NBER WORKING PAPER SERIES

PREDICTIVE RISK SCORES IN THE PUBLIC SECTOR: EXPERIMENTAL EVIDENCE FROM CHILD-PROTECTION INVESTIGATIONS

E. Jason Baron
Arkadev Ghosh
Richard Lombardo

Working Paper 35540
http://www.nber.org/papers/w35540

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2026

This project was pre-registered in the AEA RCT Registry (AEARCTR-0012623) and was approved by Duke University's Institutional Review Board (Protocol 2025-0467). We thank Adwait Sinha and Benjamin Martin for excellent research assistance. We are grateful to Rhema Vaithianathan, Emily Putnam-Hornstein, Larissa Lorimer, and the Centre for Social Data Analytics for their partnership in developing and deploying the randomized evaluation. We also thank Susan Wandalowski and Maria Torres of Northampton County for their invaluable collaboration, their institutional expertise, and their leadership in facilitating access to the administrative data and ensuring the faithful implementation of the randomized evaluation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by E. Jason Baron, Arkadev Ghosh, and Richard Lombardo. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Predictive Risk Scores in the Public Sector: Experimental Evidence from Child-Protection Investigations

E. Jason Baron, Arkadev Ghosh, and Richard Lombardo

NBER Working Paper No. 35540
July 2026
JEL No. C93, I38, J13

ABSTRACT

Many public-sector decisions require allocating scarce attention under uncertainty. We examine whether algorithmic risk assessments improve child-protection decisions, where supervisors decide which cases need closer scrutiny. In a randomized evaluation of 4,752 child referrals over 14 months in Northampton County, supervisors received an algorithmic risk score alongside standard case records. Access to the score increased foster-care placements and services for children at highest predicted risk, with little change for lower-risk cases, and it reduced subsequent maltreatment referrals. We find no evidence that the score widened racial disparities in decisions or outcomes, suggesting algorithms can improve targeting while preserving human discretion.

E. Jason Baron
Duke University
Economics Department
and NBER
jason.baron@duke.edu

Arkadev Ghosh
Duke University
Economics
and NBER
arkadev.ghosh@duke.edu

Richard Lombardo
Harvard University
Department of Economics
richardlombardo@fas.harvard.edu

A randomized controlled trials registry entry is available at https://www.socialscienceregistry.org/trials/12623

I Introduction

Many high-stakes decisions in the public sector require determining which cases warrant intervention under uncertainty. Child-protection workers must decide which investigations warrant intensive scrutiny (Bald et al., 2022; Doyle and Aizer, 2018); health-care providers must determine which patients need additional testing or treatment (Mullainathan and Obermeyer, 2022); and courts must identify which defendants pose the greatest risk (Angelova et al., 2025; Imai et al., 2023; Kleinberg et al., 2018a; Rambachan, 2024). In each case, decision-makers must allocate scarce attention and intervention resources while synthesizing large amounts of information.

Algorithmic risk assessments promise to improve these decisions by distilling complex information into an actionable signal (Albright, 2023). Yet their effects in practice are uncertain: they may reproduce existing biases, distort professional judgment, or fail to improve decisions even when their predictions are accurate (Arnold et al., 2021; Barocas and Selbst, 2016; Kleinberg et al., 2018b; Ludwig and Mullainathan, 2021; Obermeyer et al., 2019). Most existing evidence evaluates predictive performance or compares algorithmic predictions with human judgment. A smaller experimental literature examines how decision-makers actually respond when given algorithmic information, but the evidence remains limited (Agarwal et al., 2023; Angelova et al., 2025; Imai et al., 2023; Stevenson and Doleac, 2024).

Child protection is an especially important setting in which to study these questions. Nearly 40% of U.S. children experience a child-protection-services (CPS) investigation before adulthood (Kim et al., 2017). The decisions made during these investigations – whether a child can safely remain at home, whether a family should receive ongoing services, or whether foster-care placement is necessary – are among the most consequential made by public agencies (Bald et al., 2022). Both over- and under-intervention can be costly. Failing to intervene when a child is at risk may lead to continued maltreatment, with lasting consequences for children's health and well-being and, in the most severe cases, death (Currie and Widom, 2010; Currie and Tekin, 2012; Lives Cut Short, 2025). Conversely, unnecessary intervention can impose substantial costs on children and families through family separation and continued involvement with the child-welfare system (Fong, 2020).

These decisions are also made under substantial uncertainty and are highly information-intensive. Workers must synthesize extensive histories of prior referrals, investigations, services, and family relationships under considerable time pressure – precisely the type of decision environment in which algorithmic decision support may be valuable. By summarizing complex administrative records into a simple risk signal, predictive risk assessments have the potential

to help workers identify the cases most likely to require intervention. Reflecting this potential, jurisdictions across the United States have begun deploying predictive risk models, and recent federal initiatives explicitly encourage the use of predictive analytics to improve child safety and caseworker decision-making (Administration for Children, Youth and Families, 2026).

Much of the existing evidence on such models, as we discuss below, focuses on the screening phase of the child-protection system: whether a referral is accepted for investigation. We instead study algorithmic decision support *during* the investigation, when supervisors and investigators decide how intensively the agency should respond. This distinction matters because the most consequential child-protection decisions – substantiation, ongoing services, and foster-care placement – are made during the investigation rather than at screening. It also means that we study the terminal stage of a multistage decision process. Improvements to screening decisions may ultimately be undone by investigators' subsequent decisions, whereas improvements during the investigation directly determine the agency's final response to a referral (Baron et al., 2026). To our knowledge, this is the first randomized evaluation of algorithmic decision support during child-protection investigations.

We conduct a family-level randomized controlled trial in Northampton County, Pennsylvania. From December 2023 to February 2025, we randomized whether supervisors could view a five-level algorithmic complexity score for a family's referrals ($N = 4,752$ child-referral observations). The score summarized administrative case histories and predicted future foster-care placement risk. All supervisors used the same case-management system and had access to the same underlying administrative records; the experiment randomized only whether the algorithmic summary score was visible for a given family's referrals.

Access to the algorithmic risk assessment changed child-protection decisions in ways consistent with more targeted intervention. On average, making the score visible increased foster-care placement by 0.10σ ($p < 0.05$) and continuation into ongoing services by 0.07σ ($p < 0.10$), while leaving substantiation rates largely unchanged. These effects were concentrated among referrals in the highest-risk category, for which score visibility increased foster-care placement by 0.46σ ($p < 0.10$), continuation into ongoing services by 0.49σ ($p < 0.01$), and substantiation by 0.32σ ($p < 0.05$). Effects are much smaller and less systematic for lower-risk referrals. Thus, investigators intervened disproportionately in the cases the algorithm identified as highest risk.

This improved targeting was followed by fewer subsequent referrals. Access to the score reduced re-referrals within three months by 0.09σ ($p < 0.05$), a reduction of 15%. Because referrals largely originate outside the agency, and overwhelmingly from mandated reporters

such as educators and health-care providers, they provide an externally generated measure of subsequent concern about child well-being. This decline is not mechanical: it is essentially unchanged when we exclude children placed in foster care during the focal investigation. Nor is it confined to marginal reports the agency would later have dismissed: among children who are re-referred, substantiation rates are similar across treatment arms. We find no evidence that score visibility systematically increased racial disparities in intervention decisions or subsequent child-welfare outcomes. Finally, we find suggestive evidence that the reduction in subsequent referrals is concentrated among investigators who were more skilled prior to the experiment, consistent with algorithmic information complementing, rather than substituting for, professional judgment.

Taken together, the results show that providing supervisors with access to an algorithmic risk score changed investigation decisions in ways consistent with better targeting of high-stakes interventions and reduced subsequent referrals, a proxy for child safety. These findings speak directly to a central policy question: what happens when predictive risk assessments are incorporated into frontline decision-making while preserving human discretion? This question is increasingly salient as jurisdictions adopt predictive risk assessments and recent federal guidance encourages the use of predictive analytics to improve child safety and decision-making (Administration for Children, Youth and Families, 2026). By randomizing access to the score, our experiment provides causal evidence on this form of algorithmic decision support and finds that it can improve targeting while preserving human judgment.

The score may have changed decision-making by helping workers form more accurate assessments of case risk, directing attention toward referrals warranting closer scrutiny, or providing additional support for acting on concerns already present in the case records. These channels may operate jointly, and the experiment is not designed to isolate their relative importance. We do, however, find an increase in recorded supervisory activity for the highest-risk referrals, consistent with attention playing a role. More importantly, the results show that a concise algorithmic signal changed how the highest-risk cases were handled, improved the targeting of consequential interventions, and reduced subsequent referrals.

This study contributes to a growing literature on algorithmic decision support in child protection, which to date has centered on the screening stage. In a randomized evaluation of an algorithmic screening tool, Fitzpatrick et al. (2025) find that algorithmic information altered screening decisions and reduced decision-making time; Grimon and Mills (2025) show that algorithmic decision support improved the allocation of investigations, reduced child harm, and decreased disparities in CPS surveillance; and Rittenhouse et al. (2022) find that the introduction of the Allegheny Family Screening Tool reduced racial disparities in

child-protection decisions using quasi-experimental variation. We complement this important work by moving to the investigation stage, where the relevant question is not *whether* a family is investigated but *how* intensively the agency responds. Together, these studies trace how algorithmic information shapes decisions across successive stages of the child-protection system.

Our findings also speak to the broader use of predictive analytics in public-sector decision-making (Albright, 2023; Angelova et al., 2025; Hoffman et al., 2018; Imai et al., 2023; Kleinberg et al., 2018a; Mullainathan and Obermeyer, 2022; Rambachan, 2024). In our setting, the algorithm did not determine which children entered the child-welfare system, recommend a particular course of action, or automate any decisions. Instead, it summarized complex administrative histories into a risk signal that supervisors could incorporate into their own judgment. Yet this information changed behavior. Many public-sector decisions are too consequential, context-dependent, or contested to delegate fully to an algorithm. Our results suggest that predictive tools can be valuable in precisely this middle ground, supporting human discretion rather than substituting for it.

II Setting and Experiment

Child protection investigations in Northampton County

The experiment was carried out during child-protection investigations in Northampton County, Pennsylvania, home to roughly 313,000 residents in the Lehigh Valley. Like other Pennsylvania counties, Northampton administers child welfare through a county-operated Children, Youth, and Families (CYF) agency. Reports of suspected maltreatment are submitted to Pennsylvania's statewide hotline (ChildLine) and then routed to the county. After that, they are either first screened for whether they warrant investigation or sent directly to the investigation, or "Intake" phase; direct-to-Intake routing is more common for families with recent system involvement. Figure A1 summarizes this process and Appendix B.1 provides further institutional detail. Our experiment concerns only referrals that reach Intake, and the algorithmic score played no role in the earlier screening and routing decisions.

At Intake, a supervisor assigns the referral to an investigator, who reviews records, interviews the child and caregivers, and gathers evidence on the allegations, generally within 30 days. Although supervisors oversee investigations and review cases throughout the assessment process, investigators exercise substantial discretion in determining how investigations are conducted and what information is collected. Investigation outcomes are typically determined through ongoing communication between investigators and supervisors, with investigators

reporting their findings and supervisors providing guidance before final determinations are made.

An investigation may result in several possible outcomes. One possible outcome is *substantiation*, meaning that investigators conclude there is sufficient evidence to support allegations of abuse or neglect. Investigators may also decide to open a case for *ongoing services*, allowing the family to continue receiving supervision, case management, safety planning, or referrals to community-based programs after the Intake investigation ends. In the most serious cases, investigators may determine that a child cannot safely remain in the home and petition the court to remove the child and place them in *foster care*. Investigations may also be closed without substantiation, the provision of services, or foster care placement.

Algorithmic risk assessment tool

In December 2023, Northampton County introduced an Intake Supervision Tool to support oversight of active child welfare investigations. The tool displays referrals in the Intake phase – that is, referrals that have been assigned to an Intake worker but have not yet received a final determination. Figure 1 shows the overview screen, which lists all active investigations along with summary information for each referral. This information includes the individuals involved in an allegation, the number and ages of children in the household, the alleged perpetrator, the nature of the allegation, and the family's prior (if any) contact with the agency.

The central feature of the tool is a predictive *complexity score*. The score was designed to identify children at elevated risk of chronic involvement in the child-welfare system and to provide supervisors with a concise summary of information contained in extensive administrative records. It is generated by a machine-learning model (XGBoost) trained on the county's historical administrative data to predict a child's risk of entering foster care following a referral. The model's prediction is reported to supervisors as one of five ordered categories: Very Low, Low, Medium, High, or Very High. Appendix B.2 describes the model and its construction. As shown in Figure 1, during the evaluation period the complexity category was displayed for treatment referrals on the overview screen used by supervisors within the case-management system. For control referrals, the same location instead displayed "Score not shown."

Two features of the tool are central to the experiment. First, the score compresses complex administrative histories into a single interpretable signal, but every input used to generate it was already available to supervisors and investigators through the county's case-management system. As a result, the experiment altered how information was presented rather than what

information was available to decision-makers. Second, the tool was designed to support, rather than replace, professional judgment. Supervisors were instructed to communicate the complexity category to the assigned investigator whenever it was visible, but both supervisors and investigators retained full discretion over all case decisions.

Randomized evaluation

The evaluation in this paper randomized whether supervisors could view the complexity category. The study ran from December 18, 2023, through February 15, 2025, and was pre-registered in the AEA RCT Registry (#0012623).[1] Randomization was implemented at the family level: when a family first appeared in the Intake Supervision Tool, it was assigned to either the treatment group, in which the complexity category was visible to supervisors, or the control group, in which the corresponding field displayed "Score not shown." Assignment persisted across all referrals involving that family during the study period. Aside from the visibility of the complexity category, treatment and control referrals were identical within the tool (see Appendix B.3 for more details).

Importantly, referrals were randomized only *after* they entered Intake; randomized access to the Intake complexity score was never available to screening staff and played no role in whether allegations were investigated. The experiment therefore identifies the causal effect of providing supervisors with algorithmic risk information during an investigation, holding constant all prior screening decisions and other aspects of the investigation process.

III Conceptual Framework

At the Intake stage, referrals have already been accepted for investigation, so the relevant decision is not whether a family should enter the child-protection system, but how intensively the agency should respond. Investigators must decide which cases warrant closer scrutiny, whether allegations should be substantiated, whether families should receive ongoing services, and, in the most serious cases, whether children should be placed in foster care. These are high-stakes decisions made under considerable uncertainty and discretion, with potentially large costs from both over- and under-intervention. Supervisors must also allocate limited attention across many active investigations.

The algorithmic complexity score summarized extensive administrative histories – including prior referrals, investigations, services, placements, and family relationships – into a concise assessment of case risk. Because the intervention only made this summary assessment visible – and left all decisions to workers – any treatment effect captures how the algorithmic risk

[1] Appendix C describes minor deviations from the pre-registration.

information was incorporated into professional judgment while retaining discretion (Angelova et al., 2025; De-Arteaga et al., 2020; Hoffman et al., 2018; Imai et al., 2023; Stevenson and Doleac, 2024).

