A Case for Humans-in-the-Loop: Decisions in the Presence of Misestimated Algorithmic Scores

Riccardo Fogliato*
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Maria De-Arteaga*
McCombs School of Business
University of Texas at Austin
Austin, TX, 78712

Alexandra Chouldechova
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The increased use of machine learning to assist with decision-making in high-stakes domains has been met with both enthusiasm and concern. One source of ongoing debate is the effect and value of decision makers' discretionary power to override algorithmic recommendations. In this paper, we study the adoption of an algorithmic tool used to help with decisions in child maltreatment hotline screenings. By taking advantage of an implementation glitch, we investigate *corrective overrides*: whether decision makers are more likely to override algorithmic recommendations when the tool misestimates the risk score shown to call workers. We find that, after the deployment of the tool, decisions became better aligned with algorithmic assessments, but human adherence to the tool's recommendation was less likely when the displayed score was misestimated as a result of the glitch. Then, analyzing the effect of adoption and overrides on racial and socioeconomic disparities, we find that the deployment of the tool did not affect disparities with respect to the pre-deployment period. We also observe that

* Indicates equal contribution.

Corresponding authors: Riccardo Fogliato, rfogliat@andrew.cmu.edu;
Maria De-Arteaga, dearteaga@mccombs.utexas.edu.

the disparities resulting from algorithmic-informed decisions were substantially smaller than those associated with the algorithm in isolation. Together, these results make a case for the value of humans in-the-loop, showing that in high-stakes contexts, human discretionary power can mitigate the risks of algorithmic errors and reduce disparities.

1 Introduction

Algorithmic decision support tools are increasingly being incorporated into expert-driven decision-making pipelines across domains, such as inventory management (Kesavan and Kushwaha 2020), retail pricing (Phillips et al. 2015), healthcare (Caruana et al. 2015, Ganju et al. 2020), criminal justice (Kleinberg et al. 2017), education (Smith et al. 2012), and public services (Chouldechova et al. 2018). Many of these tools rely on statistical models, ranging in complexity from simple regression to deep neural networks, which distill available information about a given instance into a risk score or label reflecting the likelihood of one or more outcomes of interest. For example, a model deployed in healthcare may estimate the risk of complications from pneumonia for a patient (Caruana et al. 2015), and in the child welfare system it may estimate the risk of out-of-home placement for a child (Chouldechova et al. 2018). Although such systems have long been used in industries such as finance (Baesens et al. 2003), the uptake of machine learning has significantly expanded the application domains where these tools can be found and has brought renewed interest and scrutiny into machine learning's costs and benefits. Bolstered by decades of research showing that statistical models can outperform human experts on prediction tasks (Meehl 1954, Dawes et al. 1989, Grove et al. 2000, Ægisdóttir et al. 2006, Kleinberg et al. 2017), there is widespread optimism that these tools can increase the quality of human decisions. This optimism is tempered by evidence that human-machine complementarity can be an elusive goal because humans provided with machine predictions can fall prey both to automation bias, resulting in overreliance (Marten et al. 2004) and algorithm aversion, which induces underreliance (Dietvorst et al. 2015). Humans may also uptake information in ways that lead to increased disparities in decision outcomes (Skeem et al. 2019). Such findings raise critical questions about the role of humans in the loop. For example, after conducting a field-experiment examining the profit implications of providing discretionary power to merchants in the automotive industry, Kesavan and Kushwaha (2020) suggest that "As data-driven decision-making (DDD) emphasizes data over intuition, organizations need to reevaluate the value of providing discretionary power to managers to override DDD tools in their organizations" (Kesavan and Kushwaha 2020, p.1). And even when the desire for a human in the loop is not in question, how to design a system that complements human abilities is not well understood.

In this work, we study a child welfare decision-making context in which hotline call workers are tasked with deciding whether a call concerning potential child neglect or maltreatment should

be screened in for investigation. In 2016, Allegheny County (Pennsylvania, US) deployed a risk assessment tool, called the Allegheny Family Screening Tool (AFST), to support call workers' screening decisions. The tool uses multi-system administrative data to assess the likelihood that children in the case will experience adverse child welfare events in the near future.

Some time after the tool was deployed, the county discovered that a technical glitch caused a subset of model inputs to be incorrectly calculated in real time.[1] This glitch, in turn, resulted in the display of misestimated risk scores in some cases. Although the misestimation was often mild and the shown score generally provided reasonable risk information, the glitch gives us a rare opportunity to investigate real-world, algorithm-assisted decision making in the presence of misestimated scores.

In this work, we focus on three important questions in this area: (1) Did the tool's deployment affect call workers' decisions? (2) Did call workers treat misestimated scores differently, or did they indiscriminately adhere to recommendations? (3) How did the tool's deployment and call workers' overrides affect racial and socioeconomic disparities? We begin our analysis by determining that there was a marked change in workers' screening decisions in the post-deployment period. After establishing that an overall change in behavior did occur, we then investigate the extent to which call workers deviated from recommendations based on a misestimated risk score. We find that, although workers did exhibit a slight tendency of automation bias in cases where the tool required a supervisor's approval for screen-out (i.e. deciding not to screen in a case for investigation), they were able to appropriately override the tool's recommendations in the majority of the cases. In particular, we find that they avoided errors of omission when the score underestimated the risk. We also investigate questions of potential disparities in adherence to recommendations across racial and socioeconomic groups. Our findings show that the deployment of the tool neither significantly mitigated nor exacerbated disparities with respect to the pre-deployment period, and workers did not exhibit differential overrides across groups. Furthermore, we observe that humans in the loop mitigated disparities when compared to the algorithm in isolation, and the results indicate that this outcome was driven by human decisions that relied on factors other than the risk score alone.

2 Background

Prior research has sought to answer the questions of *whether* and *how* the deployment of algorithmic tools for decision support affects the quality of the resulting decisions. Many researchers and practitioners have advocated for the adoption of these tools on the basis of their superior predictive

[1] Before proceeding, we pause to note that these types of technical issues are not uncommon. What is uncommon is for organizations to choose to be transparent about their occurrence. We respect and value Allegheny County for its transparency and hope that this approach becomes the norm in the deployment of algorithmic systems in sensitive societal domains.

accuracy, but findings are mixed on whether integrating these predictive tools into decision-making processes significantly improves the quality of the resulting decisions. In this regard, how these tools should be integrated into decision-making pipelines is also a subject of debate, both in terms of algorithmic design and discretionary power afforded to the human.

In the operations management literature, the effectiveness of algorithmic decision support in the contexts of pricing and inventory management has attracted significant attention. Empirical evidence has shown that human overrides of centralized autonomous systems may be detrimental to overall performance (Phillips et al. 2015, Van Donselaar et al. 2010, Karlinsky-Shichor and Netzer 2019). However, the effect may be heterogeneous for different subsets of instances (Kesavan and Kushwaha 2020), and human discretion can improve performance when information is available that the algorithm ignores. For example, in the context of inventory replenishment, store managers may optimize for relevant objectives that the algorithm's optimization objective does not capture. Van Donselaar et al. (2010) find that managers can complement the algorithm by accounting for factors such as costs of in-store handling of products and potential to improve sales by maintaining a better stock. In addition to optimizing for *outcomes* that are unobserved by the algorithm, humans also may account for unobserved *attributes* that can lead to higher profits–for example, in the context of B2B personalized pricing, when the pricing involves individualized quotes with unique or complex characteristics (Karlinsky-Shichor and Netzer 2019). Importantly, there is evidence that whether inventory support can be enhanced by the presence of algorithmic tools depends on how the tools' recommendations are communicated (Acimovic et al. 2020).

Needless to say, decision making in the context of child maltreatment hotlines differs from pricing and inventory management in significant ways. Most notably, the high-stakes nature of the task implies different costs associated with different types of errors. Recent work has audited risk assessment models used in this context (Chouldechova et al. 2018), studied the public perceptions of these systems (Brown et al. 2019), and raised concerns over the risk of automating historical patterns of discrimination (Eubanks 2018). For a review of algorithms used within the U.S. Child Welfare System, we refer the reader to Saxena et al. (2020). To the best of our knowledge, we are the first to study human adherence to algorithmic recommendations in this context.[2] A related domain in which algorithmic overrides have been studied is the criminal justice system, where risk assessment models that predict future recidivism are increasingly deployed. In both domains, the decision tasks are high stakes and concern the government's allocation of goods and burdens. In the pretrial context, the main driver behind the rapid adoption of these tools was the hope of sharp

[2]The present paper builds on an earlier paper by the authors (De-Arteaga et al. 2020), which appeared in conference proceedings. We offer two novel contributions in this manuscript. First, our empirical strategy in this paper is more robust and relies on regression analyses. Second, the research questions that we investigate now include an analysis of overrides across racial and socioeconomic groups, which is key to understanding potential disparities that may arise from the use of algorithmic decision support.

and persistent reductions in incarceration rates, as indicated by policy simulations (Kleinberg et al. 2017). However, recent studies of deployed systems suggest that these reductions rarely occur (Stevenson and Doleac 2019, Sloan et al. 2018, Berk 2017). These lackluster results have been attributed, in large part, to the vast heterogeneity in judges' compliance with the tools' recommendations (Cohen and Yang 2019, Stevenson and Doleac 2019). Notably, differential compliance has been shown to be a factor driving increased poor-rich (Skeem et al. 2019) and black-white (Stevenson 2018, Albright 2019) disparities in the post-deployment period. For instance, Albright (2019) found that the increased racial gap in incarceration rates post-deployment resulted both from inter-variation (i.e., judges in whiter counties showed higher compliance with algorithm recommendations) and from intra-variations (i.e., overrides of predicted low and moderate risk were more frequent for black defendants than for white ones).

