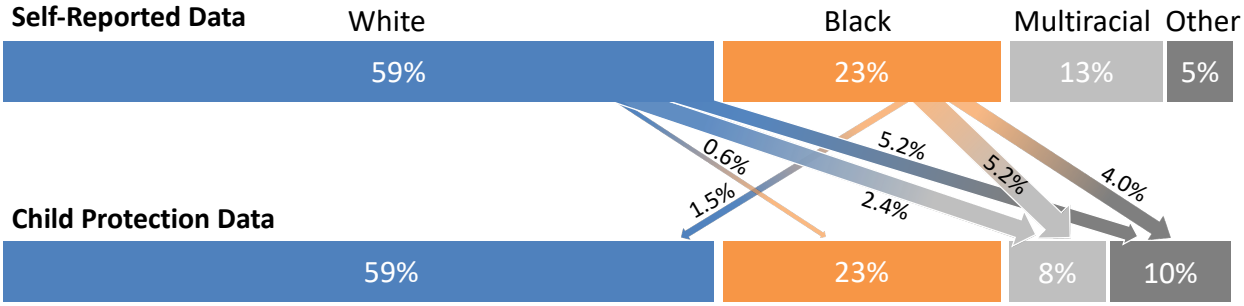
Notes: Column 1 of this table summarizes the sample for the racial misclassification analysis, consisting of children investigated by Michigan CPS in 2008-2017 who can be matched to a Michigan public school record. Columns 2-5 summarize children who are either self-identified as white or Black in the education data. Columns 2-3 summarize children whose race in the education data is correctly classified in the CPS data, while columns 4-5 summarize children whose race in the education data is misclassified in the CPS data.

education data are misclassified in the CPS data. Misclassification rates are similar in the other direction: 9% and 10% of those who are classified as white and Black, respectively, in the CPS data, are classified as another race in the education data.

The arrows in Figure 1 report the shares of self-reported white or Black children who are classified as another race in the CPS data. These show that most of the misclassification by CPS investigators occurs from children being coded as multiracial or the "other" category. Only 0.6% of self-reported white children and 1.5% of self-reported Black children are misclassified as Black and white respectively. In contrast, the share of self-reported white (Black) children coded as multiracial by CPS investigators is 2.4% (5.2%) and the corresponding share for the other category is 5.2% (4.0%).

Columns 2–5 of Table 1 compare the observable characteristics of self-reported white and Black children who are either well-classified by CPS—in the sense of also being coded as white or Black, respectively—or who are misclassified with a different racial coding. Misclassified children tend to be older and have less prior experience with the CPS system, as captured by a lower number of previous investigations. While investigation characteristics are broadly similar between the two groups, foster care placement rates differ, with misclassified Black children being 0.5 percentage points less likely to be removed from home than well-classified

Figure 1: Racial Misclassification in CPS Data



Notes: This figure shows the distribution of child race in our analysis sample according to the self-reported racial classification from the education data (in the top row) and the racial classification by investigators from the CPS data (in the bottom row). The "other" categories consist of children whose race is categorized as Hispanic, Asian, American Indian, or missing. The percentages indicate the share of self-reported white or Black children who are categorized differently by an investigator. Of individuals who self-report as white, 91% are categorized correctly. Of individuals who self-report as Black, 89% are categorized correctly.

Black children. Moreover, rates of re-investigation within six months among the children not removed from home are significantly lower for misclassified children: by 3.1 percentage points for misclassified white children and by 0.5 percentage points for misclassified Black children.

Taken together, these results suggest that CPS investigators regularly and systematically misclassify children relative to their self-reported race. The patterns in Table 1 are, however, at odds with the simple story of investigators deliberately misclassifying risky Black children as another race to hide discriminatory behavior—if anything, the misclassified Black children appear less risky. To understand the potential for systematic misclassification to affect discrimination measures, we next use these data to extend the BDEHR analysis.

Quantifying the Impact of Racial Misclassification

Building on the framework in Section 2.2, we next introduce a decomposition of UD by self-reported race into a component that is revealed by the racial coding of CPS investigators and a component that is obscured by racial misclassification in CPS records. As with the main decomposition in BDEHR, this decomposition arises as a special case of the general framework in Bohren, Hull, and Imas (2022)—developed for disentangling discrimination from direct and systemic factors—and can be applied by leveraging quasi-random investigator assignment.

We formalize this approach by letting $R_i^* \in \{b, w\}$ be the self-reported race of child i (either Black or white) and again letting R_i be race recorded by a CPS investigator. Relative to the Section 2.2 framework, we allow racial misclassification in CPS: $R_i \neq R_i^*$. Specifically, we let $R_i \in \{b, w, m\}$ where $R_i = m$ denotes that child i (who is either self-reported Black or white) is coded as multiracial or as "other" race in the CPS data. Since the share of children

misclassified as being Black or white is very small, we focus on these forms of misclassification.

Consider a population of screened-in cases, dropping the $S_i = 1$ conditioning from Section 2.2 only for notational ease. Following the BDEHR findings, we consider a measure of unwarranted disparity among children with the potential for future at-home maltreatment ($Y_i^* = 1$), now explicitly considering children who self-identify as Black or white:

$$\Delta_1^{D*} = E[D_i \mid R_i^* = b, Y_i^* = 1] - E[D_i \mid R_i^* = w, Y_i^* = 1]. \tag{9}$$

A researcher who only has access to the investigator-recorded race R_i may estimate a biased measure of such unwarranted disparity. In BDEHR, for example, unwarranted disparity is measured by comparing children with $R_i = b$ vs. $R_i = w$, discarding those coded as multiracial or another race ($R_i = m$). How does such miscoding affect estimates of Δ_1^{D*}?