By summarizing extensive case histories, the complexity score may help workers form more accurate assessments of underlying risk. It may also direct scarce supervisory attention toward referrals warranting closer scrutiny or provide additional support for acting on concerns already present in the case record, thereby affecting the effective threshold for intervention. These channels may operate jointly and potentially reinforce one another: greater scrutiny may lead workers to update their assessments of risk, while stronger assessments may affect both subsequent engagement and the likelihood of intervention.

Regardless of which channel is more important, these considerations motivate two central empirical questions. First, did access to the score improve the targeting of intervention? If so, its effects should be concentrated among children classified as highest risk rather than take the form of a uniform increase in agency involvement. We therefore examine recorded supervisory engagement via clicks in the Intake Supervision Tool, as well as three consequential investigation decisions – substantiation, ongoing service provision, and foster-care placement – across complexity categories. Second, were these changes accompanied by improvements in child safety? Because changes in intervention are not, by themselves, evidence of improved outcomes, we examine subsequent referrals to the child-protection system, which are generated primarily by external mandated reporters and serve as a proxy for child safety.

IV Data and Analysis Sample

Northampton County provided de-identified administrative data from its child-welfare case-management system. We observe referrals and report dates, allegation-level determinations, and foster-care placement records, which allow us to consider the three main investigator decisions we study: substantiation, continuation for ongoing services, and foster-care placement. Person-level identifiers let us link children across referrals and measure subsequent child-welfare involvement. Importantly, we observe the underlying complexity score for all referrals, regardless of whether the supervisors could view it during the experiment, because the score was generated for both treatment and control cases. We also observe usage logs from supervisors' interactions with the Intake Supervision Tool. We interpret these clicks as measures of active supervisory engagement rather than complete records of exposure to the score, because supervisors could view the category on the overview screen without clicking and therefore without generating a logged interaction. Finally, we observe personnel records containing investigator characteristics, which we use to examine heterogeneity across workers.

We build our analysis sample by restricting to Intake referrals reported during the experimental window (December 18, 2023–February 15, 2025), and by excluding a small set of families affected by a temporary software display issue. This yields 4,752 child-referral observations, corresponding to 3,488 unique children, 3,661 referrals, and 34 investigators. Appendix B.4 details the exact sample construction steps.

Table A1 reports summary statistics for the analysis sample. The average child is 9.2 years old and 52% are female. The sample is racially and ethnically diverse: 39% of children are recorded as White, 19% as Black, and 39% as Hispanic. About 8% of referrals involve a child with an active child-welfare investigation at the time of the focal referral. Panel B reports investigator decisions during the focal investigation. Foster-care placement is rare, occurring in approximately 2% of referrals. Approximately 5% of referrals are opened for ongoing services following the investigation, while 32% result in substantiation.

Panel C reports subsequent child welfare involvement. Approximately one-quarter of children are referred to the child protection system again within three months of the determination date, 21% have a subsequent referral screened in for investigation, and 8% have a subsequent referral substantiated. Thus, although intensive interventions such as foster care placement and ongoing services are relatively uncommon during the focal investigation, subsequent contact with the child protection system is common. These patterns underscore a central targeting challenge: identifying, during the initial investigation, which children are most likely to require additional protection or services.

Figure A2 reports covariate balance between the treatment and control groups. Across the 11 baseline child and perpetrator characteristics examined, only one difference is statistically significant at conventional levels: children in the treatment group are more likely to be recorded as White. All subsequent analyses control for child race and ethnicity, along with the other pre-treatment characteristics in the candidate control set.

V Empirical Strategy

We estimate intent-to-treat (ITT) effects of making the algorithmic complexity category visible to supervisors. Recall that treatment was assigned at the family level and persisted across all of a family's referrals during the study period. The unit of observation is a child-referral, and we cluster standard errors at the family level. Our pooled specification is

$$Y_{ifrd} = \beta T_f + X'_{ifr}\gamma + \delta_d + \theta_s + \mu_j + \varepsilon_{ifrd}, \tag{1}$$

where Y_{ifrd} denotes the outcome for child i in family f on referral r reported on date d, and T_f is the treatment indicator which equals one if supervisors could observe the complexity category. The vector X_{ifr} contains pre-treatment child and perpetrator characteristics, selected separately for each outcome using double-selection LASSO (Belloni et al., 2014) as described in detail in Appendix B.5.[2] Because treatment is randomly assigned, the treatment effect is identified without additional controls. We use double-selection LASSO solely to select pre-treatment covariates that improve precision while maintaining valid inference after data-driven control selection. The term δ_d denotes referral-date fixed effects, θ_s denotes fixed effects for whether the referral entered Intake directly or first passed through Screening, and μ_j denotes investigator fixed effects. These fixed effects are included only to improve precision.[3]

To examine heterogeneity by algorithmic risk, we interact treatment assignment with the five complexity categories displayed in the Intake Supervision Tool (Very Low to Very High). Letting G_r denote the referral's complexity category, we estimate

$$Y_{ifrd} = \sum_k \beta_k \left(T_f \times \mathbf{1}\{G_r = k\}\right) + \sum_k \lambda_k \mathbf{1}\{G_r = k\} + X'_{ifr}\gamma + \delta_d + \theta_s + \mu_j + \varepsilon_{ifrd} \quad (2)$$

where each β_k denotes the treatment effect for referrals in complexity category k. All specifications include the same controls and fixed effects as Equation (1).

We report treatment effects in units of the control-group standard deviation of each outcome to facilitate comparisons across outcomes with different base rates. For ease of interpretation, Appendix B.5 also reports the corresponding estimates in percentage-point units.

VI Main Results

The complexity score predicts adverse child-welfare outcomes

Before presenting the experimental estimates, we first examine whether the algorithmic complexity score captures meaningful variation in underlying child-welfare risk. Figure A3 plots outcome rates by complexity category for referrals *in the control group*. The complexity score was generated by an XGBoost model trained on historical administrative records to predict foster-care placement within one year of a referral. Although the model was trained to predict this specific outcome, the resulting risk classification is strongly associated with a broader set of adverse child-welfare outcomes. Foster-care placement rates increase sharply

[2]The candidate set of controls consists of pre-treatment child and perpetrator characteristics: child race and ethnicity, sex, and age; an indicator for an active investigation; and alleged-perpetrator age, sex, race, and active-placement status.

[3]Figures A6 and A7 show robustness of the point estimates to alternative specification choices.

across complexity categories, with children classified as Very High complexity substantially more likely to experience placement than those classified as Very Low complexity. Similar monotonic relationships are observed for ongoing service provision, substantiation, and subsequent referrals within three months of the investigation determination date. These results indicate that the complexity score captures meaningful variation in underlying child-welfare risk and therefore provided supervisors with a potentially informative summary of case risk during the experiment.

Because the complexity category was displayed only to supervisors, the intervention could affect investigator actions only if supervisors communicated the score to the assigned investigator. To assess whether this occurred, we incorporated questions about the complexity category into Northampton's routine supervisory review process. Beginning in early 2024, supervisors recorded the complexity category reported by the assigned investigator during the second ten-day review of each investigation. By this stage, investigators had typically been working on the case for several weeks, providing ample opportunity for the score to have been communicated during earlier supervisory interactions.

Figure A4 compares the complexity categories reported by investigators with the categories assigned to their referrals. The correspondence is high: investigators correctly identified the assigned category in the majority of cases, and most discrepancies were limited to adjacent categories. Because these assessments occurred several weeks after the investigation began, some degree of recall error is expected. Nevertheless, the results indicate that the algorithmic information was generally communicated from supervisors to investigators and remained sufficiently salient to be recalled during routine supervisory reviews.

Access to the algorithm changed investigation decisions

We next examine how access to the algorithmic complexity score affected key decisions during the focal investigation. Figure 2 reports treatment effects on foster-care placement, ongoing service provision, and substantiation.

Panel A reports the average treatment effects. Access to the complexity score increased foster-care placements by 0.10σ ($p < 0.05$) and ongoing service provision by 0.07σ ($p < 0.10$). Because both interventions are relatively rare, these estimates represent sizable proportional increases. As shown in Table A2, the estimated effect on foster-care placement is 1.2 percentage points, nearly doubling the control-group placement rate of 1.3%. In contrast, we find little evidence of an average effect on substantiation decisions.

Average treatment effects, however, mask substantial heterogeneity across algorithmic risk groups. Figure 2 therefore also reports treatment effects separately by the algorithmic

complexity category. The effects are concentrated among children classified as Very High complexity. For these children, access to the score increased foster-care placement by 0.46σ ($p < 0.10$), ongoing service provision by 0.49σ ($p < 0.01$), and substantiation by 0.32σ ($p < 0.05$). In contrast, treatment effects are generally smaller and statistically indistinguishable from zero for children in the lower complexity categories.

These results show that supervisors and investigators did not respond to the algorithm by uniformly increasing intervention. Instead, access to the complexity score primarily increased intervention among children whom the algorithm identified as highest risk, consistent with improved targeting of scarce interventions. Figure A5 shows a similar pattern in supervisory activity: while overall recorded tool usage was largely unchanged, recorded activity increased for referrals classified as Very High complexity. Although the click logs capture only part of supervisory activity, they provide complementary evidence that access to the score changed how supervisors interacted with the highest-risk referrals.

Foster-care placement and ongoing service provision are consequential interventions whose welfare effects depend critically on which children receive them. Prior research suggests that the effects of foster-care placement vary substantially with underlying child risk, with the greatest benefits accruing to children facing the highest levels of danger in the home (Baron and Gross, 2025; Doyle, 2008). The concentration of treatment effects among the highest-risk children is therefore consistent with improved targeting of these interventions.

Algorithmic information reduced subsequent referrals

Having established that access to the score changed investigation decisions, we next examine whether these changes were accompanied by differences in subsequent child-protection involvement. Our primary downstream outcome is whether a child experienced another referral to the child-protection system within three months of the focal investigation's determination date. Because referrals originate primarily outside the child-welfare agency – from educators, health-care providers, law enforcement, social-service providers, and other community members – they provide an externally generated measure of subsequent concern about child well-being. This distinguishes re-referrals from subsequent investigations or substantiations, which also depend on additional decisions made within the agency.

Figure 3 presents the main downstream results. Access to the complexity score reduced re-referrals within three months by 0.09σ ($p < 0.05$). As shown in Table A3, this corresponds to a 3.8 percentage-point decline relative to a control mean of 25.8% – a 15% reduction. One concern is that this decline could arise mechanically from the increase in foster-care placements documented above. Because placement removes a child from the home, it may

reduce the opportunity for a subsequent referral independent of any change in the child's underlying circumstances. To assess this possibility, the right panel of Figure 3 re-estimates the treatment effect after excluding children who entered foster care during the focal investigation. The estimated effect is essentially unchanged, consistent with the relatively small share of children placed in foster care during the study period. The reduction in re-referrals therefore does not appear to be driven by the mechanical effect of removal. Instead, it suggests that the observed decline reflects broader changes in the child's circumstances in the home.

Figure 3 also reports treatment effects by algorithmic complexity category. The point estimates are negative across all risk groups, although the confidence intervals are wide for the category-specific estimates. The largest reduction occurs among referrals classified as Very High complexity, for whom access to the score reduced re-referrals by 0.19σ ($p = 0.14$). This is also the group for which the algorithm produced the largest changes in investigation decisions.

We next examine whether the treatment affected the nature of subsequent referrals. One possibility is that the overall decline was driven entirely by false or minor reports that, for example, would ultimately have been screened out or not substantiated – suggesting that child safety was essentially unchanged. Conditioning on children who experienced a subsequent referral, Panel A of Figure 4 compares the agency's response to those referrals across the treatment and control groups. The rates at which subsequent referrals were screened in for investigation and substantiated are very similar across groups.[4] Subsequent foster-care placements were rare and nearly identical across treatment arms.

Panel B of Figure 4 examines the distribution of reporter types among subsequent referrals. The composition is broadly similar across the treatment and control groups, with no evidence that the reduction in re-referrals was concentrated among any particular reporter type. If anything, subsequent referrals in the treatment group were somewhat more likely to originate from educators, law enforcement, and health-care providers. Because these referrals are generated by individuals outside the child-welfare agency, this pattern suggests that the reduction in re-referrals reflects fewer subsequent concerns about child well-being among external observers rather than changes in the agency's own decision-making. Together with the evidence in Panel A, these findings are consistent with a reduction in the circumstances that give rise to subsequent reports of suspected maltreatment.

Taken together, the evidence in this section is suggestive of improved child safety. Access

[4]The relatively higher screen-in rate in this figure may reflect Direct-to-Intake routing. Because these referrals occurred shortly after the focal investigation, they generally fell within the county's informal "bounceback" window and were routed directly to the Intake phase, which we classify as screened in for investigation.

to the algorithm changed investigation decisions by concentrating intervention among the highest-risk children and was accompanied by fewer subsequent referrals, with no evidence that the decline was driven by changes in reporter composition or confined to referrals the agency would later dismiss.

VII Robustness checks and extensions

Alternative specifications. We first examine the robustness of our main findings to alternative specification choices in Figures A6 and A7. Beginning with unadjusted estimates, we sequentially add referral-date fixed effects, an indicator for Direct-to-Intake routing, pre-treatment covariates selected using double-selection LASSO, investigator fixed effects, and General Protective Services (GPS) fixed effects, which account for differences in allegation type and the corresponding investigative unit. For the investigation-decision results, we focus on foster-care placement because it is the most intensive intervention available during the Intake investigation. Across specifications, the estimated treatment effect on foster-care placement remains positive and changes little as additional controls and fixed effects are introduced (Figure A6). Figure A7 shows a similarly stable pattern for subsequent referrals: the estimated reduction remains close to 0.10σ across specifications and remains statistically significant throughout.

Re-referral response over time. We next examine the sensitivity of our results to the choice of follow-up horizon. Figure A8 plots the cumulative probability of re-referral by treatment assignment over the five months following the focal investigation's determination date. The treatment group exhibits a lower cumulative re-referral rate than the control group throughout most of the follow-up period. The gap opens gradually over the first three months before narrowing somewhat at longer horizons, although the treatment group continues to exhibit lower cumulative re-referral rates throughout the five-month follow-up period.

Heterogeneity by investigator characteristics. The effects of algorithmic decision support may depend on the characteristics of the workers using it. Using administrative personnel records, we examine treatment-effect heterogeneity for our primary downstream outcome – re-referrals within three months of the focal investigation's determination date – across investigator demographic and professional characteristics.

Figure A9 reports treatment effects by investigator subgroup. Overall, we find little evidence that the effects of access to the algorithmic complexity score vary systematically across investigator characteristics. Point estimates are always negative, and confidence intervals overlap substantially across subgroups.