More broadly, two competing tendencies have been observed in studies of human compliance with algorithmic recommendations: *algorithm aversion* and *automation bias*. Algorithm aversion is the user's tendency to ignore tool recommendations after seeing that they can be erroneous; it originates from the user's self-perceived lack of agency (Lim and O'Connor 1995, DeMichele et al. 2018) and from a lack of algorithmic transparency (Yeomans et al. 2017). Users' reliance on the system is known to vary with the observed accuracy (Yu et al. 2016, 2017) and the asserted accuracy (Yin et al. 2019) of the system. However, humans may not follow algorithmic recommendations even when they are highly reliable (Goodwin and Fildes 1999). Moreover, agents affected by algorithm aversion may prefer human judgment over algorithmic predictions, even when evidence known to both the designer and the user clearly indicates that algorithmic predictions are more accurate than human assessment (Dietvorst et al. 2015).

In contrast, users affected by automation bias may follow algorithmic recommendations despite available (but unnoticed or unconsidered) information that would indicate that the recommendation is wrong. Automation bias consists of two classes of errors: errors of omission and errors of commission. Omission errors are instances where humans fail to detect cases that should have warranted action because the cases were not flagged by the algorithmic system. A prominent example is that of pilots in high-tech cockpits, who are prone to relying only on automated cues as a heuristic replacement for vigilant information seeking (Mosier et al. 1998). Commission errors are instances in which humans take action on the basis of an erroneous algorithmic recommendation, failing to incorporate contradictory external information into the decision process. For example, in the clinical decision support context, commission errors may result in patients' being subjected to unnecessary, potentially invasive testing or treatment.

Studies analyzing factors that contribute to automation bias have found that complex tasks and time pressure may increase overreliance on decision support (Sarter and Schroeder 2001, Goddard et al. 2011). The users' experience level and confidence in their own decisions seems more likely

to lead to automation bias (Marten et al. 2004, Moray et al. 2000). Social accountability has been found to reduce automation bias (Skitka et al. 2000)—an important result in relation to decision support systems that experts who have high public visibility or who are publicly elected (e.g., judges) use. Meanwhile, studies focused on the causes of algorithm aversion have found that repeatedly seeing the algorithm make the same mistake leads to an agent's decreased reliance on the system (Dietvorst et al. 2015); giving some control over the algorithm can counter this phenomenon (Dietvorst et al. 2016).

Motivated by these challenges, recent work has explored approaches to characterizing human-machine complementarity in risk assessment contexts and devising approaches to combine the strengths of both agents. A prominent line of research has focused on the effect of algorithmic explainability on adherence to algorithmic recommendations. These studies show that the perceived accuracy of the system depends on the degree to which the explanations can be understood (Kizilcec 2016, Nourani et al. 2019), while also highlighting that explanations may induce overreliance (Lakkaraju and Bastani 2019). Importantly, findings from ethnographic research that studies the uptake of AI to assist medical diagnosis have shown that AI interrogation practices may extend beyond what can be explained; these findings highlight the need for human experts to be able to relate their own knowledge to AI predictions to make sense of how differences or disagreements arise (Lebovitz et al. 2021). Evidence assessing the effectiveness of interventions designed to mitigate underreliance and overreliance, such as querying the user's decision before showing the recommendations, has shown mixed outcomes (Buçinca et al. 2021, Fogliato et al. 2021, Gajos and Mamykina 2022). Hilgard et al. (2021) propose that, instead of generating predictions, algorithms meant for decision support should be trained with a human in the loop to incorporate the human decision process and should only report to the human user the simpler and more useful representations of the data features. Grounded in the idea of complementarity, new research proposes algorithms that are trained to optimize performance on instances that are difficult for humans, while deferring to human decision makers elsewhere (Madras et al. 2018, Wilder et al. 2020, Raghu et al. 2019). In this line of work, studies have shown that taking into account the heterogeneity of workers' expertise can yield superior human-AI performance (Gao et al. 2021). Finally, and from an algorithmic fairness perspective, recent theoretical work has emphasized the importance of considering the structure of decisions when relating machine predictions to the fairness properties of algorithms (Gillis et al. 2021).

Our work contributes to the nascent literature on human-machine complementarity by conducting a retrospective analysis of whether humans adopt a risk assessment tool and whether they indiscriminately adhere to misestimated risk scores in a real-world decision-making context. The study that is closest to ours is Bushway et al. (2012), which analyzed the effects of inconsistencies in sentencing recommendations resulting from human errors in the judicial setting in Maryland.

The authors find that in aggregate judges become more lenient in the presence of mistakenly recommended lesser sentences for violent offenses but but tend to discount recommendations that are mistakenly too high.

3 Setting and data

Call workers at Allegheny County's child welfare hotline are tasked with deciding whether a call alleging potential child maltreatment or neglect should be screened in for investigation. In making their decisions, call workers have access to the information communicated in the referral call, along with multi-system administrative data on demographics, child welfare involvement, criminal history, and other information related to the children and adults associated with the referral. The administrative data consist of hundreds of data elements. Making systematic and effective use of the administrative data in each case can therefore be challenging for workers. To help workers make better use of this data, Allegheny County introduced a risk assessment tool that distills the information contained therein into a single risk score. The risk score reflects the likelihood that the children on the referral will experience adverse child welfare-related outcomes in the months following the referral. The intended use of the tool is to help workers identify high-risk cases in instances where the information communicated in the call may be insufficient, inconclusive, or otherwise incomplete in reflecting the risk to the children. As we discuss in Section 3.2, the county created specific guidelines to strongly encourage screen-in, and thus the consequent investigation, for the highest scoring cases.

3.1 Risk assessment tool

A case associated with a call is termed by the county a *referral*, and each referral may have several *referral records* associated with it–one for each child involved in the call. The assessment tool that had been deployed was trained with all referral records collected by Allegheny County between April 2010 and July 2014. The tool comprises two distinct predictive models: one to estimate the probability of *out-of-home placement* and one to estimate the probability of a *future referral* for each child. Out-of-home placement refers to the child's being placed out of the home following an investigation, and a future referral refers to a future call involving the same child. Both models are based on features that include demographics, past welfare interaction, public welfare, adult and juvenile criminal justice involvement, and behavioral health information available on all persons associated with each referral. The predicted probabilities are then converted into an integer score ranging from 1 to 20, corresponding to the ventiles. The score shown to workers is calculated as the maximum score over both models and over all children involved in a referral. For example, if the estimated risk of re-referral is 15 and the estimated risk of out-of-home placement is 17, the resulting combined risk score will be 17. We denote this aggregated score as $S \in \{1,\ldots,20\}$. A more detailed description of the model can be found on the County's website (Allegheny County

DHS n.d.).The resulting model was deployed in August 2016.

3.2 Deployment

By design, call workers are shown risk scores only for cases with sufficient information. During the analyzed deployment period, 92.5% of referrals had an associated score shown. In addition to displaying a risk score, the tool assigns a mandatory flag to certain referrals, which means that the supervisor's approval is required to screen out the referral. Flags are placed on cases where at least one of the children associated with the referral has an out-of-home placement score of at least 18. Figure 1 presents a graphical representation of the decision-making pipeline.