Appendix A.2 derives a general decomposition of $\Delta_1^{D^*}$ into two terms. The first term arises from unwarranted disparities by self-reported race among children coded as the same race r by investigators, $E[D_i \mid R_i^* = b, R_i = r, Y_i^* = 1] - E[D_i \mid R_i^* = w, R_i = r, Y_i^* = 1]$. The second term arises from unwarranted disparities in investigator race-coding itself, $Pr(R_i = r \mid R_i^* = b, Y_i^* = 1) - Pr(R_i = r \mid R_i^* = w, Y_i^* = 1)$.[13] The appendix further shows that when—as we find in Figure 1—racial miscodings almost always occurs through multiracial or other-race codings (i.e. when $Pr(R_i = b \mid R_i^* = w) \approx 0$ and $Pr(R_i = w \mid R_i^* = b) \approx 0$), this decomposition takes an intuitive approximate form:

$$\begin{aligned}\Delta_1^{D*} \approx & \underbrace{E[D_i \mid R_i^* = b, R_i = b, Y_i^* = 1] - E[D_i \mid R_i^* = w, R_i = w, Y_i^* = 1]}_{\text{Revealed discrimination}} \omega_1^{Revealed} \\ & + \underbrace{E[D_i \mid R_i^* = b, R_i = m, Y_i^* = 1] - E[D_i \mid R_i^* = w, R_i = m, Y_i^* = 1]}_{\text{Obscured discrimination}} \omega_1^{Obscured}\end{aligned} \tag{10}$$

for some non-negative weights $\omega_1^{Revealed}$ and $\omega_1^{Obscured}$.

Equation (10) shows that unwarranted disparity by self-reported race R_i^* can be approximated as the sum of two terms. The first *revealed discrimination* term captures unwarranted disparities among children whose coded race R_i matches self-reported race. Such disparities would contribute to the discrimination measures in studies like BDEHR's, which only have access to R_i. The second *obscured discrimination* term, in contrast, would not contribute to these studies: it captures unwarranted disparities among self-reported Black vs. white children who are recorded as multiracial or other race by investigators and hence are discarded in

[13] The first term captures what Bohren, Hull, and Imas (2022) call "average direct discrimination," where R_i is taken as a signal of self-reported race R_i^*. The second term captures what they call indirect or "systemic" discrimination, arising from unwarranted disparities in the signal distribution.

BDEHR. Intuitively, Appendix A.2 shows these disparities are respectively weighted by shares of children who are well- and misclassified by CPS investigators. The appendix also shows the decomposition involves a residual term which is small in settings where misclassification rates are similar by self-reported race (as we find in Figure 1).

Like the main decomposition in BDEHR, each of the terms in Equation (10) can be estimated using quasi-random investigator assignment. Appendix A.2 formalizes this extension, which relies on the same identifying assumptions that BDEHR test extensively in this setting.

Figure 2 shows estimates of the Equation (10) decomposition in Michigan CPS. Total unwarranted racial disparity among self-identified Black and white children with future maltreatment potential, Δ_1^{D*}, is estimated at 8.3 percentage points (SE: 0.037). The decomposition shows that around 75% of this UD (6.2 percentage points; SE: 0.036), is revealed by the racial coding of CPS investigators while around 24% (2.0 percentage points; SE: 0.008) is obscured by racial misclassification. The remaining 1% is due to the decomposition residual.

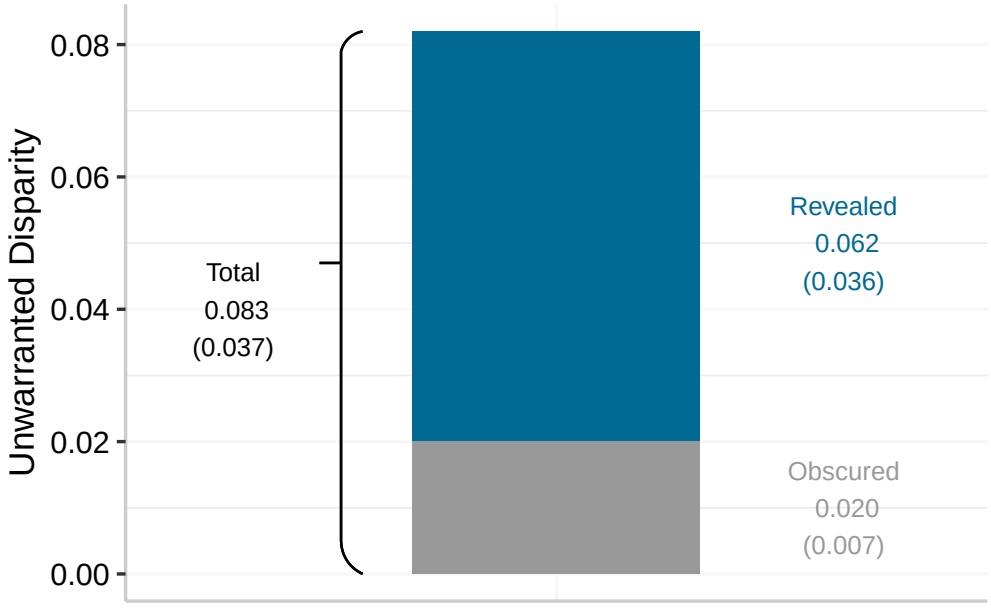
The decomposition thus shows that the majority of unwarranted racial disparity, as defined by self-reported race, is revealed by investigator-coded race. Nevertheless, a meaningful share of discrimination is obscured by CPS misclassification relative to self-reported race. Among children classified by CPS as multiracial or another race, there continues to be an unwarranted disparity across self-reported Black versus white children. Conventional discrimination analyses like BDEHR's, which condition on the endogenous CPS classification, therefore stand to understate the full extent of racial discrimination in investigator decisions.