Moving beyond investigator background characteristics, Figure A10 examines treatment-effect

heterogeneity by two pre-RCT measures of investigator skill. Panel A classifies investigators according to their historical false-negative rate: the tendency to close cases without foster-care placement or ongoing services that were subsequently followed by a re-referral within three months. Under this measure, higher-skilled investigators are those with fewer children who subsequently returned to the child-protection system without receiving an intervention.

Because this measure captures only one dimension of investigator performance, Panel B instead uses a pre-experimental performance, following Baron et al. (2024), which combines false-negative and false-positive rates into a single summary measure of investigator performance. This measure exploits the quasi-random rotational assignment of Intake cases to identify differences in false-positive rates across investigators despite the selective-labels problem. Additional details are provided in Appendix B.6.

We estimate treatment effects separately for investigators above and below the median of each skill measure. Figure A10 shows that the reduction in subsequent re-referrals is generally larger among investigators with stronger pre-experimental performance. Using the performance measure in Panel B, access to the complexity score significantly reduced three-month re-referrals among higher-performing investigators, while the corresponding estimate for lower-performing investigators is close to zero. Panel A, which classifies investigators by their historical false-negative rates, yields a similar qualitative pattern, although the difference between the two groups is not statistically significant. We find little evidence of systematic heterogeneity in investigation decisions themselves, although there is suggestive evidence that the increases in foster-care placement and ongoing service provision were somewhat larger among higher-performing investigators.

Overall, these results provide suggestive evidence that the value of the algorithmic score depended on investigators' pre-existing skill, consistent with algorithmic information complementing rather than substituting for professional judgment. However, the estimates are generally imprecise and should be interpreted cautiously, as sufficient pre-experimental data were available to estimate baseline skill for only 19 investigators.

Heterogeneity by child characteristics. Figure A11 reports treatment effects on re-referrals within three months of the focal investigation's determination date across child subgroups defined by age, gender, and prior child-protection involvement (heterogeneity by child race is examined separately in the next section). Point estimates are uniformly negative, and confidence intervals overlap substantially across subgroups. Overall, we find little evidence that the effects of access to the algorithmic risk score vary systematically across observable child characteristics.

Algorithmic information did not increase racial disparities

An important concern surrounding algorithmic risk assessments is that they may exacerbate existing racial and ethnic disparities in public-sector decision-making (Arnold et al., 2021; Barocas and Selbst, 2016; Bohren et al., 2022; Kleinberg et al., 2018b; Obermeyer et al., 2019). In child welfare, concerns have been raised that predictive models trained on historical administrative records may reinforce existing patterns of differential intervention across groups (Rittenhouse et al., 2022). Although race was not used as an input to the algorithm, the resulting risk classifications are weakly correlated with child race (Figure A12). We therefore examine whether providing supervisors with access to the algorithmic risk score affected investigation decisions or subsequent child-protection outcomes differently across racial and ethnic groups.

Figure A13 reports treatment effects separately for White, Black, and Hispanic children. Overall, we find little evidence that providing supervisors with access to the algorithmic risk score systematically widened racial or ethnic disparities in either investigation decisions or subsequent child-protection outcomes. Although point estimates vary across groups, most differences are imprecisely estimated and do not follow a consistent pattern across outcomes. One exception is ongoing service provision, where the estimated treatment effect is larger for White and Black children than for Hispanic children. However, this pattern is not reflected in foster-care placement, substantiation, or subsequent referrals.

The re-referral results tell a similar story. The estimated reduction in three-month re-referrals is largest for Black children, smaller for Hispanic children, and close to zero for White children. Thus, the group experiencing the largest increase in intervention also experienced the largest reduction in subsequent referrals. Although treatment effects vary somewhat across racial and ethnic groups, we find no systematic evidence that providing supervisors with access to the algorithmic risk score widened disparities in either investigation decisions or subsequent child-protection outcomes.

VIII Discussion and Conclusion

This paper provides experimental evidence that making an algorithmic complexity score visible during child-protection investigations changed both where workers directed attention and how they intervened. Increases in investigator intervention and recorded supervisory activity were concentrated among referrals classified as highest risk, while subsequent referrals fell by 15 percent. We find no evidence that the score increased racial disparities.

The intervention neither prescribed an action nor displaced professional discretion. Instead,

it reorganized information already available to workers into a salient assessment of risk. The score may have influenced decisions by improving workers' assessments of underlying risk, directing attention toward cases warranting closer scrutiny, or providing additional support for acting on existing concerns. These responses may reinforce each other as well: greater scrutiny can generate belief updating, while stronger assessments of risk can affect both engagement and the willingness to intervene. The increase in recorded supervisory activity for the highest-risk referrals is consistent with attention playing a role, while the concentration of downstream gains among higher-skilled investigators suggests complementarity between algorithmic information and professional judgment.

The absence of an increase in overall recorded activity is also notable. Rather than inducing supervisors to spend substantially more effort across their entire caseload, the algorithm appears to have changed how supervisors responded to a relatively small set of referrals identified as highest risk. This pattern suggests that the observed improvements in targeting did not require a broad increase in supervisory workload, which may be important for the scalability of algorithmic decision support in practice.

More broadly, our findings provide evidence that predictive risk scores can improve the targeting of interventions in high-stakes public-sector decisions while preserving professional discretion. Future research could examine how different ways of presenting and incorporating algorithmic information shape workers' assessments of risk, attention to cases, and willingness to intervene. Such work would help clarify the mechanisms through which algorithmic decision support complements professional expertise and the settings in which it is most effective.

References

Administration for Children, Youth and Families, "Modernizing Child Welfare Technologies and Tools: Opportunities for Predictive Risk Modeling to Improve Child Safety and Outcomes," Issue Brief, Administration for Children, Youth and Families, U.S. Department of Health and Human Services March 2026.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working Paper 31422, National Bureau of Economic Research 2023.

Albright, Alex, "The hidden effects of algorithmic recommendations," *Working Paper*, 2023.

Angelova, Victoria, Will Dobbie, and Crystal S Yang, "Algorithmic recommendations and human discretion," *Review of Economic Studies*, 2025.

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

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

Barocas, Solon and Andrew D Selbst, "Big data's disparate impact," *Calif. L. Rev.*, 2016, *104*, 671.

Baron, E Jason and Max Gross, "Is there a foster care-to-prison pipeline? Evidence from quasi-randomly assigned investigators," *Review of Economics and Statistics*, 2025, pp. 1–46.

Baron, E. Jason, Richard Lombardo, Joseph P. Ryan, Jeongsoo Suh, and Quitzé Valenzuela-Stookey, "Reforming Rotations," Working Paper 32369, National Bureau of Economic Research April 2024.

_ , Will Dobbie, Richard Lombardo, Ashesh Rambachan, and Joseph Ryan, "Human Decisions and Machine Predictions in Multistage Systems," *Working Paper*, 2026.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 2014, *81* (2), 608–650.

Bohren, J Aislinn, Peter Hull, and Alex Imas, "Systemic Discrimination: Theory and Measurement," 2022. NBER Working Paper 29820.

Currie, Janet and Cathy Widom, "Long-Term Consequences of Child Abuse and Neglect on Adult Economic Well-Being," *Child maltreatment*, 05 2010, *15*, 111–20.

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

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova, "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" Association for Computing Machinery 2020, pp. 1–12.

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

_ and Anna Aizer, "Economics of child protection: Maltreatment, foster care, and intimate partner violence," *Annual review of economics*, 2018, *10*, 87–108.

Fitzpatrick, Maria D, Katharine Sadowski, and Christopher Wildeman, "Algorithms and Decision-making: Evidence from Child Maltreatment Reports," *Journal of Human Resources*, 2025.

Fong, Kelley, "Getting eyes in the home: Child protective services investigations and state surveillance of family life," *American Sociological Review*, 2020, *85* (4), 610–638.

Grimon, Marie-Pascale and Christopher Mills, "Better together? a field experiment on human-algorithm interaction in child protection," *arXiv preprint arXiv:2502.08501*, 2025.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in hiring," *The Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin, "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2023, *186* (2), 167–189.

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

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

_ , Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan, "Algorithmic fairness," in "Aea papers and proceedings," Vol. 108 2018, pp. 22–27.

Lives Cut Short, "Child Maltreatment 2023: How the Government Continues to Undercount Maltreatment Fatalities and Report Misleading Trend Data," *Lives Cut Short*, February 2025.

Ludwig, Jens and Sendhil Mullainathan, "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System," *Journal of Economic Perspectives*, 2021, *35* (4), 71–96.

Mullainathan, Sendhil and Ziad Obermeyer, "Diagnosing physician error: A machine learning approach to low-value health care," *The Quarterly Journal of Economics*, 2022, *137* (2), 679–727.

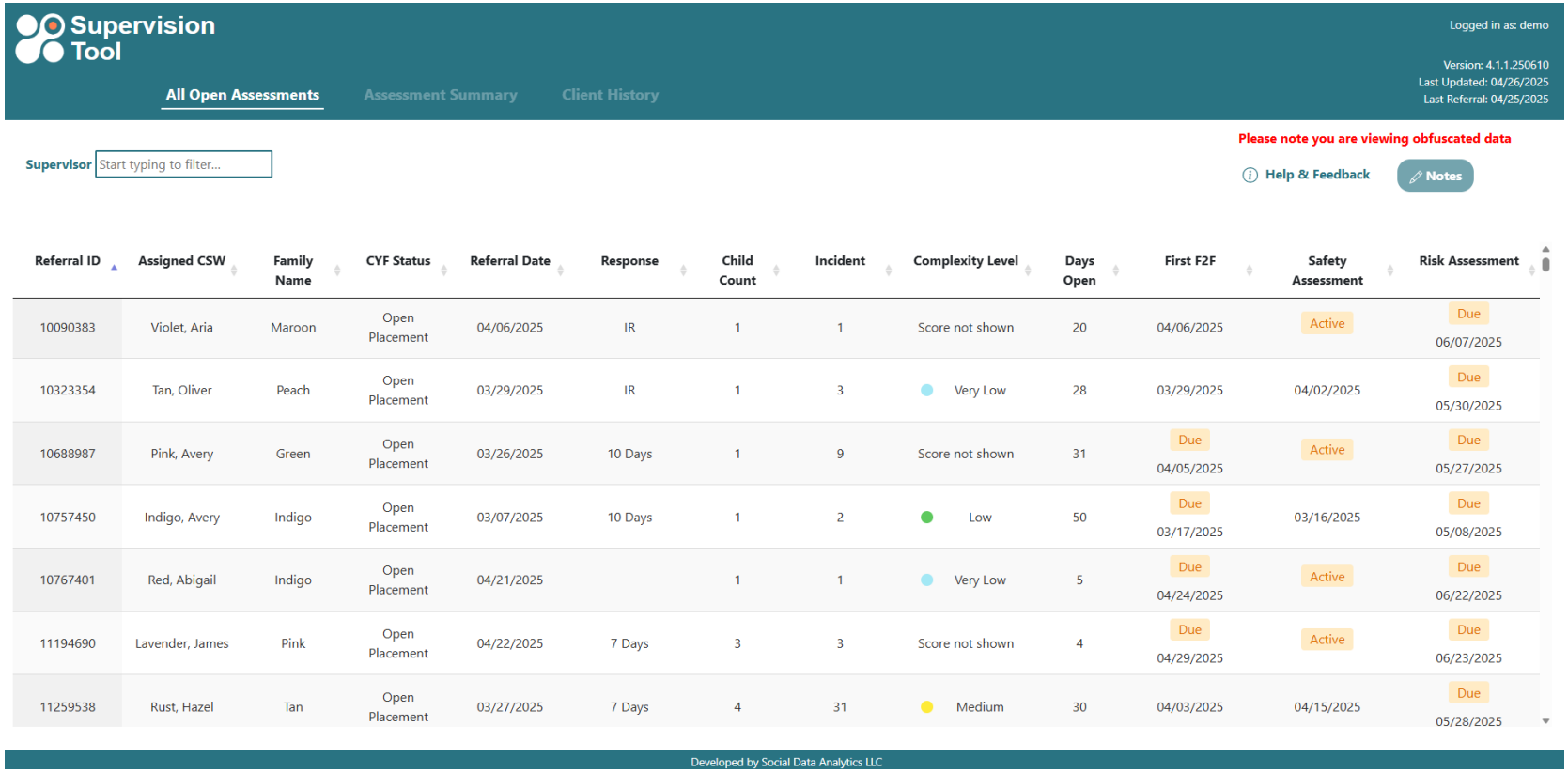
Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 2019, *366* (6464), 447–453.

Rambachan, Ashesh, "Identifying prediction mistakes in observational data," *The Quarterly Journal of Economics*, 2024, *139* (3), 1665–1711.