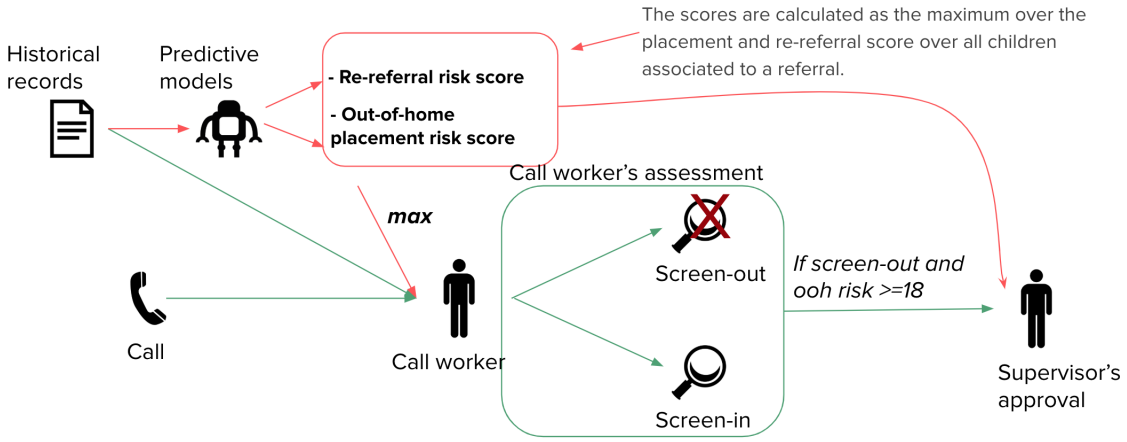

Figure 1: Decision pipeline for child welfare services in Allegheny County. Re-referral risk score and out-of-home placement risk score are calculated for all children associated with a referral. A single score–the maximum over these scores–is shown to the call worker, who also has access to historical records and the information conveyed in the call. The call worker decides whether to screen in the referral for investigation. If a call worker believes that a referral with an out-of-home placement risk score of 18 or more should be screened out, the decision requires a supervisor's approval.

3.3 Scores misestimation

During deployment, a glitch in the system caused certain model inputs to be calculated incorrectly in real time. There are two reasons for this mismatch between the assessed and the shown scores. First, the primary miscalculation issue occurred when real-time database queries erroneously returned counts and indicators of 0. Second, as cases evolved, the roles of different adults associated with the case and information about them may have changed. For instance, the individual identified as a perpetrator may change between the initial run and the retrospective analysis. This type of mismatches is less likely to have a substantial effect on the calculated score.

As a result of the glitch in the retrieval of model inputs, the score displayed (i.e., shown) to the call workers during deployment did not always correspond to the score that should have

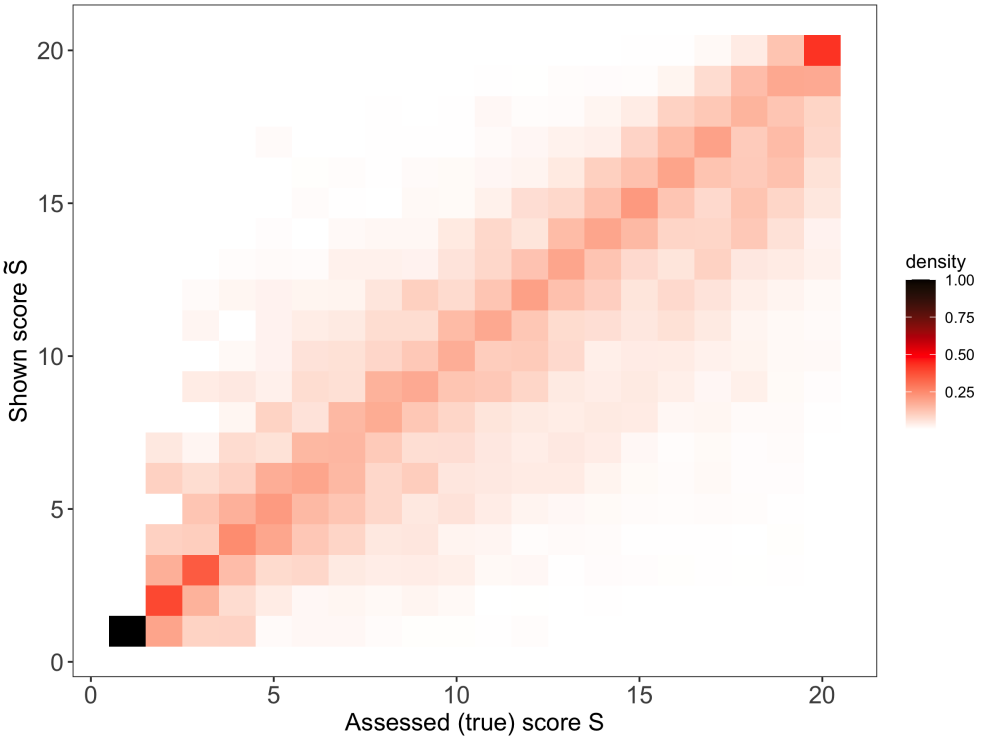

Figure 2: Heatmap of the density of shown score $\tilde{S}$, conditional on assessed score S. A cell at row r and column c shows the fraction of the times that a referral assessed at a score of $S = c$ was shown to have a score of $\tilde{S} = r$.

been shown. Figure 2 shows the distribution of the scores. The concentration of cases around the diagonal indicates that the shown scores were most often equal or very close to the assessed score that should have been displayed. These circumstances of the system's deployment allowed us to study the behavior of call workers when the shown score was inaccurate. In particular, we analyzed what happens when differences between the assessed score and the shown score result in the application of a different "mandatory screen-in" policy. The terminology and notation we use throughout the paper are detailed in Table 1.

3.4 Data

We used data from January 2015 to July 2016 to analyze the behavior of call workers before the deployment of the tool, and we used data from August 2016 until December 2017 to study their behavior after adoption. For the data from the post-deployment period, we limited our analysis to the 93% of referrals for which the system showed an associated score. Moreover, existing regulation dictates that certain calls must be investigated. Referrals that fall under this category include calls that concern bodily injury and sexual abuse. Because call workers have no discretion about whether to screen in these calls, we excluded from our analysis the referrals to which this legislation applies. We also excluded referrals associated with open investigations because call

Table 1: Terminology and notation used in the paper.

Term	Notation	Description
Assessed score	$S \in \{1,\ldots,20\}$	Score assessed by the predictive model. Maximum over assessed risk of out-of-home placement and assessed risk of re-referral.
Shown score	$\tilde{S} \in \{1,\ldots,20\}$	Score shown to the call worker, which should correspond to the assessed score. A glitch in the system resulted in non-correspondence.
Assessed mandatory screen-in	$M \in \{0,1\}$	$M = 1$ if assessed risk of out-of-home placement is greater than or equal to 18. Supervisor's approval required for screening out.
Shown mandatory screen-in	$\tilde{M} \in \{0,1\}$	Mandatory screen-in label shown to the call worker, which should correspond to *assessed mandatory screen-in*. A glitch in the system resulted in non-correspondence.

workers do not make any determination in these cases. Once we limited the data to cases that have a score shown and over which call workers had some discretion, we were left with 12,680 and 10,946 referrals in the pre- and post-deployment periods, respectively.

4 Model Specification and Estimation

We investigated the effects of the tool's deployment and of the misestimation of the scores on call workers' decisions using a regression model. In building the model, we identified three factors that likely correlate with both the decisions and the assessed score S, based on our domain knowledge and conversations with the county's staff. These factors include the month of the year, the identity of the call worker who handles the call, and the types of allegations. The first of these variables, the month, reflects seasonal variations in the nature of the calls and likely reporters. For example, during the school year, more calls from teachers–who are mandatory reporters–may be received. The second, the identity of the call worker, captures any existing heterogeneity in leniency across individuals. The third factor, the nature of the allegations, reflects the seriousness of the case, which may influence the likelihood of screen-ins. For example, the data indicate that workers more often screen in cases where the child lacks a caregiver than they do other cases. Note that each case can have multiple allegations. These considerations led us to formulate a regression model that accounts for the three different factors. For case i that occurred in month t and that was

handled by call worker c, we model the likelihood of screen-in as follows:

$$\begin{aligned} \textit{screen-in}_{itca} \sim \beta_0 + \beta_1 S_i + \beta_2 M_i + [\beta_3 + \beta_4 S_i + \beta_5 M_i] D_i \\ + [\beta_6 \mathbb{1}_{\tilde{S}_i > S_i} + \beta_7 \mathbb{1}_{\tilde{S}_i < S_i} + \beta_8 \tilde{M}_i (1 - M_i) + \beta_9 (1 - \tilde{M}_i) M] D_i + \psi_t + \mu_c + \theta_a \quad (1) \end{aligned}$$

where the βs indicate the coefficients. The dummy variables *screen-in*, D, M, and $\tilde{M}$ are indicators of whether a given case was screened in, whether it occurred in the period after the tool's deployment, whether the tool's recommendation corresponded to a mandatory screen-in, and whether a mandatory screen-in was required in practice, respectively. The model includes fixed effects for months (ψ_t) and call workers (μ_c) to account for shocks over months of the year and for the time-invariant heterogeneity across decision makers. In addition, it contains a series of dummy variables corresponding to the types of allegations (θ_a).