4 Algorithmic Policy Responses

This section explores potential policy responses to the finding of unwarranted racial disparity in foster care placement decisions, and how such policy responses might be shaped by systemic biases in the multi-phase CPS system. We focus on the scope for mitigating discrimination in foster care placement by incorporating algorithmic guidance. At least 26 states and the District of Columbia have considered using predictive analytic tools, which predict a child's potential for maltreatment from family and case characteristics, to guide CPS involvement (Samant et al., 2021).

Algorithmic tools in CPS have been piloted at both the screening phase (e.g., the Allegheny Family Screening Tool (Rittenhouse, 2022)) and the investigation phase (e.g., the Los Angeles County Risk Stratified Supervision Model). Given the findings in BDEHR that the majority of unwarranted disparity in foster care placement is driven by the investigation phase (around

Figure 2: Decomposition of Unwarranted Disparity in Investigators' Decisions (Δ_1^D)



Notes: This figure shows estimates of investigator UD by self-reported race among screened-in children with maltreatment potential (Δ_1^{D*}), as well as estimates of the two terms of decomposition (10). All estimates are based on a linear extrapolation of investigator-specific placement and subsequent maltreatment rates among children left at home, as described in Appendix A.2. Standard errors are two-way clustered by child and investigator and reported in parentheses.

81% to 87%), we begin by demonstrating how a researcher or policymaker could use guidance from a race-conscious algorithm to eliminate UD during the investigation phase. We then demonstrate how this same algorithm can be used to eliminate UD in eventual foster care placement rates by offsetting disparities from initial human screening. As before, we focus on unwarranted disparity among cases where future maltreatment potential is present (i.e. Δ_1^D and Δ_1^P) which BDEHR found to drive racial discrimination.

Using Algorithmic Guidance to Eliminate Δ^D

We first consider policy counterfactuals which eliminate investigator UD (Δ_1^D) when placement decisions are made by a race-conscious algorithm. In other words, we study how discrimination in algorithmic placement recommendations among screened-in children compares to existing human investigator UDs, and whether incorporating race-specific predictions and thresholds for placement can reduce or even eliminate UD at the investigation stage. For this analysis we take as fixed decisions occurring prior to the investigation—specifically, we condition on the prior screening-in decision.

We use a machine learning algorithm to predict future maltreatment potential Y_i^* as a function of case, child, and family characteristics observable to both the investigator and the econometrician, X_i. Importantly, we fit the algorithm separately for white and Black children. Denote this race-specific prediction by $m(X_i, R_i)$. Following Kleinberg et al. (2018a), we fit $m(X_i, r)$ with a gradient-boosted decision tree in the sample of screened-in children of race $R_i = r$ who were not placed in foster care, with Y_i^* indicating that the child has a subsequent maltreatment investigation in the home within six months of the focal investigation (the main maltreatment potential proxy in BDEHR). Features in X_i include information on the reason for the CPS report (physical abuse, medical neglect, physical neglect, domestic violence, substance abuse, or improper supervision), whether the alleged perpetrator was the child's mother or father, the child's history of CPS contact (an indicator for previous contact and the number of prior investigations), the gender and age of the child, and their residing county. We tune the algorithm to select optimal parameters in the gradient boosted decision tree using a 5-fold cross-validation technique.

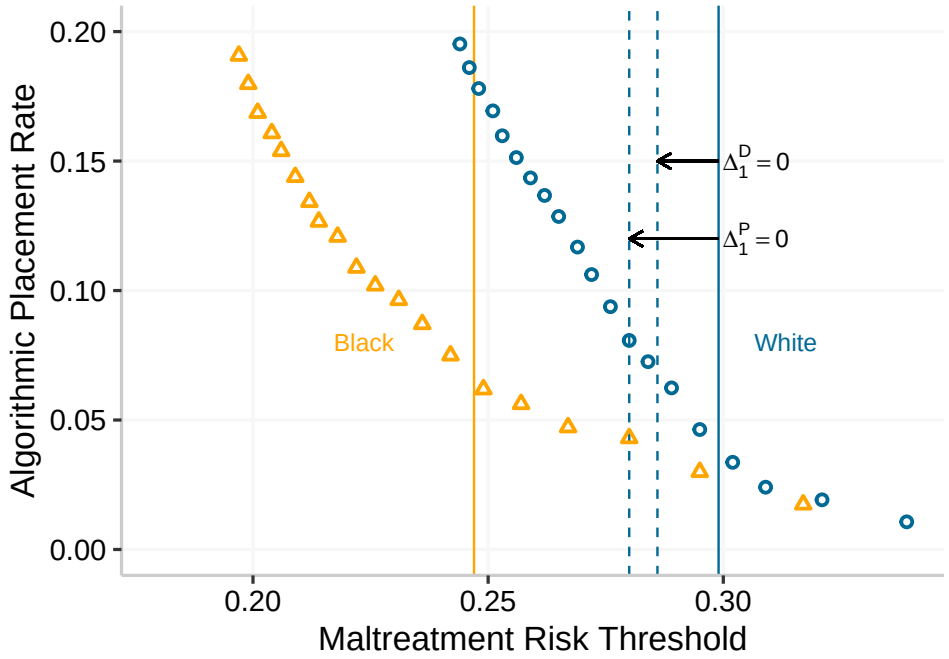
We then form foster care placement recommendations in the full evaluation dataset as $T_i = \mathbb{1}[m(X_i, R_i) > \tau(R_i)]$, where $\tau(r)$ is a risk threshold that is also allowed to depend on the child's race r. Our goal is to estimate unwarranted disparity in these algorithmic recommendations among screened-in children with maltreatment potential:

$$\Delta_1^T = E[T_i \mid S_i = 1, R_i = b, Y_i^* = 1] - E[T_i \mid S_i = 1, R_i = w, Y_i^* = 1] \tag{11}$$

Estimating discrimination in algorithmic decisions is hampered by the same fundamental selection challenge that arises when estimating discrimination in investigator decisions: data on a child's potential for subsequent maltreatment is selectively observed for individuals endogenously left at home following an investigation. Such "selective labels problem" (Kleinberg et al., 2018a) complicates the measurement of algorithmic discrimination.