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

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

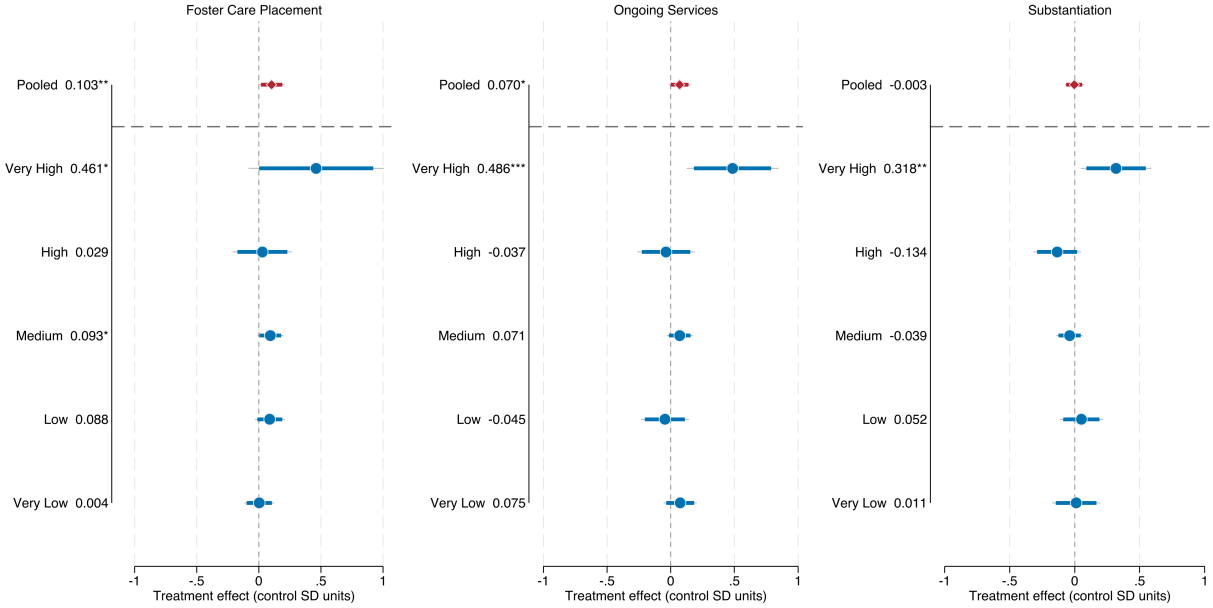
Figure 1. Northampton County's Intake Supervision Tool



Referral ID	Assigned CSW	Family Name	CYF Status	Referral Date	Response	Child Count	Incident	Complexity Level	Days Open	First F2F	Safety Assessment	Risk Assessment
10090383	Violet, Aria	Maroon	Open Placement	04/06/2025	IR	1	1	Score not shown	20	04/06/2025	Active	Due 06/07/2025
10323354	Tan, Oliver	Peach	Open Placement	03/29/2025	IR	1	3	Very Low	28	03/29/2025	04/02/2025	Due 05/30/2025
10688987	Pink, Avery	Green	Open Placement	03/26/2025	10 Days	1	9	Score not shown	31	Due 04/05/2025	Active	Due 05/27/2025
10757450	Indigo, Avery	Indigo	Open Placement	03/07/2025	10 Days	1	2	Low	50	Due 03/17/2025	03/16/2025	Due 05/08/2025
10767401	Red, Abigail	Indigo	Open Placement	04/21/2025		1	1	Very Low	5	Due 04/24/2025	Active	Due 06/22/2025
11194690	Lavender, James	Pink	Open Placement	04/22/2025	7 Days	3	3	Score not shown	4	Due 04/29/2025	Active	Due 06/23/2025
11259538	Rust, Hazel	Tan	Open Placement	03/27/2025	7 Days	4	31	Medium	30	04/03/2025	04/15/2025	Due 05/28/2025

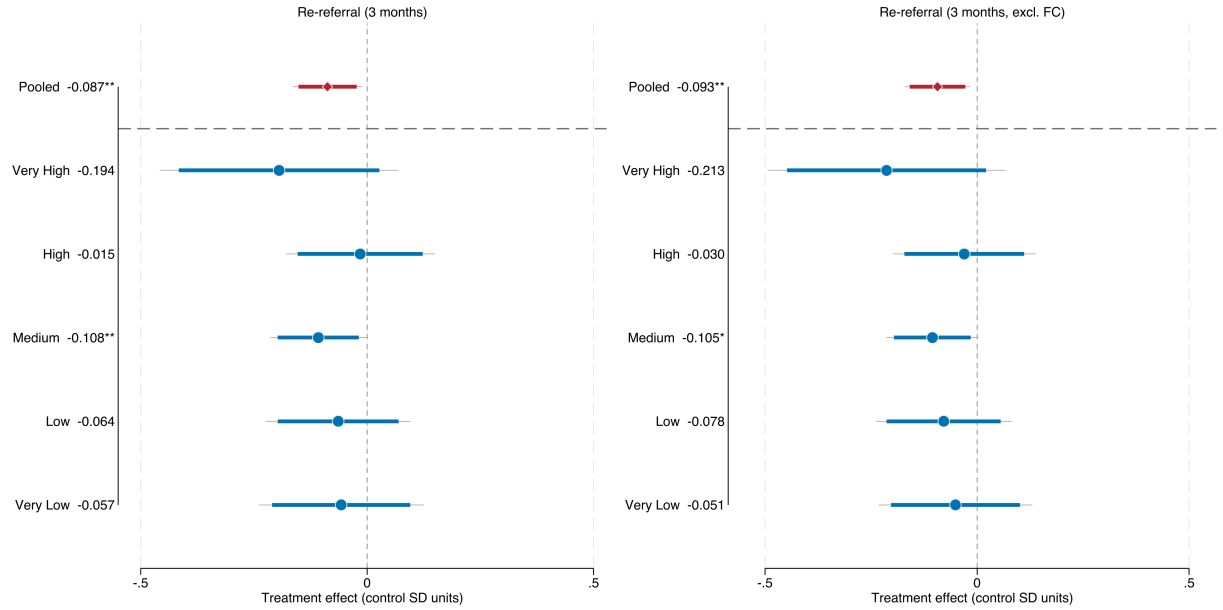
Notes. The figure displays Northampton County's Intake Supervision Tool. The algorithmic complexity category was visible only for referrals assigned to treatment; for control referrals, the category said "Score not shown." The intervention studied in this paper was the randomized visibility of this category to supervisors.

Figure 2. Effect of Algorithmic Risk Information on Investigation Decisions



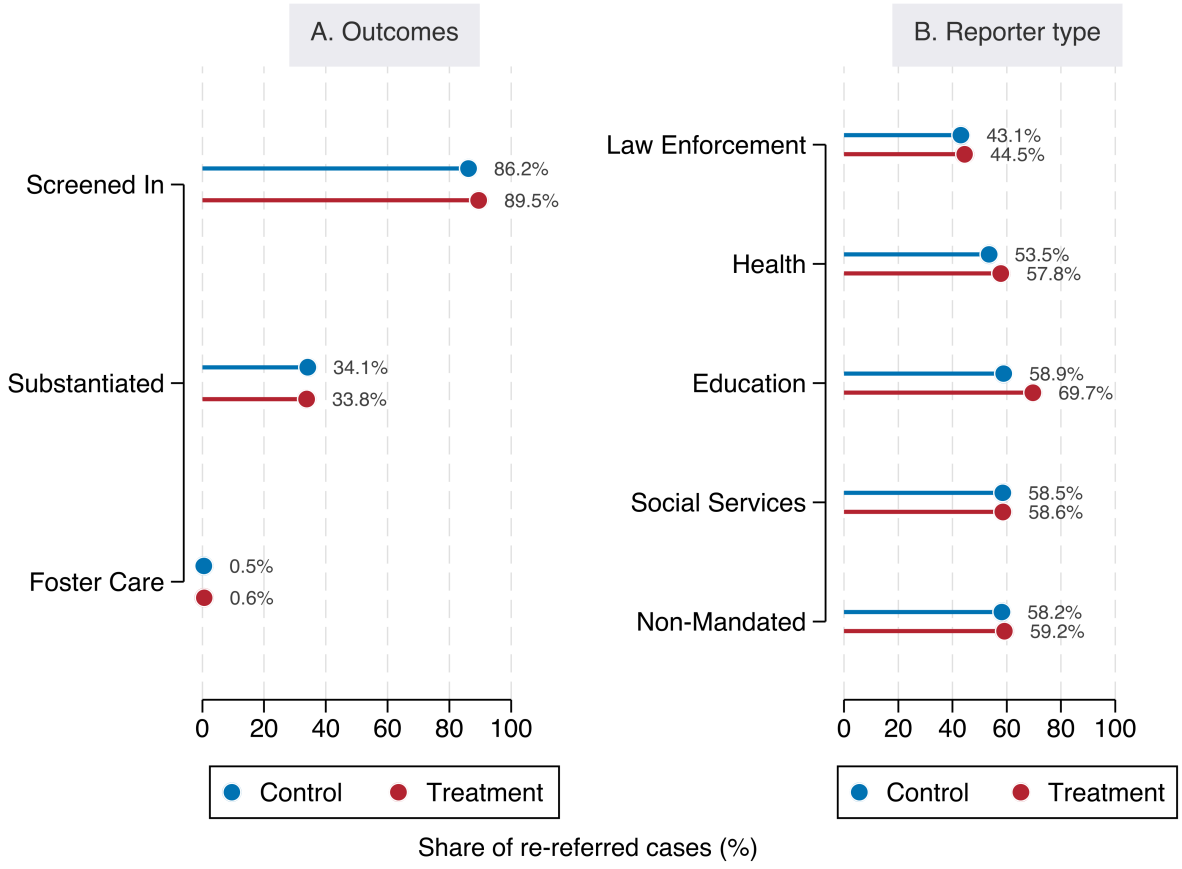
Notes. This figure reports intent-to-treat (ITT) estimates of providing supervisors with access to the algorithmic complexity score. The red diamonds report the pooled treatment effects, while the blue circles report treatment effects separately by algorithmic complexity category (Very Low to Very High). All regressions include referral-date fixed effects, investigator fixed effects, an indicator for Direct-to-Intake routing, and pre-treatment covariates selected using double-selection LASSO (Belloni et al., 2014). Standard errors are clustered at the family level. Effects are standardized by the control-group standard deviation of the corresponding outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively.$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 3. Effect of Algorithmic Risk Information on Subsequent Referrals



Notes. This figure reports intent-to-treat (ITT) estimates of providing supervisors with access to the algorithmic complexity score on re-referrals to the child-protection system within three months of the focal investigation's determination date. Panel A reports estimates for the full sample. Panel B excludes referrals in which the child entered foster care during the focal investigation to assess whether the results are driven mechanically by removal from the home. The red diamonds report the pooled treatment effects, while the blue circles report treatment effects separately by algorithmic complexity category (Very Low to Very High). Regressions include referral-date fixed effects, investigator fixed effects, an indicator for Direct-to-Intake routing, and pre-treatment covariates selected using double-selection LASSO (Belloni et al., 2014). Standard errors are clustered at the family level. Effects are standardized by the control-group standard deviation of the corresponding outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 4. Agency Responses and Reporter Types for Subsequent Referrals



Notes. This figure describes subsequent referrals occurring within three months of the focal investigation's determination date. The sample is restricted to children who experienced a subsequent referral during this follow-up period. Panel A reports the share of subsequent referrals that were screened in for investigation, substantiated, or resulted in foster-care placement. Panel B reports the share of subsequent referrals associated with each reporter type; reporter categories are not mutually exclusive. Bars report raw means separately for the treatment and control groups.

Online Appendix

A Supplemental Tables and Figures

Table A1. Descriptive Statistics of the Analysis Sample

	(1) Mean	(2) Missing Share
Panel A: Child Characteristics		
White	0.39	0.08
Black	0.19	0.08
Hispanic	0.39	0.04
Female	0.52	0.02
Active Investigation?	0.08	0.02
Age	9.22	0.02
Perpetrator Age	37.50	0.13
Any Perpetrator with Active Placement?	0.01	0.02
Any Black Perpetrator?	0.25	0.22
Any White Perpetrator?	0.79	0.22
Any Female Perpetrator?	0.75	0.02
Panel B: Investigator Decision Variables		
Foster Care Placement Rate	0.02	–
Ongoing Services	0.05	–
Substantiation Rate	0.32	–
Panel C: Outcomes (Within 3 months of Determination Date)		
Re-Referral Rate	0.24	–
Screened-in Re-Referral Rate	0.21	–
Substantiated Re-Referral Rate	0.08	–
Observations	4,752	
No. of Unique Children	3,488	
No. of Unique Workers	34	
No. of Unique Supervisors	4	

Notes. This table reports means for characteristics and outcomes in our analysis sample. Column 1 reports the mean among observations with non-missing values for each variable; Column 2 reports the share of observations with missing values for each variable. Active investigation indicates that the child was involved in an active investigation at the time of the focal referral. Perpetrator characteristics are constructed from the county's administrative predictor fields. For referrals involving multiple alleged perpetrators, perpetrator age is the average age across perpetrators, while the race, gender, and active-placement indicators equal one if any alleged perpetrator has the listed characteristic.

Table A2. Effect of Algorithmic Risk Information on Investigation Decisions

	(1) Foster Care	(2) Ongoing Services	(3) Substantiated
Panel A: Pooled			
Treatment	0.012**	0.015*	-0.001
	(0.006)	(0.009)	(0.017)
Control Mean	0.013	0.048	0.318
Panel B: By Complexity Category			
Treatment × Very High	0.052*	0.104***	0.148**
	(0.032)	(0.039)	(0.065)
Treatment × High	0.003	-0.008	-0.063
	(0.014)	(0.025)	(0.044)
Treatment × Medium	0.010*	0.015	-0.018
	(0.006)	(0.011)	(0.024)
Treatment × Low	0.010	-0.010	0.024
	(0.007)	(0.020)	(0.039)
Treatment × Very Low	0.000	0.016	0.005
	(0.007)	(0.014)	(0.044)
Control Mean			
Very High	0.041	0.090	0.342
High	0.024	0.097	0.393
Medium	0.008	0.035	0.322
Low	0.003	0.026	0.271
Very Low	0.000	0.004	0.220
Observations	4,752	4,752	4,752

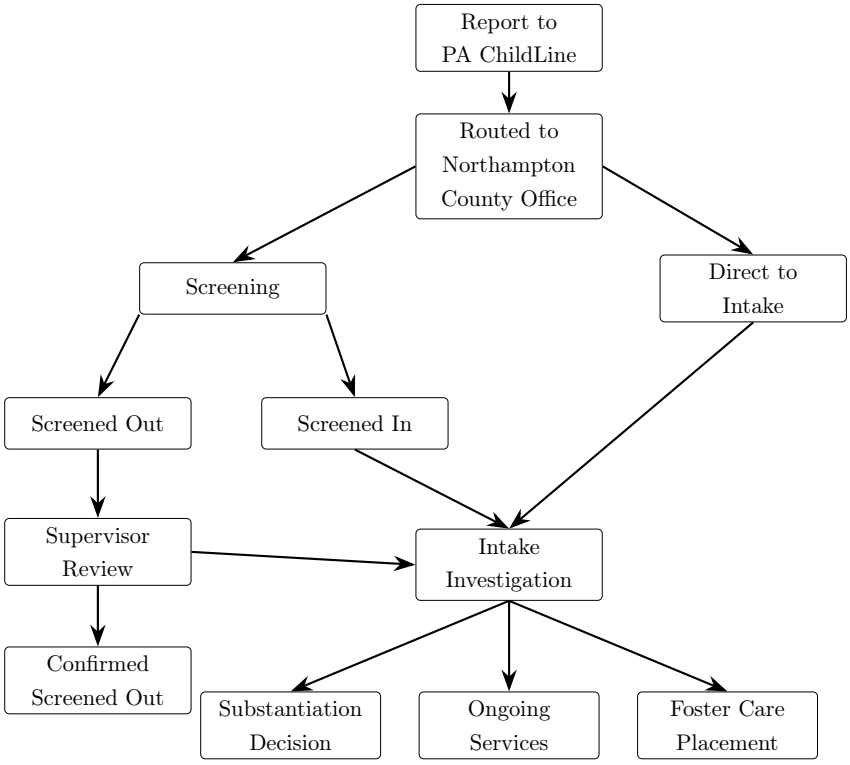
Notes. This table reports the treatment effects shown in Figure 2 in percentage points rather than standardized units. Panel A reports the pooled intent-to-treat (ITT) estimates. Panel B reports treatment effects separately by algorithmic complexity category. All regressions include referral-date fixed effects, investigator fixed effects, an indicator for Direct-to-Intake routing, and pre-treatment covariates selected using double-selection LASSO (Belloni et al., 2014). Standard errors, reported in parentheses, are clustered at the family level. * p<0.10, ** p<0.05, *** p<0.01.