According to our model specification, decisions depend, ceteris paribus, on the assessed score S both in the pre- and post-deployment periods.[3] The strength of this association between decision and score is allowed to vary between the two periods. The dummy variables M and DM are included to incorporate the potentially higher screen-in rates for cases where children are assessed as being at highest risk of out-of-home placement. We handle the misestimation of the scores and its effect on decisions through two sets of components that capture the discrepancy between shown $\tilde{S}$ and assessed scores S, and the interplay between M and $\tilde{M}$. Regarding the former, we distinguish between cases where the assessed score S is underestimated and cases where it is overestimated using the indicator functions $\mathbb{1}(\tilde{S} < S)$ and $\mathbb{1}(\tilde{S} > S)$, respectively. We include each of these terms separately because call workers may be more prone to commit errors of omission or errors of commission, and the distinction between overestimated and underestimated scores allows us to differentiate between these two types of errors. In terms of the latter discrepancy (i.e., between shown and assessed mandatory screen-ins), we include in the model two dummy variables: one that flags cases for which mandatory flags were incorrectly shown, $\tilde{M}(1-M)D$, and one that flags cases that, according to assessed scores, should have had a mandatory flag but did not have it shown, $(1-\tilde{M})MD$. To facilitate the interpretation of these coefficients, the distribution of cases across $(S, \tilde{S}, M, \tilde{M})$ shown in Table 4 in the Appendix is useful.

To assess racial and poverty disparities in screen-in rates, we use models that include all terms in Equation (1), as well as interactions between each of these terms (excluding the fixed effects)

[3] A common econometrics approach used to assess the effect of risk assessment tools on decisions is the regression discontinuity design. This approach leverages the fact that risk scores are continuous but are presented to users in rounded buckets (Cowgill 2018). However, the regression discontinuity design is not suitable for our domain because the discrete nature of the features yields scores that are not continuous. Thus, cases whose risk scores are right below or above a given threshold may differ in key ways, and thus differences in the treatment they receive could be explained by factors other than the score shown.

and a dummy variable corresponding to a demographic attribute. We consider one model for race and one for poverty. The dummy variable for race indicates whether any of the children present in the referral are black, and the dummy variable for poverty indicates whether any of the children involved in the call live in a neighborhood where 20% or more of the households in that neighborhood are below the poverty level. We use one model to assess each type of disparity, and to simplify the notation, we refer to the demographic attribute variable in the generic form of A, which is equal to race or to poverty depending on the model. Our inclusion of the terms corresponding to $[1+S+M][1+D]A$ in the model is motivated by the fact that the use of the tool may have affected heterogeneity in screen-in rates across racial and socioeconomic groups. Thus, including them allows us to study whether workers likelihood to override a recommendation varied across demographic groups. We also include interactions with the terms corresponding to the treatment of misestimated scores–that is, $[\mathbb{1}(\tilde{S}<S)+\mathbb{1}(\tilde{S}>S)+(\tilde{M}(1-M)D))+((1-\tilde{M})MD)][1+D]A$– to ensure that our model accounts for potential differences in adherence to misestimated scores. However, we do not report these terms in the regression results because the sample size is too small to allow for precise estimation and a meaningful interpretation of the coefficients.

We use the same model specification to analyze the quality of the decision-making process, changing the dependent variable from *screen-in* to a relevant outcome that indicates the appropriateness of call workers' decisions. Cases that are screened in are subject to further scrutiny, including home visits by a social worker. As a result of this process, if the referral is connected to a previously closed case, a decision may be made to reopen the investigation. The case also may be "accepted for services," which means that county services and risk-reduction interventions are offered to the family. Both outcomes indicate that the screen-in decision was warranted. We use model specification (1) to investigate whether and how the share of referrals that either were accepted for services or were connected to previous referrals varied after the deployment of the tool, while controlling for the fixed effects. In this analysis, we limit the data to screen-in decisions so that we consider only observed outcomes and make no assumptions based on the counterfactuals.

To estimate the coefficients of model specification in Equation (1), as well as the analogous model using case outcomes as dependent variable, we use logistic regression and compute sandwich standard errors clustered at the call workers' levels. For the statistical tests on these coefficient estimates, we consider a significance level of 0.01. To test the statistical significance of individual coefficients or linear combinations thereof, we use Wald tests.

4.1 Robustness tests

We conduct a series of robustness checks to assess whether our results are threatened by the posited model specification and estimation strategy. First, we estimate the coefficients in Equation (1) using ordinary least squares (OLS) regression and probit regression, i.e., assuming different link

functions. Second, we fit one logistic regression model on data of referrals occurring in the three months prior to and the three months following the tool's deployment, and another on data only from the post-deployment period. Third, we use the *focal reweighting variable* graphical model diagnostics proposed by Buja et al. (2019) to investigate changes in the coefficient estimates when the distribution of the regressors varies. For this analysis, we study the estimates when the model is fitted on data relative to a single month of the year, on referrals handled by an individual call worker, or on cases characterized by similar values of the assessed scores. Fourth, we use a model specification different from that in Equation (1). Here, we assume that the dependent variable (e.g., decisions) depends on a smooth function of the assessed scores S and on the difference between the shown and assessed scores, $\tilde{S} - S$. For the estimation, we use a generalized additive model (GAM) with a logit link function (Hastie and Tibshirani 2017).

The set of results delivered by the robustness tests, using alternative estimation strategies and different time periods, is virtually analogous to the results that we discuss in Section 5 of the paper. Thus, these findings corroborate the conclusions that we draw in this paper. Additionally, the reweighting diagnostics reveal that the regression model in our main analysis is, not surprisingly, misspecified. Similarly, the GAM reveals that logistic regression may not be flexible enough to accurately describe the phenomenon. Nonetheless, our interpretation of the results does change in light of these findings. We discuss the findings and our interpretations in the next sections. In addition, we describe the results of the robustness tests when they are relevant to our discussion but explain most of their details in Section 8 of the Appendix.

5 Analysis

In our analysis, we answer three central questions: (1) *Algorithm aversion:* Did call workers' decisions change after the tool was deployed, or were the recommendations ignored? (2) *Automation bias:* Did call workers treat misestimated scores differently, or did they indiscriminately follow recommendations? and (3) *Disparate treatment:* Did call workers' adherence to algorithmic recommendations vary across racial and socioeconomic groups? If so, did this exacerbate or mitigate existing disparities?

5.1 Algorithm adoption: Change in call workers' behavior

First, we investigated whether the association between assessed scores and call workers' decisions changed after the tool's deployment.[4] If call workers had ignored algorithmic recommendations, the association between the screen-in rates and the assessed score S should have remained the same, ceteris paribus. Meanwhile, if the tool's deployment had affected the likelihood of a case being screened in, we would expect the association between the screen-in rates and the assessed

[4]We focused this analysis on assessed score rather than on shown score because the glitch could not be reproduced. Thus, we could not estimate what shown scores $\tilde{S}$ would have been prior to the tool's deployment.

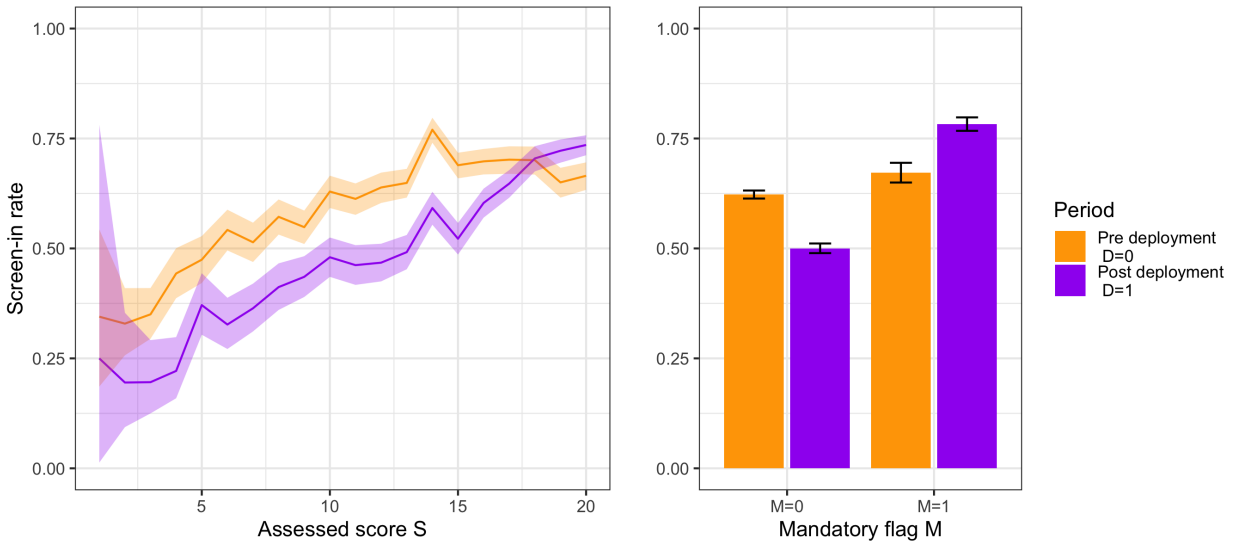

Figure 3: Screen-in rates by assessed score S before and after deployment. Shaded regions and error bars in the left and right figures, respectively, indicate 95% confidence intervals for the estimates.