We follow the methods developed in Arnold, Dobbie, and Hull (2021) to overcome this selection challenge and estimate UD in our algorithm's placement recommendations. These methods follow the quasi-experimental approach in Section 2.2. By varying the $\tau(w)$ and $\tau(b)$ thresholds we trace out the possibilities frontier of the race-specific placement rates among children with and without future maltreatment potential, and the corresponding disparity, Δ_1^T. Specifically, we estimate $\delta_{1r} = Pr(T_i = 1 | S_i = 1, Y_i^* = 1, R_i = r)$ for $r \in \{w, b\}$ by combining our baseline estimates of race-specific maltreatment risk $\mu_r^{S=1} = E[Y_i^* \mid S_i = 1, R_i = r]$ with estimates of the conditional second moments $\rho_r = E[Y_i^* T_i \mid S_i = 1, R_i = r]$. The latter estimates come from a local linear extrapolation of the conditional moments $E[Y_i^* T_i \mid S_i = 1, R_i = r, D_{ij} = 1]$, following the logic in Appendix A.1.

Figure 3: Algorithmic Placement Rates at Different Risk Thresholds



Notes. The figure shows estimates of the algorithmic placement rates among screened-in children with maltreatment potential of race r as a function of various race-specific placement thresholds $\tau(r)$. The vertical lines show the thresholds that would match the placement rates by human investigators in our sample. The first dashed line (going from right to left) shows the policy change that reduces $\tau(w)$ from 29.9% to 28.6%, and would equalize placement rates among screened-in white and Black children with maltreatment potential (and thus yield a Δ_1^D equal to zero). The second dashed line shows the policy change that reduces $\tau(w)$ from 29.9% to 28%, and would equalize placement rates among the full population of calls involving white and Black children with maltreatment potential (and thus yield a Δ_1^P equal to zero).

We bring these ideas to investigations from the analysis sample in BDEHR from 2008 to 2017, consistent with the sample period in Section 3 (see BDEHR for a description of this analysis sample). Figure 3 shows estimates of placement rates (δ_r) for each race r that would arise from race-specific thresholds, $\tau(r)$. The vertical lines show the race-specific placement threshold $\tau(r)$ that would match the observed placement rates in our sample: predicted-risk thresholds of 29.9% for white children and 24.2% for Black children. The difference in the heights of the blue and orange curves at the thresholds illustrate estimates of Δ_1^T for status quo thresholds. As we move the thresholds around, we simulate how the measures of algorithmic discrimination would change.

We find that average algorithmic discrimination among children with subsequent maltreatment potential would be around 31% lower than the average human investigator UD: we estimate Δ_1^T to be around 4 percentage points, vs. 5.8 percentage points for Δ_1^D (see Table 3 in BDEHR). Thus, incorporating algorithmic guidance in this setting meaningfully reduces unwarranted racial disparity, even while keeping average placement rates the same.

In line with our main findings in BDEHR that white children may be "under-placed" in high-risk situations relative to Black children, we next simulate whether algorithmic discrimination during the investigation phase could be eliminated by slightly lowering the threshold for white children. Panel A of Figure 3 shows that reducing $\tau(w)$ by 4.3%, from 0.299 (the solid blue vertical line in the figure) to 0.286 (the dashed blue vertical line) would set the placement rates for screened-in white and Black children approximately the same in cases with maltreatment potential ($\delta_{w1} \approx \delta_{b1}$), and therefore $\Delta_1^T \approx 0$.

Using Algorithmic Guidance to Eliminate Δ^P

The exercise above is a natural starting point for a policy response to unwarranted disparity, especially given the findings in BDEHR that most discrimination in eventual foster care placement is driven by the investigation phase. The exercise is also policy-relevant, given current efforts to use algorithmic guidance to reduce racial disparities during the investigation phase. However, in a multi-phase system such as CPS, such an exercise may leave significant unwarranted disparity on the table if discrimination exists at earlier stages in the system. Recall that Δ_1^P measures unwarranted disparities among the full set of calls to CPS, incorporating UD at both the screener and investigation stages. Per Equation (8), an algorithm designed to set $\Delta_1^D = 0$ need not set $\Delta_1^P = 0$ because $\Delta_1^S > 0$. In other words, even if disparities during the investigation stage are eliminated, unwarranted disparity in foster care placement will remain because of disparities introduced during the initial human screening. We next demonstrate how to extend the above framework to explore algorithmic counterfactuals that eliminate total discrimination.

To do so, we define a multi-phase discrimination decomposition for the algorithm-based placement rates, analogous to the Equation (8) decomposition:

$$\Delta_1^P = \Delta_1^S \tilde{\omega}_1^S + \Delta_1^D \tilde{\omega}_1^D \tag{12}$$

where, as before:

$$\Delta_1^S = E[S_i | R_i = b, Y_i^* = 1] - E[S_i | R_i = w, Y_i^* = 1] \tag{13}$$

and

$$\tilde{\omega}_1^D = E[S_i | R_i = w, Y_i^* = 1], \tag{14}$$

but now:

$$\Delta_1^D = E[T_i | R_i = b, Y_i^* = 1, S_i = 1] - E[T_i | R_i = w, Y_i^* = 1, S_i = 1] \tag{15}$$

and

$$\tilde{\omega}_1^S = E[T_i|R_i = b, Y_i^* = 1, S_i = 1], \tag{16}$$

where the algorithmic placement decisions T_i are implicitly tied to given thresholds $\tau(b)$ and $\tau(w)$. A policymaker may now, for example, wish to find the value of $\tau(w)$ that sets Δ_1^P to zero, holding fixed the value of $\tau(b)$ at the solid orange line in Figure 3.