Table A3. Effect of Algorithmic Risk Information on Subsequent Referrals

	(1) Re-Referral (3 Months)	(2) Re-Referral (3 Months, Excl. Focal FC)
Panel A: Pooled		
Treatment	-0.038**	-0.041**
	(0.017)	(0.017)
Control Mean	0.258	0.255
Panel B: By Complexity Category		
Treatment × Very High	-0.085	-0.093
	(0.059)	(0.062)
Treatment × High	-0.007	-0.013
	(0.037)	(0.037)
Treatment × Medium	-0.047**	-0.046*
	(0.024)	(0.024)
Treatment × Low	-0.028	-0.034
	(0.035)	(0.036)
Treatment × Very Low	-0.025	-0.022
	(0.041)	(0.040)
Control Mean		
Very High	0.346	0.341
High	0.300	0.301
Medium	0.258	0.253
Low	0.190	0.191
Very Low	0.193	0.193
Observations	4,752	4,668

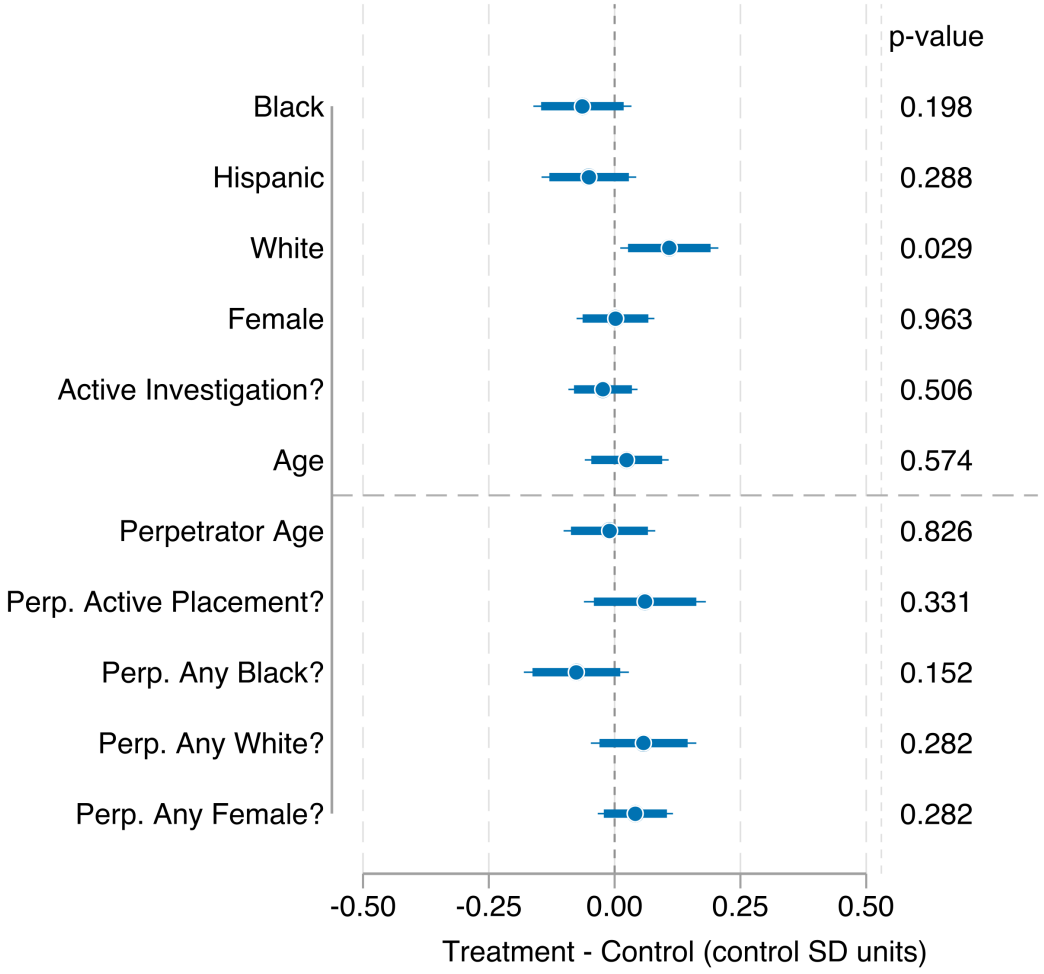
Notes. This table reports the treatment effects shown in Figure 3 in percentage points rather than standardized units. Column (1) reports estimates for the full analysis sample. Column (2) excludes referrals in which the child entered foster care during the focal investigation. Panel A reports the pooled intent-to-treat (ITT) estimates, while Panel B reports treatment effects separately by algorithmic complexity category. All regressions include referral-date fixed effects, investigator fixed effects, an indicator for Direct-to-Intake routing, and pre-treatment covariates selected using double-selection LASSO (Belloni et al., 2014). Standard errors, reported in parentheses, are clustered at the family level. * p<0.10, ** p<0.05, *** p<0.01.

Figure A1. Child-protection screening and investigation workflow in Northampton County



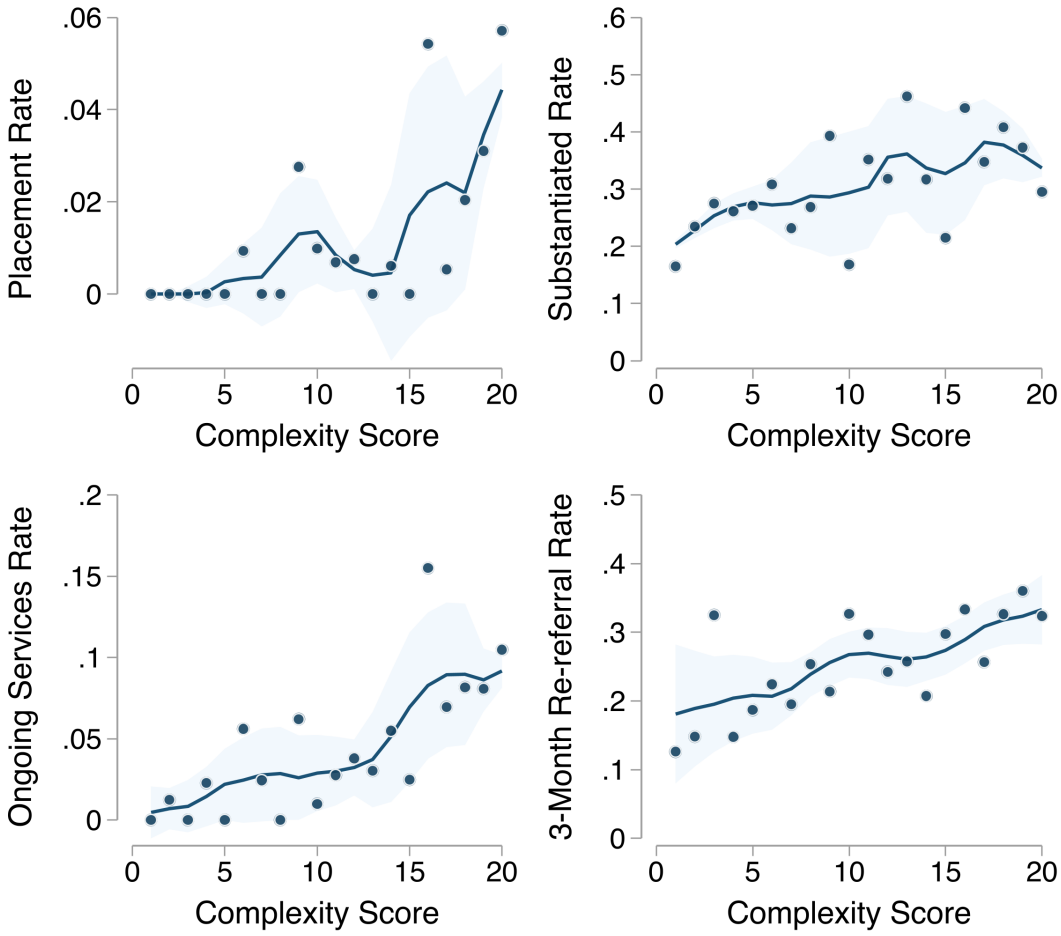
Notes. This figure summarizes the child-protection process in Northampton County. Reports of suspected maltreatment are submitted to Pennsylvania's statewide hotline and routed to the county child-welfare agency. Referrals may enter the Intake unit directly or first pass through Screening, where they are either screened out or accepted for investigation. The randomized intervention studied in this paper occurs during the Intake stage, after referrals have been accepted for investigation. Supervisors use an algorithmic risk assessment tool to support decisions regarding substantiation, ongoing services, and foster-care placement.

Figure A2. Balance Test of Treatment Assignment



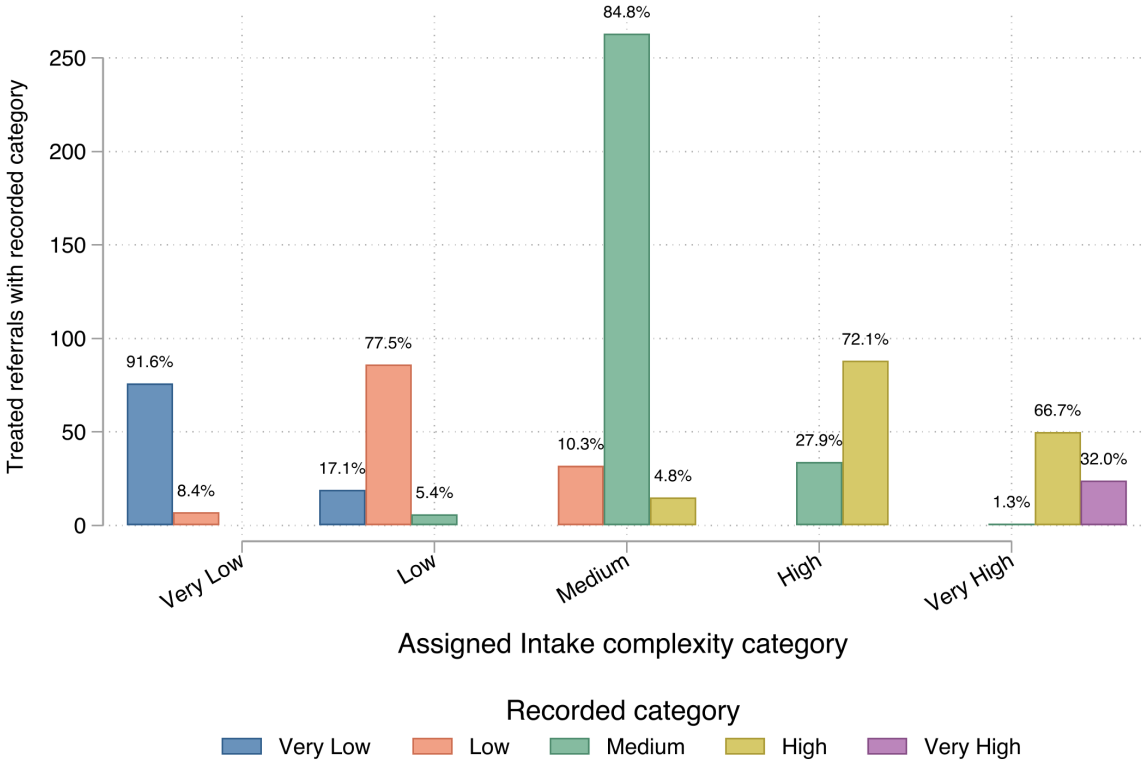
Notes. Each row reports the treatment-control difference for the indicated baseline covariate, standardized by the control-group standard deviation. Estimates are obtained from regressions of the covariate on treatment, controlling for referral-date fixed effects, investigator fixed effects, and an indicator for Direct-to-Intake routing; standard errors are clustered at the family level. Regressions are estimated using observations with non-missing values for the corresponding covariate. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. The dashed vertical line indicates no treatment-control difference, and the right-hand column reports the corresponding p-value. Active investigation indicates that the child was involved in an active investigation at the time of the focal referral. Perpetrator characteristics are constructed from the county's administrative predictor fields. For referrals involving multiple alleged perpetrators, perpetrator age is the average age across perpetrators, while the race, gender, and active-placement indicators equal one if any alleged perpetrator has the listed characteristic. The dashed horizontal line separates child- and perpetrator-level characteristics.

Figure A3. Child-Welfare Outcomes by Algorithmic Risk Category (Control Group)



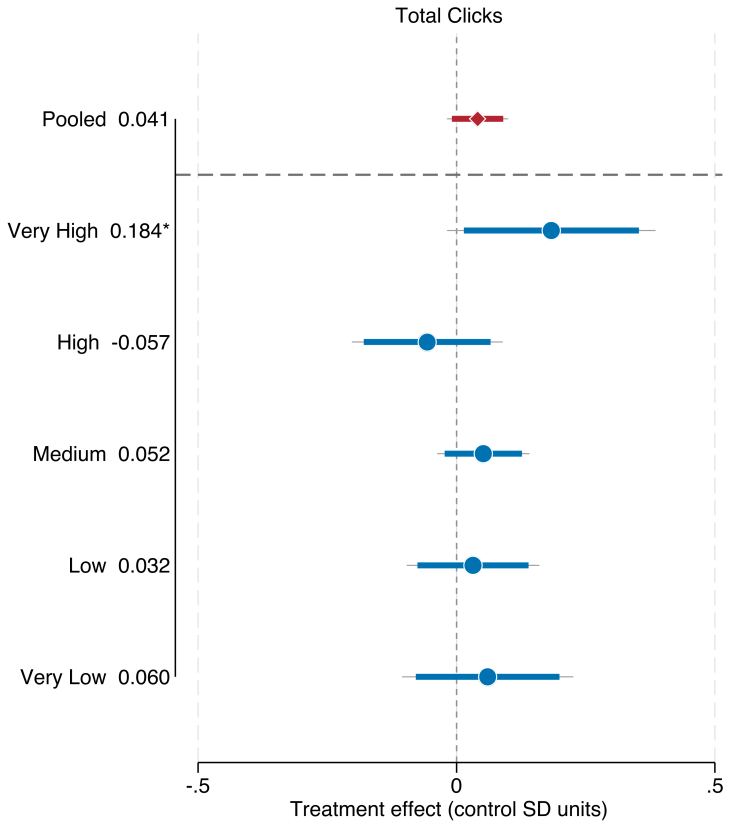
Notes. This figure plots child-welfare outcomes by the algorithmic complexity score among referrals assigned to the control group. Restricting to the control group ensures that the observed relationships are not affected by the experimental intervention. Points report mean outcome rates after collapsing referrals by the exact complexity-score value; the solid lines show LOWESS fits. Outcomes are foster-care placement, substantiation, and ongoing service provision during the focal investigation, along with any subsequent referral within three months of the investigation determination date.