score S to become stronger after the tool's deployment, ceteris paribus. In the left panel of Figure 3, note that screen-in rates in the post-deployment period appear to be slightly more aligned with the score than screen-in rates in the pre-deployment period. Pearson's correlation coefficients in pre-deployment and post-deployment periods are consistently 0.90 and 0.98, respectively. We found a substantial increase in the screen-in rates, post-deployment, for cases with large values of the assessed score. The plot in the right panel reveals that, for the set of cases with an assessed mandatory flag $M = 1$, screen-in rates increased from 65% in the pre-deployment period to 77% after the deployment. In contrast, screen-in rates for cases without an assessed mandatory flag went down from 61% to 49%. Although the overall screen-in rates remained relatively stable–63% and 57% in pre- and post-deployment periods, respectively–we saw a shift in the types of cases that were screened in. The small change in the screen-in rates can be explained by resource constraints, which limit the number of cases that the county can investigate.

To assess whether the presence of underlying temporal trends may explain these findings, we conducted a placebo test by comparing screen-in rates across score values on data from two periods preceding the tool's deployment: 2015 and the pre-deployment period of 2016 (Figure 6 in the Appendix). Screen-in rates in the two periods were similar across the entire range of scores, indicating that the calibration change seen post deployment was driven by the tool rather than being explained by general temporal trends. We conducted an additional test by comparing the rates before and after deployment only for the call workers that reviewed at least 500 cases in both periods (Figure 6 in the Appendix). On this subset of the data scores and screen-in rates also

became more aligned in the post-deployment period compared to the pre-deployment period.

However, that these patterns might be explained by other factors is possible, and thus, we applied the empirical strategy described in Section 4. For this analysis, we were interested in the estimates of the coefficients in model specification in Equation (1), which are shown in column (1) of Table 2. In particular, we studied the coefficient values concerning the assessed scores S and mandatory flag M, relative to the pre- and post-deployment periods. We observe that, ceteris paribus, cases with larger values of the assessed score S were significantly more likely to be screened in, both pre- and post-deployment (coefficient of S=0.064, 95% confidence interval [0.054, 0.075]). Meanwhile, the coefficient associated with deployment D is negative and statistically significant (-0.62, [-0.984, -0.257]). The two terms that capture the effect of the deployment on the alignment between assessed scores and decisions are SD and MD. These terms estimate the effect of the assessed scores and mandatory flags on screening decisions in the post-deployment period. The coefficient estimate of SD is close to zero and not statistically significant (-0.002, [-0.026, 0.023]). The coefficient estimate of MD, instead, is positive and statistically significant, which suggests that screen-in rates for the subset of cases that were estimated as being at the highest risk for out-of-home placement substantially increased from the pre-deployment period to the post-deployment period, ceteris paribus (1.170, [0.985, 1.355]). Because the screen-in rates remained approximately constant across the pre- and post-deployment periods, the increase in screen-in rates for these cases was associated with the observed decrease in screen-ins for the remaining cases (coefficient of D).

These results are consistent with all of our robustness checks. However, the re-weighting diagnostics reveal that the estimate of the coefficient relative to S is positive and large for low values of S, and close to zero for higher values of the assessed score. Thus, in the pre-deployment period changes in the estimated level of risk were associated with larger changes in the likelihood of screen-in for cases in the low-score range (recall that during this period call workers did not see the score). Meanwhile, the coefficient estimate of SD is positive for cases in the high-score range, which suggests that communicating the risk scores to call workers may have influenced more their post-deployment decisions on this set of referrals, which is consistent with what is observed in Figure 3. The estimates of the coefficients relative to M and MD were approximately constant across the re-weighting diagnostics. We find analogous patterns in the model specification estimated using the GAM.

We next analyzed whether the deployment of the tool affected the quality of the outcomes of the decision-making process. We did so by studying whether the share of referrals that were accepted for services or connected to existing cases varied after the tool's deployment. Before the tool's deployment, 14% of the screened-in referrals were accepted for services; meanwhile, 11% were connected to an existing case involving the same family, and the investigation was reopened.

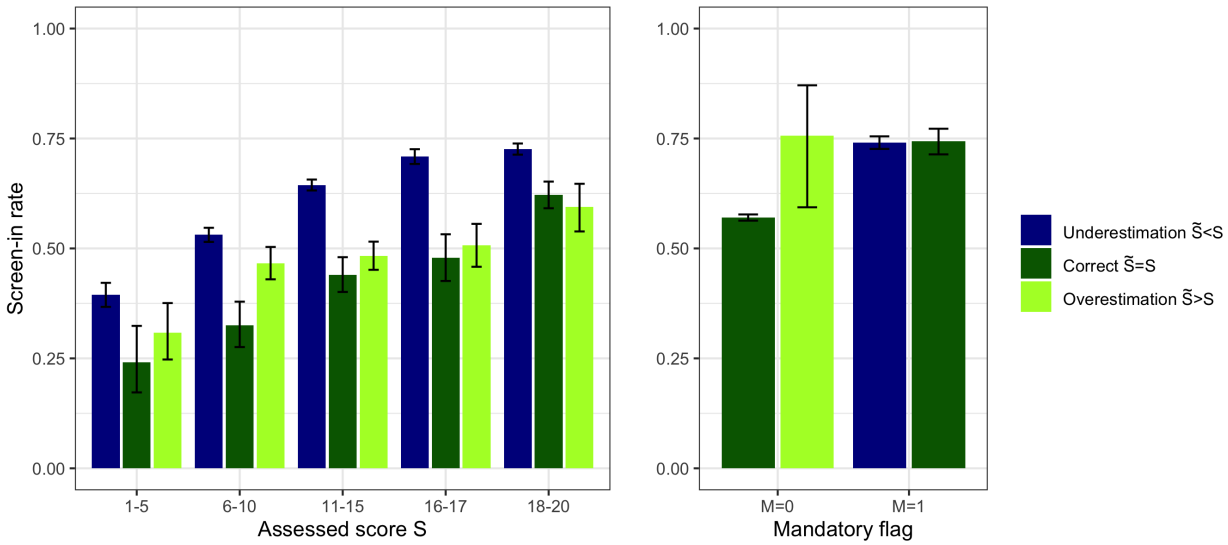

Figure 4: Analysis of screen-in rates across misestimated scores. In the left panel, screen-in rates by assessed score S and shown score $\tilde{S}$. In the right panel, screen-in rates by assessed mandatory flag M and presence of such flag $\tilde{M}$. The summary statistics are computed on data from the post-deployment period only. Error bars indicate 95% confidence intervals.

In the post-deployment period, these rates increased to 17% and 16%, respectively. The regression results in column (2) of Table 2 similarly suggest that, despite the the post-deployment period's substantial increase in screen-in rates for cases with an assessed mandatory flag, the likelihood that they would be accepted for services or connected to other existing referrals was even higher (0.326, [0.034, 0.618]), although not statistically significant. This finding, which is consistent across all of our robustness tests, indicates that the additional screened-in cases did merit an investigation at the same rate (or higher) as previously screened-in cases that had a similar estimated risk of out-of-home-placement.

5.2 Corrective overrides: Decisions in the presence of misestimated scores

Having established that call workers did update their behavior after the deployment of the tool, we investigated whether they indiscriminately adhered to algorithmic recommendations. We had some evidence that they did not do so because we had observed that not all cases that were assigned a mandatory flag were screened in. We performed a more nuanced analysis by studying decisions in the presence of misestimated scores. For this analysis, we used the fact that, because of the glitch that occurred during the post-deployment period (described in Section 3.3), the shown score did not always correspond to the assessed score. As already noted, we distinguish between two types of automation bias: errors of omission and errors of commission. For the former, we investigated call workers' behavior when the shown score underestimated the assessed score–that is, $\tilde{S} < S$ or $\tilde{M} < M$). For the latter, we analyzed cases where the shown score overestimated the assessed

Table 2: Regression results for the assessment of the tool's deployment on call workers' decision-making and case outcomes.