From the findings in BDEHR we obtain estimates of $\Delta_1^S = 0.038$ and $\tilde{\omega}_1^D = 0.595$ (see their Table 3 and Table A8). We can also infer the value of $\tilde{\omega}_1^S = E[T_i|R_i = b, Y_i^* = 1, S_i = 1]$ for $\tau(b)$ such that $E[T_i|R_i = b] = E[D_i|R_i = b]$. This value is given by the intersection of the orange triangles and the solid vertical line in Panel A of Figure 3, and it is equal to 0.07.

From Equation (12), setting $\Delta_1^P = 0$ requires:

$$\Delta_1^D = -\frac{\Delta_1^S \tilde{\omega}_1^S}{\tilde{\omega}_1^D}. \tag{17}$$

Plugging in the above values, this condition is obtained with $E[T_i|R_i = w, Y_i^* = 1, S_i = 1] = 0.075$. A grid search based on the curves plotted in Figure 3 find that setting $\tau(w) = 0.280$ achieves this conditional placement rate for white children. Panel B of Figure 3 illustrates this result: reducing $\tau(w)$ by an additional 2% (from 0.286 to 0.280 – the new dashed vertical blue line in the figure) would set the placement rate for white children with subsequent maltreatment potential at 7.5%, compared to the algorithmic placement rate for Black children of 7%. Intuitively, the placement rate for screened-in white children needs to be larger than the placement rate for screened-in Black children in order to mitigate unwarranted disparities in the initial screening decision of 3.8 percentage points.

These results show how researchers or policymakers can employ algorithmic guidance to mitigate unwarranted disparities in foster care placement, leveraging quasi-experimental variation as needed to address OVB. In multi-phase systems such as CPS, an analysis of discrimination that follows the framework developed in BDEHR can furthermore be combined with algorithmic counterfactuals that counteract disparities introduced by human decision-makers at earlier points in the system. In our setting this exercise shows that policy responses to unwarranted racial disparity are meaningfully affected by the earlier decisions.

We conclude this section by noting that our policy counterfactuals have focused entirely on eliminating unwarranted disparities in foster care placement, without consideration of the full welfare implications of such policies. Given the findings of BDEHR, that disparities are driven by an under-placement of white children in high-risk situations, we focused on counterfactuals that decrease the threshold for placing white children. A policy that places additional white

children with maltreatment potential is likely to be welfare-improving, especially in our context, where the causal effects of foster care for marginal children are positive. However, it is important to recognize that, in practice, such a policy would impact both children with and without maltreatment potential. That is, since one cannot set maltreatment-potential-specific thresholds, such a policy could also result in additional white children without maltreatment potential being placed in foster care. Because the causal effects of foster care for these children are less clear in our context, we are unable to make conclusive welfare statements.

5 Conclusion

This study contributes to a growing literature that seeks to understand how conventional discrimination measures may obscure additional forms of "systemic" discrimination (Bohren, Hull, and Imas, 2022) by conditioning on endogenous or upstream factors. We develop quasi-experimental tools to (i) quantify the impact of racial misclassification on measures of unwarranted disparity, and (ii) design policy responses that consider systemic biases in multi-stage decisions. Applying these methods to the context of foster care placement by CPS investigators, we showed that racial misclassification obscures around 24% of the unwarranted racial disparity in foster care placement documented in earlier work. We also illustrated the design of policy responses that could potentially mitigate unwarranted disparities in foster care placement. These responses address discrimination not only during the investigation phase but also in preceding phases leading up to the investigation.

The methods developed in this paper may be useful in future studies of unwarranted disparity. While the study's application is to the child protection context, we note that this system shares several similarities with other high-stakes settings. Our methods for estimating the impact of racial misclassification on measures of unwarranted disparity, for example, could be applied in studies of the criminal justice system. Racial categorization in administrative datasets frequently used in estimating racial discrimination within the criminal justice system typically relies on coding by law enforcement officers or other involved decision-makers. Moreover, recent data collection efforts by the Criminal Justice Administrative Records System may enable researchers to cross-reference these records with self-reported racial and ethnic data from alternative sources, such as Census data (e.g., Finlay, Luh, and Mueller-Smith (2024)).

Second, akin to child protection, the criminal justice system operates across multiple phases. For instance, consider the potential involvement of police officers and prosecutors in the charging decision. Simply implementing a policy counterfactual to eliminate discrimination in prosecutor decisions, while overlooking the possibility of prior discrimination in police

officers' decisions to arrest, may leave remaining unwarranted disparity in charging rates. With the increasing availability of datasets tracking cases throughout the criminal justice system (Harrington and Shaffer, 2022), our methods can be used to construct algorithmic counterfactuals that address discrimination from a "systems-based" perspective. Similar arguments can also be extended to multi-phase systems in housing, lending, and employment. Consequently, future research could integrate the methods from this paper with recent advancements in such settings, which have developed various quasi-experimental tools for estimating racial discrimination (e.g., Arnold, Dobbie, and Yang (2018); Arnold, Dobbie, and Hull (2022); Goncalves and Mello (2021); Lodermeier (2023)).