Figure A4. Agreement Between Assigned and Reported Algorithmic Complexity Categories



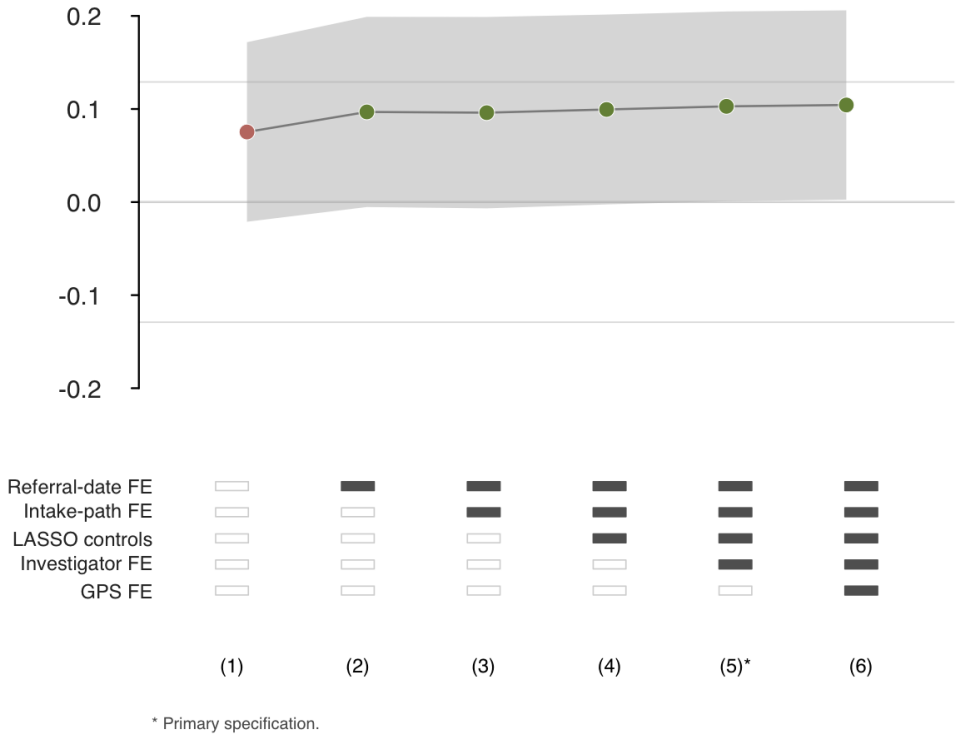
Notes. This figure compares the algorithmic complexity category assigned to each treated referral with the category reported by the assigned investigator during a routine supervisory review. Beginning in early 2024, supervisors recorded the category reported by the investigator during the second ten-day review of each investigation. The x-axis shows the assigned complexity category, while the colored bars indicate the category reported by the investigator. Percent labels are calculated within assigned complexity category. Most reported categories match the assigned category, with most discrepancies occurring between adjacent categories.

Figure A5. Effect of Algorithmic Risk Information on Supervisor Activity



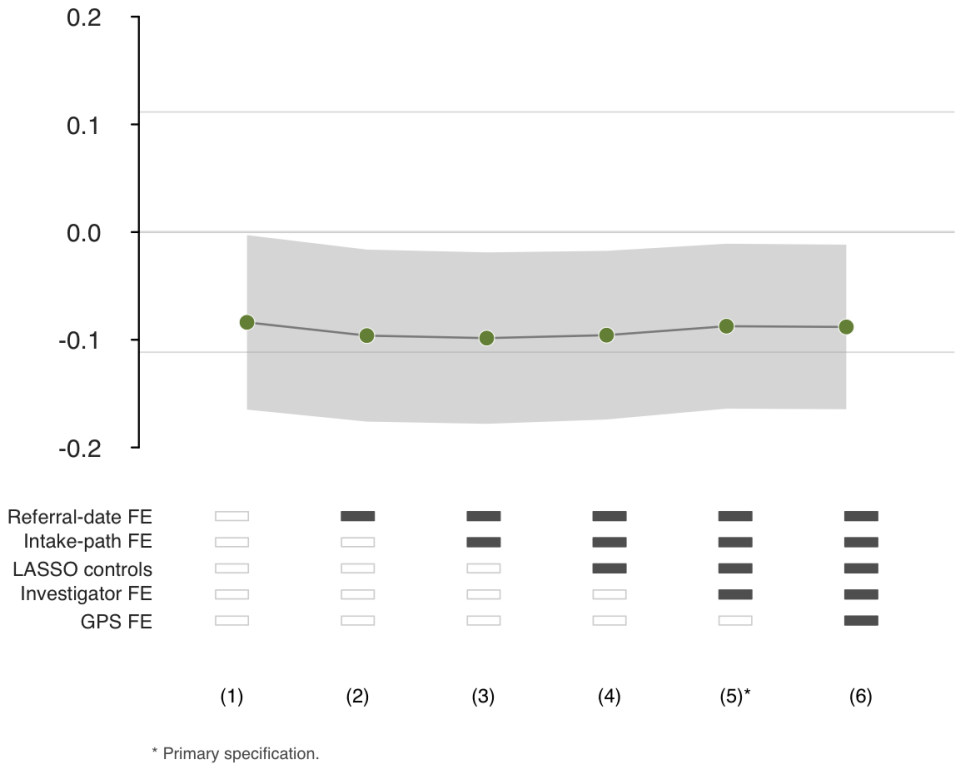
Notes. This figure reports intent-to-treat (ITT) estimates of providing supervisors with access to the algorithmic complexity score on supervisor activity recorded in the Intake Supervision Tool. The figure reports effects on the total number of recorded interactions. The red diamonds report the pooled treatment effects, while the blue circles report treatment effects separately by algorithmic complexity category (Very Low to Very High). Regressions include referral-date fixed effects, investigator fixed effects, an indicator for Direct-to-Intake routing, and pre-treatment covariates selected using double-selection LASSO (Belloni et al., 2014). Standard errors are clustered at the family level. Effects are standardized by the control-group standard deviation of the corresponding outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A6. Robustness of Foster-Care Placement Estimates to Alternative Specifications



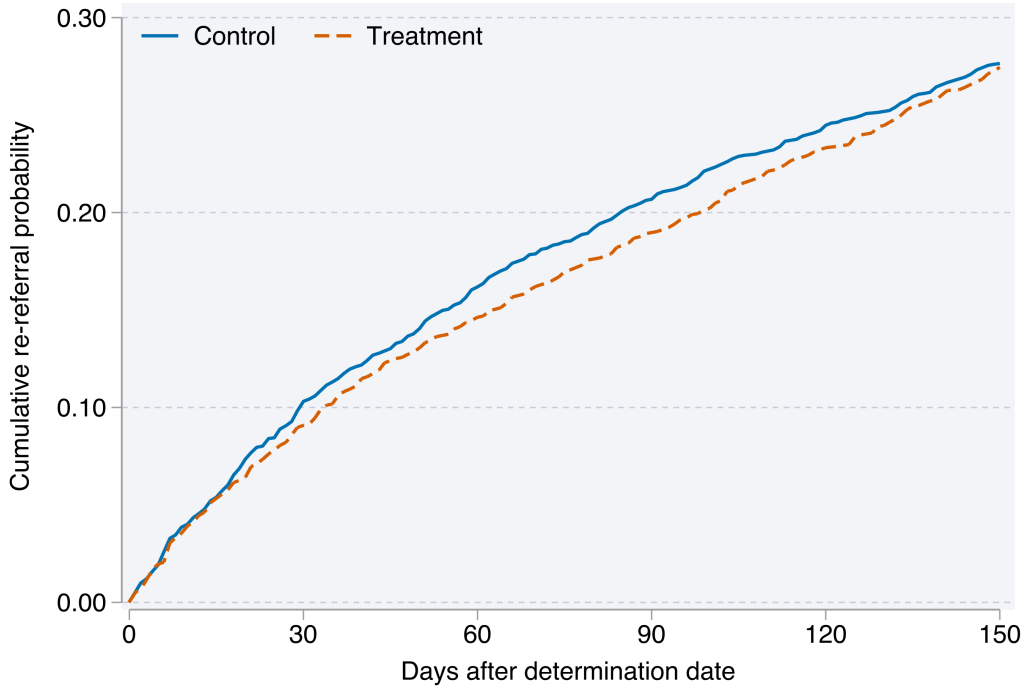
Notes. This figure reports the estimated treatment effect on foster-care placement under a sequence of alternative regression specifications. Beginning with an unadjusted specification, referral-date fixed effects, an indicator for Direct-to-Intake routing, pre-treatment covariates selected using double-selection LASSO, investigator fixed effects, and General Protective Services (GPS) fixed effects are added sequentially. Effects are standardized by the control-group standard deviation of the outcome. Gray bands denote 95% confidence intervals. Green and red points indicate estimates that are, respectively, statistically significant and not statistically significant at the 10% level.

Figure A7. Robustness of Re-Referral Estimates to Alternative Specifications



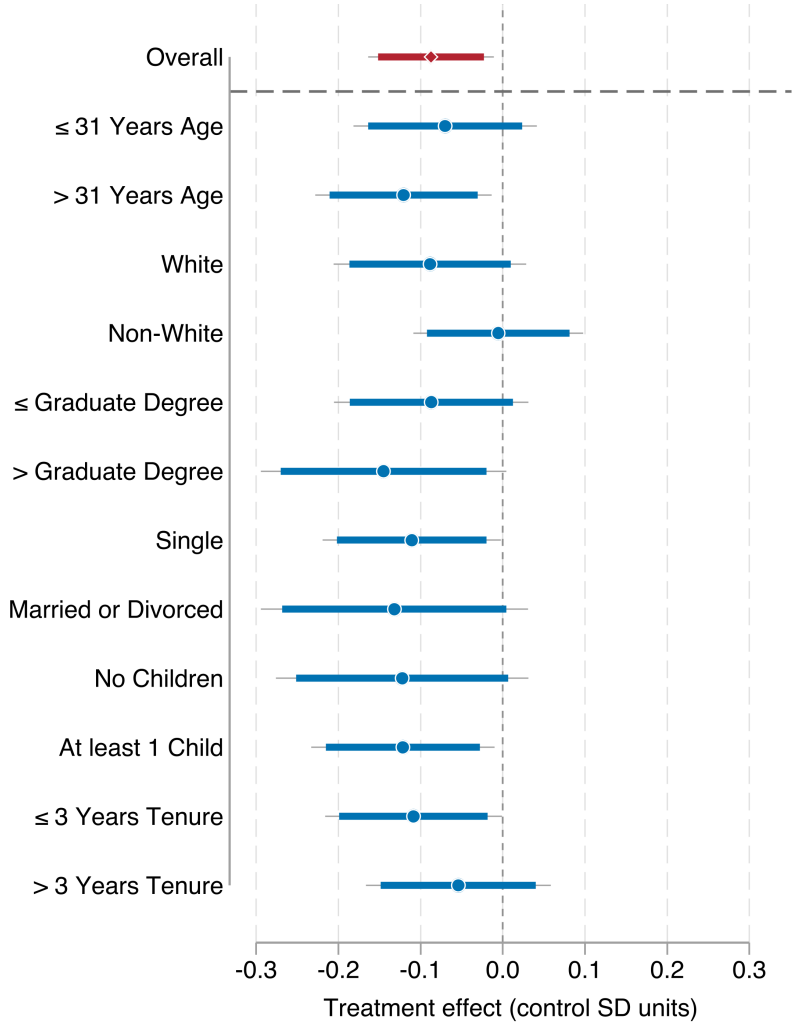
Notes. This figure reports the estimated treatment effect on re-referrals within three months under a sequence of alternative regression specifications. Beginning with an unadjusted specification, referral-date fixed effects, an indicator for Direct-to-Intake routing, pre-treatment covariates selected using double-selection LASSO, investigator fixed effects, and General Protective Services (GPS) fixed effects are added sequentially. Effects are standardized by the control-group standard deviation of the outcome. Gray bands denote 95% confidence intervals. Green and red points indicate estimates that are, respectively, statistically significant and not statistically significant at the 10% level.

Figure A8. Cumulative Probability of Re-Referral by Treatment Assignment



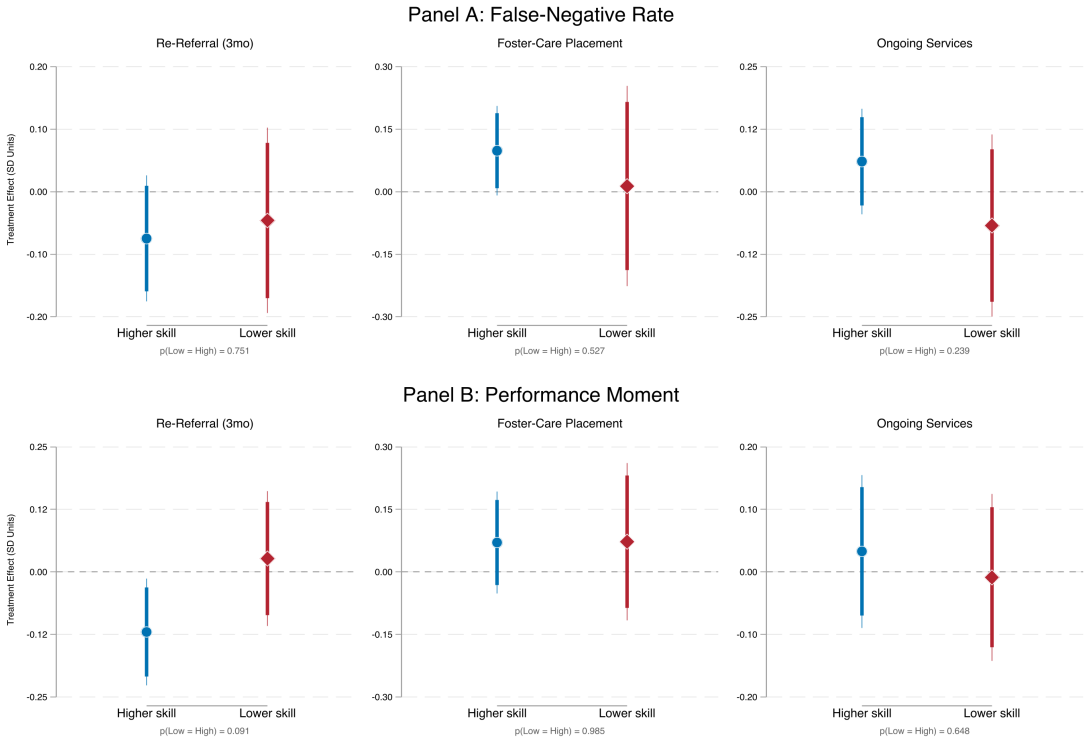
Notes. This figure plots Kaplan–Meier estimates of the cumulative probability of re-referral to the child-protection system following the determination date of the focal investigation, separately for the treatment and control groups. Children who did not experience a subsequent referral during the observation period are right-censored at the administrative data cutoff date.