	Dependent variable:	
	Call worker's decision	Case outcome
	(1)	(2)
S	0.064***	0.082***
	(0.005)	(0.008)
D	-0.620***	-2.094***
	(0.185)	(0.193)
SD	-0.002	0.087***
	(0.013)	(0.013)
M	-0.216***	0.054
	(0.049)	(0.114)
MD	1.170***	0.326**
	(0.094)	(0.149)
$\mathbb{1}(\tilde{S} > S)D$	0.174***	0.203*
	(0.052)	(0.120)
$\mathbb{1}(\tilde{S} < S)D$	0.530***	1.015***
	(0.057)	(0.084)
$\tilde{M}(1-M)D$	0.721*	-1.445**
	(0.379)	(0.564)
$(1-\tilde{M})MD$	-0.220***	0.125
	(0.073)	(0.107)
Month FE	Yes	Yes
Call worker FE	Yes	Yes
Allegation type FE	Yes	Yes
Observations	23,626	14,263

Notes: Months of referral, call workers' identity, and allegation types fixed effects (FE) are included in the model but excluded from the table. Standard errors are clustered at the call workers' level. Intercept coefficients are omitted from the table. The dependent variable in the regression in the left column indicates whether the case was screened in (coding=1) or not (0). The dependent variable 'case outcome' indicates whether the case was accepted for service or a closed case was reopened (coding=1), or the case was not accepted for service (0). For the regression where case outcomes are used as the dependent variable, only cases that have been screened in are considered.

*p<0.1; **p<0.05; ***p<0.01

score–that is, $\tilde{S} > S$ or $\tilde{M} > M$.

A crucial limitation to keep in mind for the interpretation of these results is that the glitch did not affect cases completely at random; the misestimate depended, at least in part, on the characteristics of the case.[5] Thus, call workers may have treated correctly estimated and misestimated cases differently because of changes in the underlying level of risk, despite similar values of the assessed score S. This limitation allows for the possibility that call workers screen out a larger share of the cases for which the assessed score is underestimated precisely because these cases are effectively less risky than those for which no misestimation occurred. Analogously, they may screen in a larger share of the cases for which the assessed score is overestimated precisely because these cases are more risky. If and when such a pattern is observed, an analysis of call workers' decisions alone cannot explain the cause of the differences in screen-in rates between misestimated and correctly estimated scores. However, the model still allows us to study whether call workers indiscriminately adhere to the shown score. In other words, based on this model alone, we cannot claim the *presence* of automation bias just because the regression analysis suggests it, but the model does lend itself to studying the *absence* of automation bias. The latter can be inferred if we observe that call workers treat correctly estimated and misestimated cases differently, indicating that they integrate other sources of information into their decisions.

To overcome the limitation resulting from the non-random nature of the glitch, we complemented our analysis with the second regression model, the results of which are reported in column (2) of Table 2; the table shows the effect of algorithmic deployment on an outcome of interest: whether the referral is accepted for service or results in a reopened investigation. In cases where the first regression analysis suggested possible automation bias, this second model allowed us to study whether the level of risk was effectively different. Conversely, in cases where evidence showed that call workers did not indiscriminately adhere to the score, the second model enabled us to study whether overrides improved decision quality. The four interactions that capture potential automation bias are $\mathbb{1}(\tilde{S} > S)D$, $\mathbb{1}(\tilde{S} < S)D$, $\tilde{M}(1-M)D$, and $(1-\tilde{M})MD$. In the remainder of this section we focus on these interactions and consider both models shown in Table 2.

5.2.1 Misestimation of the score

The left panel of Figure 4 shows the screen-in rates for overestimated, underestimated, and correctly estimated assessed scores S. In this exploratory data analysis, we observe that cases with

[5] To assess whether the misestimation depended on other features present in the data, we tried to predict the difference between the assessed and shown scores using the data that were available to us. We trained two models: a Lasso regression and a random forest; their hyperparameters were tuned, via cross-validation, on a superset of the features that were available to the model. The two fitted models achieved moderate predictive accuracy, which was substantially higher than random guessing. This evidence suggests that the misestimation of the score did not occur completely at random. However, neither the county nor our research team has been able to reproduce the glitch. Thus, we cannot retroactively compute the misestimated scores for referrals occurring before the deployment of the tool.

overestimated and correctly estimated scores ($\tilde{S} > S$ and $\tilde{S} = S$, respectively) are screened in at similar rates. Interestingly, cases with underestimated scores ($\tilde{S} < S$) are screened in at *higher* rates. This apparently counterintuitive finding can be potentially explained by the non-random nature of the glitch, as already discussed; that is, cases with underestimated scores may be more risky. Consistent with this hypothesis, we find that this set of cases, when screened in, is accepted for service more frequently than screened-in cases that have lower, correctly estimated scores (see Figure 5 in the Appendix).

Next, we next turn to our regression analysis. The estimate of the coefficient associated with $\mathbb{1}(\tilde{S} < S)D$ in column (1) of Table 2, which is both positive and large, corroborates that cases with underestimated scores were more likely to be screened in than cases with correctly estimated scores S (0.53, [0.419, 0.642]). The results in column (2) of Table 2 show that this set of cases also was significantly more likely to be accepted for service (1.015, [0.850, 1.181]), indicating that they were indeed higher risk. This analysis suggests that call workers avoided errors of omission by successfully integrating other sources of information into their decision-making. The estimated effect of the overestimation of the assessed score $\mathbb{1}(\tilde{S} > S)D$ on decisions is positive and statistically significant, yet small (0.174, [0.071, 0.277]). This analysis suggests possible errors of commission, meaning that call workers would screen in these cases at higher rates than their risk level warranted. If this were the case, we should observe in the second regression model that cases with overestimated scores are associated with a lower likelihood of screen-in in the post-deployment period. However, results from this regression reveal that screened-in cases with overestimated scores were not less likely to be accepted for service than comparable cases with correctly estimated scores (0.203, [-0.033, 0.439]); thus, the regression with respect to case outcomes indicates that call workers' decisions may have been warranted. Two results from the robustness tests are noteworthy here. First, the reweighting diagnostics show that the estimate of the coefficient for $\mathbb{1}(\tilde{S} > S)D$ is positive for low values of the assessed score and that it is close to 0 for higher values of the assessed score. Because the association between screen-in decisions and assessed scores is strongest for cases involving low scores, this measure might suggest some degree of automation bias. Second, the estimates using the GAM show that, as the difference $\tilde{S} - S$ increases, the likelihood that the case would be screened in increases, and the likelihood that the screened-in cases would be accepted for service remains approximately constant, consistent with our results from the main model. To summarize, in this analysis of the effect of the misestimation of the assessed score S on call workers' decisions, we find little evidence of automation bias and significant indication that call workers avoided errors of omission.

5.2.2 Misestimation of the mandatory flag

In the presence of automation bias, a mandatory flag that resulted from a misestimated score should have increased the likelihood of screen-in. Conversely, the likelihood of screen-in should have decreased if a mandatory flag was mistakenly absent absent. The right panel of Figure 4 shows the screen-in rates for cases with the assessed and shown mandatory flags and cases without them– that is, across the values of $(\tilde{M}, M)$. The screen-in rate is constant across cases involving the assessed mandatory flag $M = 1$, regardless of whether the flag was actually shown. This result is reassuring because it suggests that the cases at high risk of out-of-home placement were screened in at high rates, despite the underestimation. Our exploratory analysis also shows that cases that were mistakenly assigned a mandatory flag ($\tilde{M} = 1, M = 0$) were screened in at higher rates than those that were not. Note, however, that the confidence intervals around the former estimate are wide because of the small sample size (41 cases).

The regression analysis in Table 2 reveals similar patterns. The coefficient relative to the lack of mandatory flags $(1-\tilde{M})MD$ is negative and statistically significant (-0.22, [-0.364, -0.077]). However, the majority of these cases (86%) also are such that $\mathbb{1}(\tilde{S} < S) = 1$. Thus, this coefficient must be interpreted jointly with the coefficient associated with $\mathbb{1}(\tilde{S} < S)D$ (0.530, [0.419, 0.642]). For these cases, the positive and statistically significant sum of the coefficient estimates relative to $(1-\tilde{M})MD$ and $\mathbb{1}(\tilde{S} < S)D$ (0.31, p-value<0.01) implies that the cases that mistakenly lacked the mandatory flag were *more* likely to be screened in, compared to cases for which the scores were correctly estimated. Applying an analogous argument to an interpretation of the results of the regression model for case outcomes in column (2) of Table 2, we observe that the coefficient estimate of $(1-\tilde{M})MD$ is close to zero and not statistically significant (0.125, [-0.084, 0.334]), while the coefficient estimate of $\mathbb{1}(\tilde{S} < S)D$ is positive and statistically significant (1.015, [0.850, 1.181]). Thus, these results indicate that call workers avoided errors of omission, screening in cases with underestimated scores at even higher rates than correctly estimated cases, and evidence suggests that these screen-ins were warranted, based on the likelihood of their being accepted for service.