Econometric Appendix

A.1 Identification Strategy in BDEHR

This appendix summarizes the quasi-experimental identification strategy in BDEHR. We begin by re-writing Equations (1)–(8) in terms of a set of directly estimable moments and a set of unknown parameters which capture the average risk of subsequent maltreatment among various populations of Black and white children. We then show how one can estimate these parameters leveraging the quasi-random assignment of screeners and investigators. We demonstrate this approach for estimating the screener UD measures here, and we then discuss how this approach can be easily extended to estimate both the investigator UD measures and the decompositions of placement UDs; BDEHR also discuss an alternative bounding approach, which we omit here for brevity.

First, note that the screener UD measures (1)-(2) can be rewritten as:

$$E[S_i \mid R_i = r, Y_i^* = 1] = 1 - \frac{E[(1 - S_i)Y_i^* \mid R_i = r]}{E[Y_i^* \mid R_i = r]} \tag{18}$$

and

$$E[S_i \mid R_i = r, Y_i^* = 0] = 1 - \frac{E[(1 - S_i)(1 - Y_i^*) \mid R_i = r]}{1 - E[Y_i^* \mid R_i = r]}, \tag{19}$$

Screening decisions S_i are observed; since the subsequent maltreatment outcomes of screened-out children (with $S_i = 0$) directly reveal their subsequent maltreatment potential, the numerators in both expressions do not suffer from the selective observability of Y_i^*. The challenge of estimating screener UDs thus reduces to the challenge of estimating the parameters in the denominators: the average subsequent maltreatment risk of Black and white children in

the full population of calls, $\mu_b = E[Y_i^* \mid R_i = b]$ and $\mu_w = E[Y_i^* \mid R_i = w]$. These parameters are not directly estimable because of the core selection challenge: Y_i^* is unobserved if child i is placed into foster care.

BDEHR's approach for estimating these key parameters builds on Arnold, Dobbie, and Hull (2021, 2022): they leverage variation in screener-specific maltreatment rates among children who are not placed in foster care. The intuition is as follows: consider a hypothetical screener whose calls have an eventual placement rate near zero. This could happen, for example, if this screener either screens-out all calls or if most of their screened-in calls do not result in foster care placement. If we were to randomly assign a subset of calls to this screener, the subsequent maltreatment rate among these calls would not suffer from the selection challenge. Moreover, by random assignment, the average Y_i^* among calls involving Black and white children in this subset of cases would be the same as the averages in the full population of calls.

BDEHR approximate this idealized experiment. Absent such a hypothetical screener, they estimate μ_r by extrapolating variation in observed subsequent maltreatment rates across quasi-randomly assigned screeners with very low placement rates. Specifically, they use regression adjustment to account for the non-random variation in screener assignment and either linear, quadratic, or local linear approximation to estimate μ_r from screener- and race-specific estimates of at-home maltreatment rates. With estimates of μ_r and the screener-specific numerators in Equations (18)-(19), they estimate screener-specific UD measures Δ_{j0}^S and Δ_{j1}^S following Equations (1) and (2). They then aggregate these screener-specific measures, weighting by caseloads, to estimate average screener UDs Δ_0^S and Δ_1^S. Finally, they average these measures weighting by $\bar{\mu} = \mu_b Pr(R_i = b) + \mu_w Pr(R_i = w)$ to estimate overall average screener UD. They use the same approach to estimate investigator UDs, placement UDs in the full set of calls, and the UD decompositions.

This empirical strategy relies on two primary assumptions: the as-good-as-random assignment of screeners and investigators and an exclusion restriction requiring that the assignment of screeners and investigators only impacts subsequent maltreatment potential through the decision to place children in foster care. BDEHR discuss and probe these assumptions empirically, showing that they are both plausible in the Michigan CPS context.

A.2 Race Misclassification Decomposition

This appendix derives the approximate decomposition in Equation (10) from an exact decomposition based on Bohren, Hull, and Imas (2022) and discusses identification. For

the latter, since for each $r^* \in \{b, w\}$ and $y \in \{0, 1\}$:

$$E[D_i \mid R_i^* = r^*, Y_i^* = y] = \sum_{r \in \{b,w,m\}} E[D_i \mid R_i^* = r^*, R_i = r, Y_i^* = y] Pr(R_i = r \mid R_i^* = r^*, Y_i^* = y),$$

we have by adding and subtracting quantities of the form $E[D_i \mid R_i^* = r^*, R_i = r, Y_i^* = y]$ $Pr(R_i = r \mid R_i^* = r', Y_i^* = y)$ for $r' \neq r^*$, and rearranging terms:

$$\begin{aligned}
&\Delta_y^* \equiv E[D_i \mid R_i^* = b, Y_i^* = y] - E[D_i \mid R_i^* = w, Y_i^* = y] \\
&= \sum_{r \in \{b,w,m\}} \left(E[D_i \mid R_i^* = b, R_i = r, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = r, Y_i^* = y]\right) Pr(R_i = r \mid R_i^* = w, Y_i^* = y) \\
&+ \sum_{r \in \{b,w,m\}} \left(Pr(R_i = r \mid R_i^* = b, Y_i^* = y) - Pr(R_i = r \mid R_i^* = w, Y_i^* = y)\right) E[D_i \mid R_i^* = b, R_i = r, Y_i^*].
\end{aligned}$$