Figure A9. Treatment Effects on Re-Referrals by Investigator Characteristics



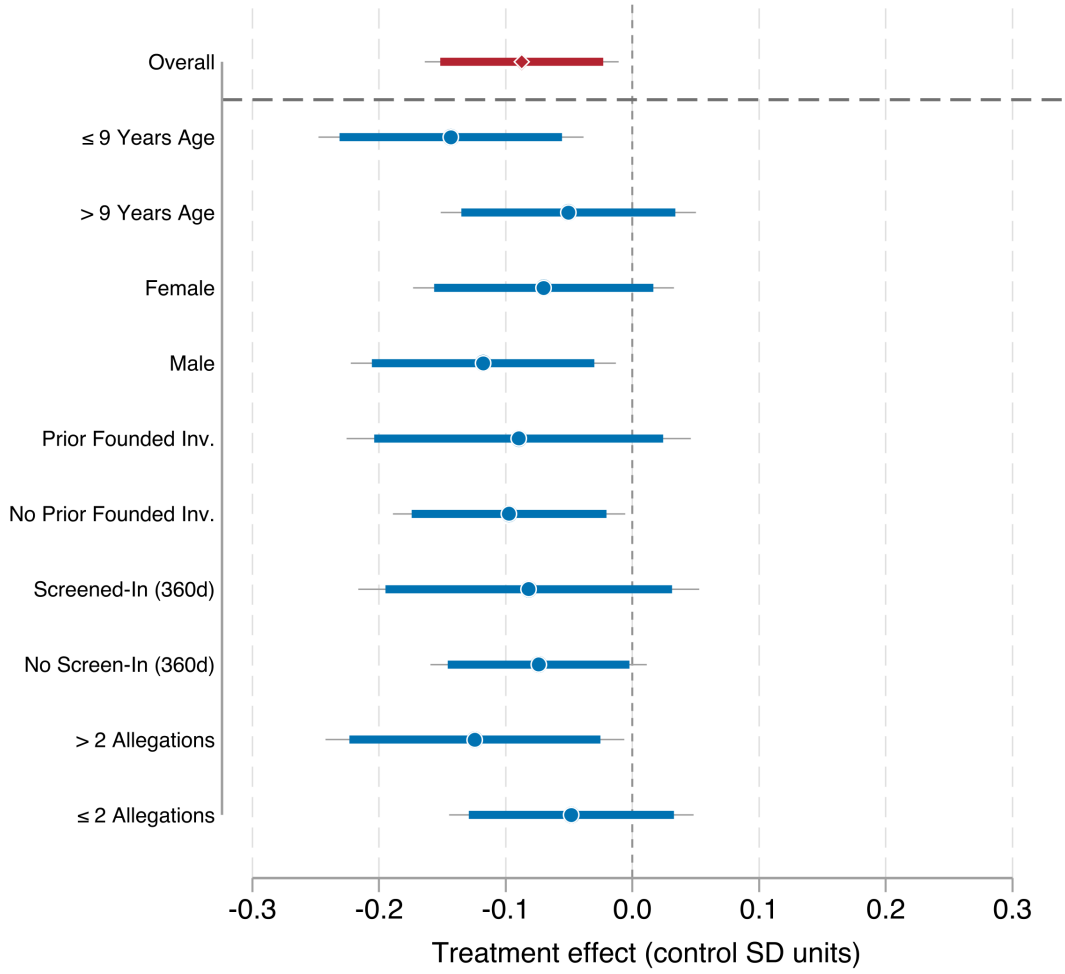
Notes. This figure reports treatment effects on re-referrals within three months of the focal investigation's determination date across investigator subgroups. Investigator characteristics are obtained from Northampton County administrative personnel records and include age, race, education, parental status, marital status, tenure, and other observable characteristics. Estimates are obtained from regressions that interact treatment assignment with the indicated investigator characteristic and include the same controls and fixed effects as the main specification. Effects are standardized by the control-group standard deviation of the outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively.

Figure A10. Treatment Effects by Pre-RCT Investigator Skill



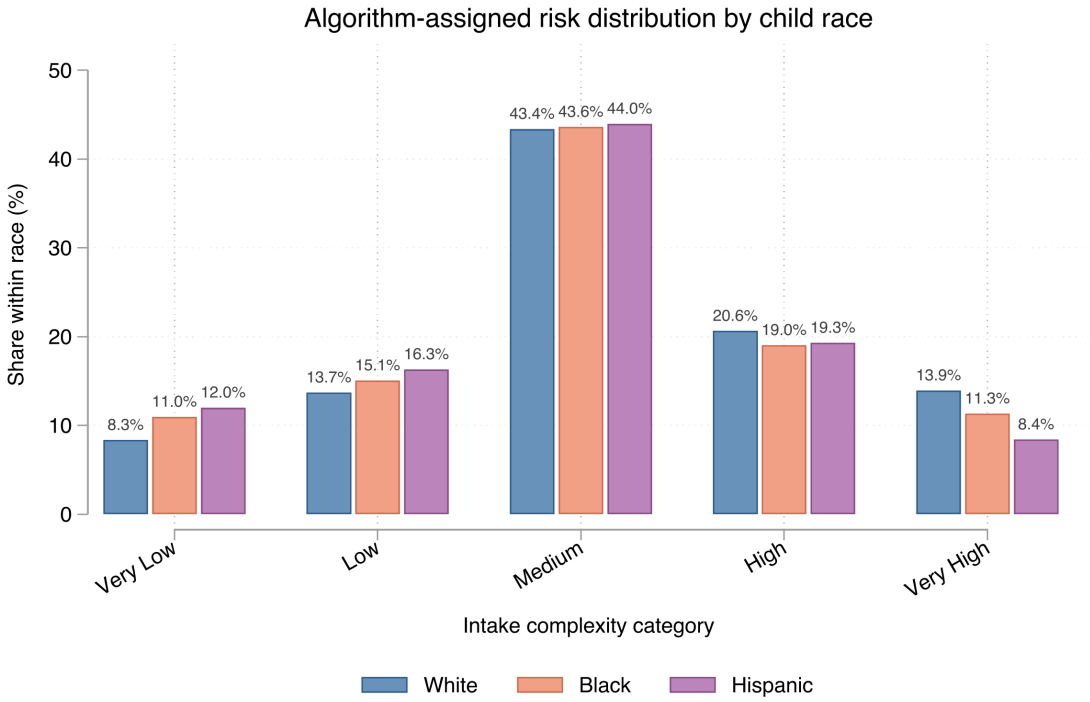
Notes. This figure reports treatment effects separately for investigators above and below the median of two pre-experimental measures of investigator performance, using the 19 investigators active both before and during the experiment ($N = 3{,}666$ child-referral observations; 1,979 family clusters). Panel A classifies investigators by their historical false-negative rate, defined as referrals closed without foster-care placement or ongoing services that were followed by a re-referral within three months. Panel B classifies investigators using the pre-experimental performance measure in Baron et al. (2024), which combines false-negative and false-positive rates into a single measure of investigator performance. Estimates are obtained from regressions that interact treatment assignment with the corresponding investigator-performance measure and include the same controls and fixed effects as the main specification. Effects are standardized by the control-group standard deviation of the corresponding outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. Standard errors are clustered at the family level. Reported p-values test the equality of treatment effects across the higher- and lower-performance groups

Figure A11. Effect of Algorithmic Risk Information on Re-Referrals by Child Characteristics



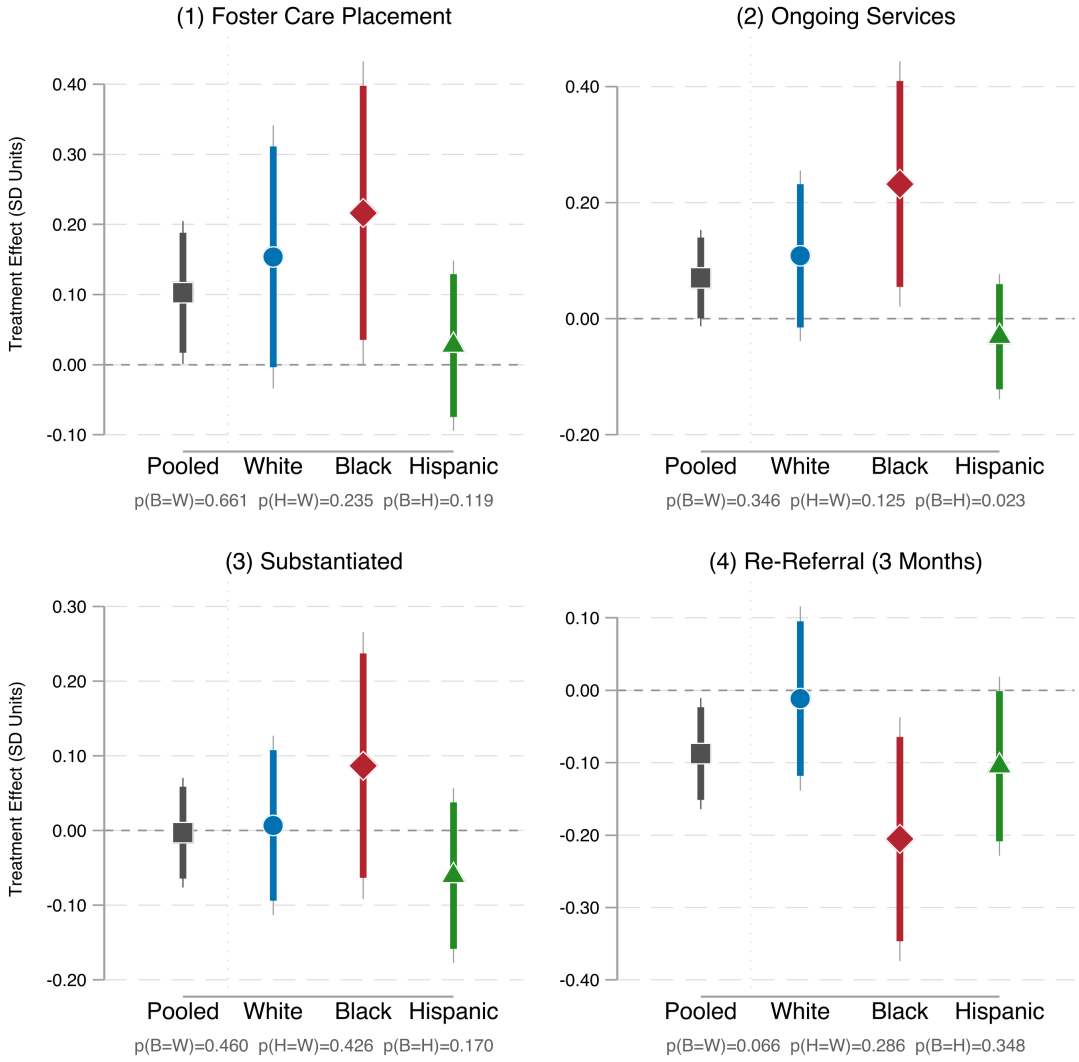
Notes. This figure reports treatment effects on re-referrals within three months of the focal investigation's determination date for subgroups defined by child characteristics measured at the time of referral. *Prior Founded Inv.* indicates that the child had at least one founded or validated investigation before the focal referral. *Screened-In (360d)* indicates that the child had at least one referral screened in for investigation during the 360 days preceding the focal referral. > 2 *Allegations* indicates that the focal referral contained more than two allegations (the sample median). Race is omitted from this figure and examined separately in Figure A13. Estimates are obtained using the same specification as the main analysis. Effects are standardized by the control-group standard deviation of the outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A12. Algorithm-Assigned Risk Categories by Child Race



Notes. This figure plots the distribution of algorithm-assigned Intake complexity categories for White, Black, and Hispanic children in the analysis sample. Bars report the share of children within each race assigned to each complexity category; shares sum to one within race.

Figure A13. Treatment Effect Heterogeneity by Child Race and Ethnicity



Notes. This figure reports treatment effects separately by child race and ethnicity. The gray square labeled "Pooled" reports the overall treatment effect from the main specification, while the remaining estimates are obtained from a single regression that fully interacts treatment assignment with race and ethnicity indicators. Estimates are obtained using the same controls and fixed effects as the main specification. Effects are standardized by the control-group standard deviation of the corresponding outcome. Thick and thin spikes denote 90% and 95% confidence intervals, respectively. Reported p-values test the equality of treatment effects across Black (B), White (W), and Hispanic (H) children.

B Additional Institutional Details

B.1 Child welfare investigation process in Northampton County

Reports of suspected maltreatment are initially received by the ChildLine, Pennsylvania's statewide child-maltreatment hotline and referral system. Referrals involving children from Northampton County are routed to the Northampton Children, Youth, and Families (CYF) agency. Upon reaching the Northampton CYF office, referrals are assigned to one of two paths: Screening or Direct-to-Intake.

If a referral is sent to the Screening unit, screening staff first review the referral and gather any necessary additional information before deciding whether the referral should be screened out or subsequently sent to Intake. Otherwise, the referral is sent directly to Intake, which is more common when a family has substantial recent child-welfare involvement. Approximately 40 percent of referrals that ultimately reach Intake first pass through Screening, while the remaining 60 percent are routed directly to Intake. Our experiment is focused on referrals that reached the Intake phase, regardless of whether they were routed through screening first or sent directly to Intake.

The Screening process is a separate unit from the Intake process studied in this paper. In this county, screeners have up to 30 days to decide whether to send the report to Intake or screen out the report, and will often make contact with the family to assist their decision. Screening workers do not have access to the Intake Supervision Tool, and thus do not observe the intake complexity score.

For referrals that pass through Screening, the screener makes an initial screen-in or screen-out recommendation. This is then sent to the screening supervisor, who has authority to override these recommendations. The screening supervisor has access to a separate Northampton Decision Aid Tool (NDAT) risk prediction score when making override decisions. Only the screening supervisor has access to this score, and they do not share it with screening workers. For screened-in calls that are sent to Intake, the NDAT score is also not shared with Intake workers or supervisors; Screening and Intake are separate units which use different case-management systems in this county, and so Intake supervisors did not have access to the NDAT screening interface nor the NDAT risk score. The screening supervisor can use the NDAT score to inform their override decision. We do not observe screening override decisions in our data, but interviews with the county Screening supervisor suggest most screening override decisions result in screening-in high-risk referrals that were initially screened-out. Any use of the separate NDAT risk score during the Screening process is not part of the

randomized treatment studied in this paper. The randomized intervention begins only after a referral enters the Intake phase. After the screening supervisor makes the final screen-in or screen-out decision, screened-in referrals are sent to the Intake phase while CYF involvement ends for screened-out referrals.