In analyzing errors of commission, we observe that in column (1) of Table 2, the estimate of the coefficient corresponding to the overestimated flags $\tilde{M}(1-M)D$ is positive and large, but not statistically significant (0.721, [-0.021, 1.464]), as a result of the small sample size. The change in the likelihood of screen-in in the presence of an erroneously assigned flag (given by the sum of the coefficients estimates of $\tilde{M}(1-M)D$ and $\mathbb{1}(\tilde{S} > S)D$), compared to the absence of the flag, is positive but not statistically significant (0.895, p-value=0.02). Similarly, the coefficient associated with the term $\tilde{M}(1-M)D$ in column (2) is negative but not statistically significant (-1.445, [-2.550, -0.334]). The negative sum of this coefficient estimate and the coefficient estimate of $\mathbb{1}(\tilde{S} > S)D$, indicates that cases that were mistakenly assigned the mandatory flag were less likely to result in

investigations or to be accepted for service than similar cases. Although the small sample size does not allow us to conclusively identify the presence of automation bias, note that this result does hint at the possibility that the presence of a flag requiring a supervisor's approval to screen-out a call might have induced errors of commission.

Together, the analysis of decisions in the presence of misestimated scores shows compelling evidence indicating that call workers avoided errors of omission, while their behavior showed a tendency toward automation bias when a mandatory flag was shown.

5.3 Disparate treatment

Bias in algorithmic-informed decisions may stem from biased predictions or from disparate algorithmic uptake. A comprehensive evaluation of the fairness properties of the tool in isolation has been conducted by Chouldechova et al. (2018). In this section, we investigate whether the adherence to the tool's assessment by the call workers varied on the basis of a child's race and of a family's socioeconomic status.

5.3.1 Racial disparities

We first focus on the analysis of racial disparities. When considering disparities in algorithmic-informed decisions, two points of reference are relevant: the human decisions prior to the tool's deployment and the algorithmic predictions in isolation (i.e. what the decisions would have been if the tool made autonomous decisions). Among all cases involving Black children ($A = 1$), 68% and 62% were screened in in the periods before and after the deployment of the tool, respectively. Among the remaining cases ($A = 0$), the screen-in rates were 59% and 53%. Thus, we observe that screen-in rates for Black children were higher both before and after deployment, and that the differences between groups remained largely the same. If the tool were making autonomous decisions and screening in the highest scoring cases while maintaining the same screen-in rate, 66% of the cases involving Black children would be screened in (based on the shown score $\tilde{S}$), while 51% of other cases would be. Thus, human decisions in isolation and algorithmic-informed human decisions led to smaller disparities than those associated with autonomous algorithmic decisions.

To understand what is driving the reduced disparities when humans are involved in the decisions, and whether the likelihood of call workers' adhering to algorithmic recommendations varies based on a child's race, we turn to the regression analysis presented in column (1) of Table 3. The positive and statistically significant coefficient estimate for the race intercept A (0.426, [0.220, 0.632]), together with the near-zero estimate for AS ([-0.019 , 0.018]), indicates that Black children were more likely to be screened in regardless of the level of risk. This disparity was unaffected by the tool's deployment, as indicated by the coefficient associated with AD (-0.035, [-0.455, 0.385]) and by ASD's being close to zero and not statistically significant (-0.007, [-0.036 0.022]). These coefficients estimates suggest that call workers interpreted changes in the assessed scores

Table 3: Regression results for the assessment of racial and socioeconomic disparities in screen-in rates.

	Dependent variable: Screen-in decision	
	Demographic attribute A	
	A=1 if at least one child in referral is black	A=1 if % families living below poverty threshold $\geq$ 20%
	(1)	(2)
A	0.426***	-0.320*
	(0.105)	(0.168)
S	0.055***	0.054***
	(0.005)	(0.005)
AS	-0.0004	0.031**
	(0.010)	(0.013)
D	-0.675***	-0.607***
	(0.239)	(0.185)
AD	-0.035	-0.020
	(0.214)	(0.220)
SD	0.005	0.002
	(0.015)	(0.013)
ASD	-0.007	-0.012
	(0.015)	(0.016)
M	-0.169**	-0.203***
	(0.083)	(0.057)
AM	-0.097	-0.118
	(0.122)	(0.113)
MD	1.397***	1.205***
	(0.169)	(0.128)
AMD	-0.311	-0.003
	(0.219)	(0.221)
Month FE	Yes	Yes
Call worker FE	Yes	Yes
Allegation type FE	Yes	Yes
Observations	23,626	23,626

Notes: Month of referral's, call workers', and allegation types' fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercept coefficients are omitted from the table. Terms relative to the shown scores also are included in the model, but the corresponding coefficient estimates are omitted from the table.

*p<0.1; **p<0.05; ***p<0.01

in a similar manner, regardless of the racial identity of the child. This result, in turn, indicates that the racial differences in screen-in rates, whether comparing the algorithm in isolation or the algorithm-informed decisions, are driven not by differential adherence to recommendations but by human decisions relying on factors other than the risk score alone.

5.3.2 Socioeconomic disparities

The second part of our analysis targets socioeconomic disparities. Here, A indicates whether the child resides in a neighborhood where more than 20% of the families live under the poverty threshold ($A = 1$). For cases with $A = 1$, we observe a reduction in the screen-in rates–from 66% to 58%–from the pre-deployment to the post-deployment period. The remaining cases ($A = 0$) are characterized by lower screen-in rates in the pre-deployment period (61%), but they are screened in at similar levels (57%) after deployment. Thus, after the deployment, these two types of cases were screened in at similar rates, despite having different screen-in rates before deployment. This similarity in the post-deployment period is surprising: If the tool were used to make autonomous decisions, 65% of the cases with $A = 1$ and 54% of those with $A = 0$ would have been screened in. Thus, we again observe that algorithmic-informed human decisions exhibit less disparities than the algorithm in isolation.

In the regression results reported in column (2) of Table 3, we observe that none of the interactions with the demographic attribute are statistically significant at the 0.01 level. Nonetheless, the estimates of the coefficients associated with A, which is negative (-0.32, [-0.649, 0.009]), and with AS, which is positive (0.031, [0.006, 0.056]), would indicate that, regardless of the time period, the relationship between the assessed score and the screen-in decisions differed across socioeconomic groups: Increases in the assessed score were associated with larger incremental increases in screen-in rates if the child lived in a neighborhood with more than 20% of families living below the poverty threshold. Thus, ceteris paribus, these children were more likely to be screened in than their counterparts in other neighborhoods when the level of risk was higher than average (S>10). Because the estimates of the coefficients relative to the deployment of the tool were close to zero and not statistically significant, they indicate that the deployment did not have a substantial effect on decisions for either of the two groups–neither because of the access to the score nor because of the presence of the mandatory flag. This result, again, suggests that algorithmic-informed decisions exhibit smaller disparities than the algorithm in isolation because the call workers integrate other sources of information into their decisions.

6 Discussion

Our research has analyzed algorithmic adoption, overrides, and disparities in the context of a risk assessment tool that assists call workers tasked with deciding which calls concerning potential child neglect or abuse should be screened in for further investigation. We focused the first part of

our analysis on investigating whether humans changed their behavior when the tool was deployed and whether they were more likely to deviate from recommendations that resulted from misestimated algorithmic scores. We found that call workers did change their behavior when the tool was deployed, showing partial adherence to the tool's recommendations. We also found that, despite evidence of a slight tendency toward automation bias in the presence of a mandatory flag, call workers did make corrective overrides. In particular, strong evidence suggested that call workers avoided errors of omission, dismissing recommendations that would have led them to erroneously screen out calls when the technical glitch resulted in an underestimation of the score.

When considering how humans make use of recommendations provided by an algorithmic system, the phenomena of algorithm aversion and automation bias can be seen as two ends of a broad spectrum, and both of them are undesirable. On the one end, algorithm aversion leads humans to completely disregard the machine, even when the recommendation may be providing useful information. On the other end, automation bias results in humans blindly following the machine's recommendations, failing to make use of other sources of information and their own judgment to disregard the recommendation when evidence suggests that the machine may be mistaken. The call workers in our study were found to be at neither extremum: They changed their decision-making following the deployment of the tool, but they did not indiscriminately adhere to the tool's recommendations.