Suppose now that no child who self-reports as Black is miscoded as white and vice-versa, such that $Pr(R_i = w \mid R_i^* = b) = Pr(R_i = b \mid R_i^* = w) = 0$. This eliminates several of the terms of the exact decomposition and yields an alternative decomposition by again adding, subtracting, and rearranging terms:

$$\begin{aligned}
\Delta_y^* =& E[D_i \mid R_i^* = b, R_i = b, Y_i^* = y] Pr(R_i = b \mid R_i^* = b, Y_i^* = y) \\
&+ E[D_i \mid R_i^* = b, R_i = m, Y_i^* = y] Pr(R_i = m \mid R_i^* = b, Y_i^* = y) \\
&- E[D_i \mid R_i^* = w, R_i = m, Y_i^* = y] Pr(R_i = m \mid R_i^* = w, Y_i^* = y) \\
&- E[D_i \mid R_i^* = w, R_i = w, Y_i^* = y] Pr(R_i = w \mid R_i^* = w, Y_i^* = y) \\
=& \left(E[D_i \mid R_i^* = b, R_i = m, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = m, Y_i^* = y]\right) Pr(R_i = m \mid R_i^* = b, Y_i^* = y) \\
&+ \left(Pr(R_i = m \mid R_i^* = b, Y_i^* = y) - Pr(R_i = m \mid R_i^* = w, Y_i^* = y)\right) E[D_i \mid R_i^* = w, R_i = m, Y_i^* = y] \\
&+ \left(Pr(R_i = b \mid R_i^* = b, Y_i^* = y) - Pr(R_i = w \mid R_i^* = w, Y_i^* = y)\right) E[D_i \mid R_i^* = w, R_i = w, Y_i^* = y] \\
&+ \left(E[D_i \mid R_i^* = b, R_i = b, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = w, Y_i^* = y]\right) Pr(R_i = b \mid R_i^* = b, Y_i^* = y) \\
=& \underbrace{\left(E[D_i \mid R_i^* = b, R_i = m, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = m, Y_i^* = y]\right)}_{\text{Obscured discrimination}} Pr(R_i = m \mid R_i^* = b, Y_i^* = y) \\
&+ \underbrace{\left(E[D_i \mid R_i^* = b, R_i = b, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = w, Y_i^* = y]\right)}_{\text{Revealed discrimination}} Pr(R_i = b \mid R_i^* = b, Y_i^* = y) \\
&+ \underbrace{\left(Pr(R_i = w \mid R_i^* = w, Y_i^* = y) - Pr(R_i = b \mid R_i^* = b, Y_i^* = y)\right) + \right. \\
&\left. \times \left(E[D_i \mid R_i^* = w, R_i = m, Y_i^* = y] - E[D_i \mid R_i^* = w, R_i = w, Y_i^* = y]\right)}_{\text{Residual}}.
\end{aligned}$$

Equation (10) approximates this decomposition by assuming $Pr(R_i = w \mid R_i^* = b) \approx 0$ and $Pr(R_i = b \mid R_i^* = w) \approx 0$ and ignoring the final residual term. In practice we find

only a small share of self-reported Black and white children are misclassified as white and Black and that misclassification rates are similar by self-reported race. These facts imply $Pr(R_i = w \mid R_i^* = w, Y_i^* = y) \approx Pr(R_i = b \mid R_i^* = b, Y_i^* = y)$, making the residual small.

Identification of these decompositions follow similarly to the BDEHR strategy discussed above. Each term of either decomposition can be written:

$$
\begin{aligned}
E[D_i \mid R_i^* = r^*, R_i = r, Y_i^* = 1] &= 1 - \frac{E[(1 - D_i)Y_i^* \mid R_i^* = r^*, R_i = r]}{E[Y_i^* \mid R_i^* = r^*, R_i = r]} \\
E[D_i \mid R_i^* = r^*, R_i = r, Y_i^* = 0] &= 1 - \frac{E[(1 - D_i)(1 - Y_i^*) \mid R_i^* = r^*, R_i = r]}{1 - E[Y_i^* \mid R_i^* = r^*, R_i = r]} \\
Pr(R_i = r \mid R_i^* = r^*, Y_i^* = 1) &= E[Y_i^* \mid R_i^* = r^*, R_i = r] \frac{Pr(R_i^* = r^*, R_i = r)}{Pr(R_i^* = r^*, Y_i^* = 1)} \\
&= E[Y_i^* \mid R_i^* = r^*, R_i = r] \frac{Pr(R_i^* = r^*, R_i = r)}{E[Y_i^* \mid R_i^* = r^*] Pr(R_i^* = r)} \\
&= E[Y_i^* \mid R_i^* = r^*, R_i = r] \frac{Pr(R_i = r \mid R_i^* = r)}{E[Y_i^* \mid R_i^* = r^*]} \\
Pr(R_i = r \mid R_i^* = r^*, Y_i^* = 0) &= (1 - E[Y_i^* \mid R_i^* = r^*, R_i = r]) \frac{Pr(R_i^* = r^*, R_i = r)}{Pr(R_i^* = r^*, Y_i^* = 0)} \\
&= (1 - E[Y_i^* \mid R_i^* = r^*, R_i = r]) \frac{Pr(R_i^* = r^*, R_i = r)}{(1 - E[Y_i^* \mid R_i^* = r^*]) Pr(R_i^* = r^*)} \\
&= (1 - E[Y_i^* \mid R_i^* = r^*, R_i = r]) \frac{Pr(R_i = r \mid R_i^* = r)}{1 - E[Y_i^* \mid R_i^* = r^*]}
\end{aligned}
$$

The first two equations are of the form of the UD components BDEHR estimate, just with a "combined" race category of $\{R_i^* = r, R_i = r^*\}$. We estimate them following the steps in Appendix A.1. The second two equations depend on the mean risk parameters $E[Y_i^* \mid R_i^* = r, R_i = r^*]$ that are estimated in this procedure, directly estimable racial (mis-)categorization probabilities $Pr(R_i = r \mid R_i^* = r)$, and additional mean risk parameters $E[Y_i^* \mid R_i^* = r]$ which can be estimated again by following the extrapolation approach in BDEHR.

References

Agadjanian, Alexander. 2022. "How Many Americans Change Their Racial Identification over Time?" *Socius* 8 23780231221098547.