Once a referral enters Intake, a supervisor assigns it to an Intake worker for assessment. In cases involving a recent closure or open case, assignment may follow a bounce-back rule that returns the referral to the worker who recently handled the family. Outside these cases, assignment follows a rotational system. Intake workers then conduct the assessment, gather additional information, and work with supervisors to determine the agency's response. Intake workers determine whether the referral should be dismissed, substantiated, ongoing services assigned, or whether they should petition the courts for foster care placement.

B.2 Risk prediction algorithm construction

We next describe the construction of the Intake algorithmic complexity score. The complexity score is a predicted probability that a child experiences an adverse child-welfare outcome following a referral: foster care placement within one year of the referral. The model was estimated on historical Northampton County administrative records from CAPS, the county's child-welfare information system.

The model uses a rich set of predictors describing the child, the alleged perpetrator(s), other household adults, and the family's prior involvement with the child-welfare system. These include child demographic characteristics and age; the allegation types associated with the referral; counts of prior referrals, screen-ins, investigations, and substantiations; prior placements and removals (e.g., kinship, non-kinship, court-ordered, emergency, and voluntary); prior service receipt; and analogous histories for alleged perpetrators and parents. Child race and ethnicity were not used as model inputs. The model was estimated using a gradient-boosting decision tree (XGBoost).

For interpretability, the model's continuous predicted risk was not displayed directly to supervisors. Instead, predictions were discretized into five ordered complexity categories – Very Low, Low, Medium, High, and Very High – which are the labels shown in the Intake Supervision Tool and used throughout our analysis. The model generates separate risk predictions for each child involved in the referral. For referrals with multiple children, the displayed complexity category is the maximum category across children in that referral. Thus, throughout our analysis, we use the maximum category risk across children in the referral as our risk category.

B.3 Treatment implementation and randomization

The randomized evaluation was implemented within Northampton County's Intake Supervision Tool. The experiment began on December 18, 2023 and continued through February 15, 2025. During this period, referrals entering the Intake Supervision Tool were assigned to either a treatment condition, in which the algorithmic complexity category was visible to Intake supervisors, or a control condition, in which the complexity category was hidden. After the evaluation period ended, the complexity category was made visible for all referrals.

Randomization occurred only after a referral entered the Intake stage. For referrals sent directly to Intake, the randomizer is applied when the referral enters the Intake Supervision Tool. For referrals that first pass through Screening, the randomizer is applied only after the referral is screened in and enters the Intake Supervision Tool. Thus, for referrals sent first through the Screening phase, treatment assignment occurs after the county screening decision. The randomized treatment therefore did not affect ChildLine routing, county direct-to-Intake routing, or the county screening decision. Instead, the experiment randomized whether Intake supervisors could observe the complexity category after the referral had entered Intake.

Assignment was implemented at the family level. When a family first appeared in the Intake Supervision Tool during the evaluation period, the tool assigned the family to treatment or control. This assignment then persisted for all subsequent referrals involving the same family during the study period. Families with open Intake referrals at the start of the experiment were randomized into treatment or control at the start of the experiment. Family-level assignment avoids situations in which supervisors could observe the complexity category for one referral involving a family but not for another referral involving the same family. In the analysis, all child-referral observations associated with a treated family are assigned to treatment, and all child-referral observations associated with a control family are assigned to control.

The treatment changed only the visibility of the algorithmic complexity category in the Intake Supervision tool. For treatment referrals, the Intake Supervision Tool displayed the category to supervisors on the overview screen. For control referrals, the corresponding field showed "Score not shown." Treatment and control referrals otherwise appeared in the same Intake Supervision Tool, and supervisors retained access to the same underlying administrative records through the county's case-management system. The complexity category was displayed to supervisors through the case-management system, and supervisors were instructed to communicate the category to assigned investigators when it was visible.

B.4 Analysis sample construction and outcomes

Our analysis uses de-identified administrative records from Northampton County's child-welfare case-management system. These records include referral- and incident-level information (including report dates, determination dates, and assigned investigators), allegation-level records describing the allegations and investigation outcomes, family-level treatment assignment from the randomized evaluation, and complete child-level placement histories. Person identifiers allow us to link children across referrals and measure subsequent child-protection involvement over time.

We construct our analysis sample by restricting attention to referrals that entered the Intake stage during the experimental period. Treatment is assigned at the family level and remains fixed across all referrals involving the same family throughout the study. We begin with 6,884 child-referral observations with a report date after the start of the experiment and a determination date before the end of the experimental period. We then restrict the sample to referrals that reached the Intake stage, dropping referrals never assigned to an Intake caseworker ($N = 1,715$) and dropping observations missing an Intake risk score ($N = 289$).[5] We additionally exclude 107 observations affected by a temporary software display issue (described below). Finally, because our primary specifications include referral-date, Direct-to-Intake routing, and investigator fixed effects, we drop 21 singleton observations. The resulting analysis sample contains 4,752 child-referral observations, corresponding to 3,488 unique children, 3,661 referrals, 2,422 families, and 34 Intake investigators.

A temporary software error in the Intake Supervision Tool caused the algorithmic complexity category to be displayed without regard to randomized assignment for a small set of newly appearing families. The error occurred between March 26 and April 5, 2024, and affected only families that had not yet been assigned a randomization status; previously randomized families were unaffected. Northampton County identified the 67 affected families (37 treatment and 30 control), and we excluded all observations associated with these families from the analysis to preserve the integrity of the randomized evaluation.

We study three key decisions made during the focal investigation: substantiation, assignment to ongoing services, and foster-care placement. We define a referral as substantiated if it received a determination of "Validated," "Founded," or "Indicated." We directly observe whether a referral resulted in assignment to ongoing services following the Intake investigation. Foster-care placement requires additional construction because the administrative data record placement histories rather than the investigation that led to the placement. We observe

[5] An Intake risk score is not generated if there is no victim, allegation or perpetrator identified.

complete placement histories for all children in Northampton County and classify a referral as resulting in foster-care placement if either (i) the child entered foster care within 90 days of the focal investigation's determination date, or (ii) the child entered foster care on or after the determination date but before the child's next recorded referral. This definition is designed to capture placements resulting from the focal investigation while allowing for the fact that court approval and the execution of a removal may occur after the investigation has concluded.

Our primary downstream outcome is whether a child experienced another referral to the child-protection system within three months of the focal investigation's determination date. Because these referrals are generated primarily by external reporters rather than by the child-welfare agency itself, they may provide an independent measure of subsequent concern about child well-being. We also examine the nature of subsequent child-protection involvement by constructing indicators for whether a subsequent referral was screened in for investigation (using the same Intake definition as in our sample restrictions), substantiated, or resulted in a subsequent foster-care placement. Each outcome is measured over the corresponding follow-up window.

B.5 Empirical specifications and control selection

Our primary estimand is the intent-to-treat (ITT) effect of making the algorithmic complexity category visible to supervisors. The unit of observation is a child-referral. We estimate both the pooled specification in Equation (1) and the heterogeneous-effects specification in Equation (2), which interacts treatment with the five algorithmic complexity categories. Both specifications include referral-date fixed effects, an indicator for whether the referral entered Intake directly or first passed through Screening, investigator fixed effects, and a set of pre-treatment covariates selected as described below.[6] Because treatment is randomly assigned, the treatment effect is identified without additional controls or fixed effects. We include these variables solely to improve precision. Finally, because treatment is assigned at the family level, we cluster standard errors at the family level throughout to account for within-family correlation across multiple child-referral observations.

We select pre-treatment controls separately for each outcome using the double-selection LASSO procedure of Belloni et al. (2014). The candidate set consists of child and alleged-perpetrator characteristics measured prior to treatment, including child race and ethnicity,

[6]Referral-date fixed effects absorb common temporal shocks over the study period. The Direct-to-Intake indicator accounts for systematic differences between referrals routed directly to Intake and those that first passed through Screening, while investigator fixed effects absorb persistent differences in investigator behavior. Figures A6 and A7 show that our results are robust to alternative specification choices.

sex, age, prior active investigation involvement, and alleged-perpetrator age, sex, race, and active-placement status. For covariates with missing values, we set missing observations to zero and include a corresponding missing-value indicator in the candidate set. Following Belloni et al. (2014), we select covariates that predict either the outcome or treatment assignment using the plug-in penalty, with the selection step clustered at the family level. The union of the selected covariates is then included linearly in the estimating regression alongside the fixed effects. Because selection is performed separately for each outcome, the resulting set of controls may differ across specifications.

To facilitate comparisons across outcomes with different base rates, we report treatment effects in units of the control-group standard deviation of each outcome. Specifically, we divide each estimated coefficient and its standard error by the control-group standard deviation of the corresponding outcome. In the risk-category heterogeneity analysis, all estimates are standardized using the same full-sample control-group standard deviation, allowing direct comparisons across complexity categories. For ease of interpretation, Tables A2 and A3 also report the corresponding estimates in percentage-point units.

B.6 Measuring Investigator Skill

Figure A10 classifies investigators using two measures of pre-experimental investigator performance, both constructed from cases handled before the experiment began. We use historical Northampton County administrative records for referrals reported before the December 18, 2023 randomization start, restricting attention to the 19 investigators active both before and during the experiment. Both measures exploit the rotational assignment of Intake cases to investigators (Appendix B.1). Conditional on referral timing and routing, assignment is as good as random, allowing systematic differences in outcomes across investigators to be attributed to investigator performance rather than differences in the cases they receive.

Panel A classifies investigators according to their historical false-negative rate. We define a false negative as a referral closed without foster-care placement or ongoing services that is followed by a re-referral within three months. We estimate investigator-specific false-negative rates by regressing this indicator on investigator indicators, report-month fixed effects, and an indicator for Direct-to-Intake routing. The estimated investigator fixed effects provide strata-adjusted measures of false-negative rates, with larger values indicating a greater tendency to close cases that subsequently return to the child-protection system. Throughout this exercise, subsequent re-referral within three months serves as our measure of downstream child-protection concern, consistent with the primary outcome used in the experimental analysis.

Panel B instead uses the pre-experimental performance measure in Baron et al. (2024), which combines information on investigators' false-negative and false-positive rates. A false positive is an intervention – foster-care placement or assignment to ongoing services – for a child who would not otherwise have subsequently returned to the child-protection system. Because these counterfactual outcomes are unobserved, investigator-specific false-positive rates cannot be directly estimated. However, Baron et al. (2024) show that under the random assignment of investigators, differences in false-positive rates across investigators are identified.

Following Baron et al. (2024), we combine the strata-adjusted false-negative rates from Panel A with analogous strata-adjusted intervention rates to construct a summary measure of investigator performance. The two components are weighted according to the assumed social costs of false positives and false negatives, assigning a weight of 1 to false positives and 0.25 to false negatives. Intuitively, this weighting places greater emphasis on avoiding unnecessary interventions while still penalizing failures to intervene when children subsequently return to the child-protection system. The resulting measure ranks investigators according to their pre-experimental performance.

C Pre-Registration Differences

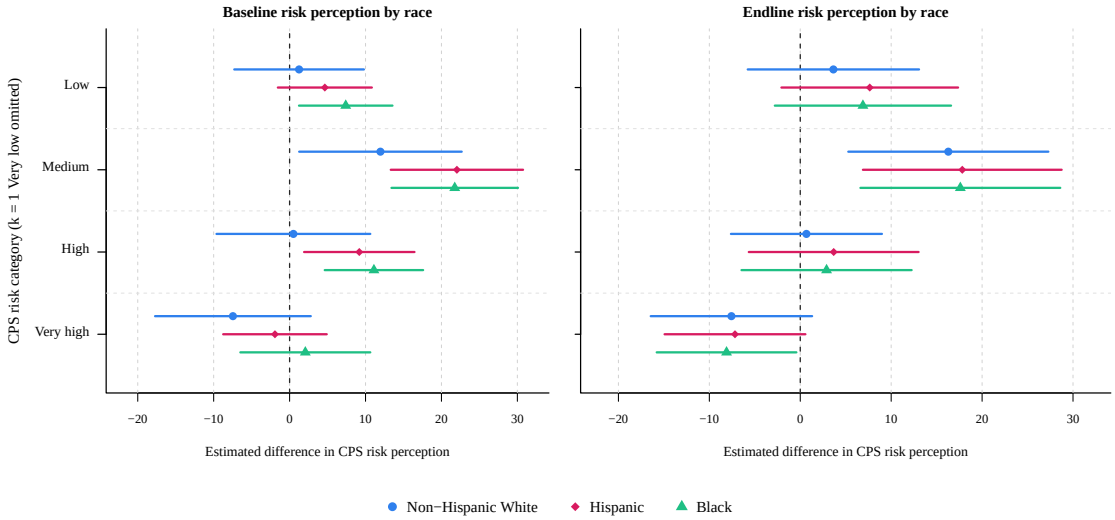
This project was pre-registered in the AEA RCT Registry (#0012623) before the intervention began. In this appendix, we describe the main differences between the pre-registration and the analysis presented in the paper.

First, the pre-registration listed hospitalization records as a secondary outcome, conditional on data availability. These data were ultimately not made available to the research team, and we therefore do not include hospitalization outcomes in the analysis. The paper instead focuses on the child-welfare administrative outcomes available from the agency, including re-referrals, subsequent substantiations, subsequent foster-care placements, and agency decisions during the focal investigation.

Second, the pre-registration included worker beliefs as an outcome. We collected baseline and endline survey data on worker beliefs about the distribution of child-welfare risk across racial and ethnic groups. In the survey, workers were asked, out of 100 children from each racial or ethnic group, how many they expected to fall into each algorithmic risk category. Because the number of caseworkers is small, and because the composition of respondents changed between the baseline and endline surveys, we do not use these data to estimate individual-level belief updating. Instead, we present descriptive office-level evidence on whether perceived racial differences in risk were present at baseline and whether they remained at endline.

Figure A14 reports these results. At baseline, workers' perceived risk distributions differed by child race and ethnicity, indicating that there were racial differences in beliefs about CPS risk at the office level. By endline, these differences were smaller and less systematic. We interpret this evidence descriptively. Since the survey respondent composition changed over time and the study was not powered to estimate worker-level belief updating, the figure should not be interpreted as a causal estimate of individual belief change. Rather, it documents that the belief outcome specified in the pre-registration was measured, but the data are not detailed or large enough to support the individual-level belief updating exercise cleanly.

Figure A14. Worker Beliefs About Child-Protection Risk by Race and Ethnicity



Notes. This figure summarizes baseline and endline survey responses on workers' perceptions of child-protection risk across racial and ethnic groups. In each survey, respondents were asked to allocate 100 hypothetical children within each racial or ethnic group across the five algorithmic risk categories. Points report estimated differences in the perceived share of children assigned to each risk category, with horizontal bars denoting 95% confidence intervals. The Very Low risk category is omitted because the reported shares across categories sum to 100 within each racial or ethnic group. Because the number of respondents is small and the composition of respondents differs between the baseline and endline surveys, the figure should be interpreted as descriptive office-level evidence rather than as evidence of individual-level belief updating.