Agan, Amanda Y, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias." Working Paper 30981, National Bureau of Economic Research.

Antle, Becky F, Anita P Barbee, Dana N Christensen, and Dana J Sullivan. 2009. "The prevention of child maltreatment recidivism through the solution-based casework model of child welfare practice." *Children and Youth Services Review* 31 (12): 1346–1351.

Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings* 111 49–54.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review* 112 (9): 2992–3038.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–1932.

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

Baron, E Jason, and Max Gross. 2022. "Is There a Foster Care-To-Prison Pipeline? Evidence from Quasi-Randomly Assigned Investigators." National Bureau of Economic Research Working Paper 29922.

Baron, Jason, Joseph Doyle, Natalia Emanuel, Peter Hull, and Joseph Ryan. 2024. "Discrimination in Multi-Phase Systems: Evidence from Child Protection." *NBER Working Paper #31490.*

Benson, Cassandra, Maria D Fitzpatrick, and Samuel Bondurant. 2022. "Beyond Reading, Writing, and Arithmetic: The Role of Teachers and Schools in Reporting Child Maltreatment." *Journal of Human Resources* forthcoming.

Bohren, J Aislinn, Peter Hull, and Alex Imas. 2022. "Systemic Discrimination: Theory and Measurement." National Bureau of Economic Research Working Paper 29820.

Canay, Ivan A, Magne Mogstad, and Jack Mountjoy. 2022. "On the Use of Outcome Tests for Detecting Bias in Decision Making." National Bureau of Economic Research Working Paper 27802.

Conrick, Kelsey M, Brianna Mills, Astrid B Schreuder et al. 2023. "Disparities in misclassification of race and ethnicity in electronic medical records among patients with traumatic injury." *Journal of racial and ethnic health disparities* 1–5.

Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff. 2019. "The causes and consequences of test score manipulation: Evidence from the New York regents examinations." *American Economic Journal: Applied Economics* 11 (3): 382–423.

DeFina, Robert, and Lance Hannon. 2016. "Social status attainment and racial category selection in the contemporary United States." *Research in Social Stratification and Mobility* 44 91–97.

Dinerstein, Michael, and Isaac M Opper. 2022. "Screening with Multitasking: Theory and Empirical Evidence from Teacher Tenure Reform." Working Paper 30310, National Bureau of Economic Research.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania. 2021. "Measuring Bias in Consumer Lending." *Review of Economic Studies* 88 (6): 2799–2832.

Feigenberg, Benjamin, and Conrad Miller. 2022. "Would Eliminating Racial Disparities in Motor Vehicle Searches have Efficiency Costs?" *Quarterly Journal of Economics* 137 (1): 49–113.

Finlay, Keith, Elizabeth Luh, and Michael Mueller-Smith. 2024. "Implications of Race and Ethnicity (Mis)Measurement in the U.S. Criminal Justice System." *Working Paper*.

Goncalves, Felipe, and Steven Mello. 2021. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review* 111 (5): 1406–1441.

Grimon, Marie-Pascale, and Christopher Mills. 2022. "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial." Job market paper.

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

Harrington, Emma, and Hannah Shaffer. 2022. "Brokers of Bias in the Criminal Justice System: Do Prosecutors Compound or Attenuate Racial Disparities in Policing?", Working paper.

Hull, Peter. 2021. "What Marginal Outcome Tests Can Tell Us about Racially Biased Decision-Making." National Bureau of Economic Research Working Paper 28503.

Jacob, Brian A, and Steven D Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *The Quarterly Journal of Economics* 118 (3): 843–877.

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. 2018a. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018b. "Algorithmic Fairness." *AEA Papers and Proceedings* 108 22–27. 10.1257/pandp.20181018.

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.

Lodermeier, Alison. 2023. "Racial Discrimination in Eviction Filing."

Luh, Elizabeth. 2022. "Not so Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports." *Working Paper*.

Putnam-Hornstein, Emily, John Prindle, and Ivy Hammond. 2021. "Engaging Families in Voluntary Prevention Services to Reduce Future Child Abuse and Neglect: A Randomized Controlled Trial." *Prevention Science* 22 (7): 856–865.

Rambachan, Ashesh. 2022. "Identifying Prediction Mistakes in Observational Data." Working paper.

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig. 2021. "An Economic Approach to Regulating Algorithms." National Bureau of Economic Research Working Paper 27111.

Rittenhouse, Katherine. 2022. "Income and Child Maltreatment: Evidence from a Discontinuity in Tax Benefits." https://krittenh.github.io/katherine-rittenhouse.com/RittenhouseJMP_current.pdf, Working Paper.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2022. "Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool." Working paper.

Rose, Evan K. 2023. "A constructivist perspective on empirical discrimination research." *Journal of Economic Literature* 61 (3): 906–923.

Roth, Wendy D. 2012. *Race migrations: Latinos and the cultural transformation of race*. Stanford University Press.

Samant, Anjana, Aaron Horowitz, Sophie Beiers, and Kath Xu. 2021. "Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Heard Of." *ACLU*, https://www.aclu.org/news/womens-rights/family-surveillance-by-algorithm-the-rapidly-spreading-tools-few-have-heard-of.

Smith, Ning, Rajan L Iyer, Annette Langer-Gould et al. 2010. "Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children." *BMC Health Services Research* 10 (1): 1–7.

Telles, Edward, and Tianna Paschel. 2014. "Who is black, white, or mixed race? How skin color, status, and nation shape racial classification in Latin America." *American Journal of Sociology* 120 (3): 864–907.

Yi, Youngmin, Frank R Edwards, and Christopher Wildeman. 2020. "Cumulative prevalence of confirmed maltreatment and foster care placement for US children by race/ethnicity, 2011–2016." *American Journal of Public Health* 110 (5): 704–709.