Perhaps one surprising result is that cases with assessed mandatory flags M were more likely to be screened in, even though the glitch very often resulted in the flags not being shown (as seen in Table 4). Here, recall that most of these cases still had relatively high shown scores; only 10% of cases were misestimated by more than 5 points. Thus, the deployment of the tool resulted in the availability of new information indicating that these cases possibly involved high risk to the child. Furthermore, the deployment of the tool might have shifted the decision-making criteria, as suggested by recent research (Green and Chen 2021), and this shift might have led to increased attention to historical indicators of risk in the administrative data.

The risk of reproducing and exacerbating societal disparities is a central concern around the deployment of algorithmic tools to assist with experts' decisions (Barocas and Selbst 2016). As discussed in Section 2, differential adherence to recommendations has been shown to be a potential source of disparities. In this work, we have studied whether call workers' adherence to recommendations varied on the basis of the race and socioeconomic status of the children involved in the call. Our findings differ from the results of research that studies the adoption of risk assessment tools in other domains, such as recidivism prediction for bail decisions. We find that racial or socioeconomic disparities in screen-in rates were not affected by the deployment of the tool. Our findings also show that, when comparing humans' algorithmic-informed decisions with algorithmic predictions in isolation, humans in the loop mitigated disparities. This effect was not driven by disparate

adherence to recommendations. Instead, it was driven by the way call workers integrated factors other than the score into their decisions. This finding highlights that humans' discretionary power and ability to integrate information that is unobserved by the algorithm can have important fairness implications.

The nature of our study has the advantage of inherent field validity, by virtue of its being a study of a system deployed in the real world. Although crowdsourcing studies and randomized field experiments allow for controlled variations of different factors, even the best-conceived randomized trials can fail to have field validity in social policy settings (Nagin and Sampson 2019). Moreover, given the high-stakes nature of the task, behaviors displayed by call workers when given hypothetical calls in a lab setting may differ from their behavior when faced with a real allegation of child abuse. Thus, studying human decisions in the presence of a technical glitch offered a rare opportunity to study a phenomenon that would not be ethically feasible to investigate as a randomized field trial. Nonetheless, the retrospective nature of our study represents a clear limitation.

What contributed to the desirable human behaviors we observed? The retrospective nature of our analysis means that we were unable to identify how different elements of the decision-making framework impacted decision outcomes. However, through our analysis and discussions with jurisdiction staff, we were able to identify certain elements of the deployment setup that could have influenced the observed behavior, which we elaborate on next. An important direction for future research is to further investigate these and other factors in more controlled settings to gain a better understanding of their influence on commission and omission errors in algorithm-assisted decision-making.

We discuss three elements of our particular setup. First, a key property of this deployment setup is that, throughout the process, call workers continued to have access to both the referral calls and the administrative data system. This access provided a different and broader view of the case than what was being pulled into the risk score calculation. In particular, even when inputs related to past child welfare history were being miscalculated in real time, workers still had access to the correct information in the data system. Having access to the raw features and the time to inspect them may have played an important role. Importantly, many call workers had also been previously trained to make decisions without the aid of a risk assessment tool. Therefore, they had experience in parsing and interpreting the raw data themselves. A question that arises is whether this previous experience played an important role, and whether similar decision-making could be expected from call workers who started working after the tool's deployment. The answer to this question could inform decisions about the need to train experts on how to use other data sources appropriately, to avoid an over-reliance on algorithmic recommendations.

Second, the risk tool provides workers only with a score and does not "explain" its predictions, nor does it display values of any of the features involved in the score calculation. If this additional

information had been provided, the glitch might have been detected by workers. However, this distillation of the data also might have been trusted by workers, who in turn would have been dissuaded from examining the original data. In the latter case, the explanations might have induced over-reliance.

Research has shown that in some settings algorithms deployed in isolation may outperform human-in-the-loop systems, leading some to call for complete automation. Instead, the results presented in this paper highlight one of the beneficial roles that trained and experienced humans-in-the-loop can play, guarding against harmful effects that can result from erroneous algorithmic recommendations. Providing humans with autonomy to override the machine mitigated the effects of miscalculated scores in the child maltreatment call screening context, and also mitigated racial and socioeconomic disparities that would have resulted from algorithmic autonomous decisions. Given that technical glitches, such as the one studied in this paper, always represent a potential risk, and that any statistical model will make inaccurate predictions, design should focus on augmenting the human's ability to identify and correct those mistakes. Future research in controlled settings that evaluates the effect of individual elements of the decision-making pipeline described in this paper could identify specific design practices that effectively strengthen the human's role.

References

Acimovic J, Parker C, F Drake D, Balasubramanian K (2020) Show or tell? improving inventory support for agent-based businesses at the base of the pyramid. *Manufacturing & Service Operations Management* .

Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, Nichols CN, Lampropoulos GK, Walker BS, Cohen G, et al. (2006) The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist* 34(3):341–382.

Albright A (2019) If you give a judge a risk score: evidence from kentucky bail decisions. *Harvard John M. Olin Fellow's Discussion Paper* 85.

Allegheny County DHS (n.d.) Allegheny family screening tool. URL `https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx`.

Baesens B, Setiono R, Mues C, Vanthienen J (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science* 49(3):312–329.

Barocas S, Selbst AD (2016) Big data's disparate impact. *California Law Review* 104:671–732.

Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13(2):193–216.

Brown A, Chouldechova A, Putnam-Hornstein E, Tobin A, Vaithianathan R (2019) Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic

decision-making in child welfare services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 41 (ACM).

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–21.

Buja A, Brown L, Kuchibhotla AK, Berk R, George E, Zhao L (2019) Models as approximations ii: A model-free theory of parametric regression. *Statistical Science* 34(4):545–565.

Bushway SD, Owens EG, Piehl AM (2012) Sentencing guidelines and judicial discretion: Quasi-experimental evidence from human calculation errors. *Journal of Empirical Legal Studies* 9(2):291–319.

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730 (ACM).

Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of ACM FAccT*, 134–148.

Cohen A, Yang CS (2019) Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy* 11(1):160–91.

Cowgill B (2018) The impact of algorithms on judicial discretion: Evidence from regression discontinuities. Technical report, Working paper.

Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgment. *Science* 243(4899):1668–1674.

De-Arteaga M, Fogliato R, Chouldechova A (2020) A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

DeMichele M, Baumgartner P, Barrick K, Comfort M, Scaggs S, Misra S (2018) What do criminal justice professionals think about risk assessment at pretrial? Available at SSRN 3168490.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Dietvorst BJ, Simmons JP, Massey C (2016) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Eubanks V (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor* (St. Martin's Press).

Fogliato R, Chouldechova A, Lipton Z (2021) The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–24.

Gajos KZ, Mamykina L (2022) Do people engage cognitively with ai? impact of ai assistance on incidental learning. *arXiv preprint arXiv:2202.05402* .

Ganju KK, Atasoy H, McCullough J, Greenwood B (2020) The role of decision support systems in attenuating racial biases in healthcare delivery. *Management Science* 66(11):5171–5181.

Gao R, Saar-Tsechansky M, De-Arteaga M, Han L, Lee MK, Lease M (2021) Human-ai collaboration with bandit feedback. *International Joint Conferences on Artificial Intelligence (IJCAI)* .

Gillis T, McLaughlin B, Spiess J (2021) On the fairness of machine-assisted human decisions. *arXiv preprint arXiv:2110.15310* .

Goddard K, Roudsari A, Wyatt JC (2011) Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1):121–127.

Goodwin P, Fildes R (1999) Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making* 12(1):37–53.

Green B, Chen Y (2021) Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–33.

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12(1):19.

Hastie TJ, Tibshirani RJ (2017) *Generalized additive models* (Routledge).

Hilgard S, Rosenfeld N, Banaji MR, Cao J, Parkes D (2021) Learning representations by humans, for humans 139:4227–4238.

Karlinsky-Shichor Y, Netzer O (2019) Automating the b2b salesperson pricing decisions: Can machines replace humans and when. Available at SSRN:3368402.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182–5190.

Kizilcec RF (2016) How much information?: Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395 (ACM).

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. *The quarterly journal of economics* 133(1):237–293.

Lakkaraju H, Bastani O (2019) " how do i fool you?": Manipulating user trust via misleading black box explanations. *arXiv preprint arXiv:1911.06473* .

Lebovitz S, Lifshitz-Assaf H, Levina N (2021) To engage or not to engage with ai for critical judgments: How professionals deal with opacity when using ai for medical diagnosis. *at Organization Science Special Issue on Theorizing Emerging Technologies* .

Lim JS, O'Connor M (1995) Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making* 8(3):149–168.

Madras D, Pitassi T, Zemel R (2018) Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 6147–6157.

Marten K, Seyfarth T, Auer F, Wiener E, Grillhösl A, Obenauer S, Rummeny EJ, Engelke C (2004) Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. *European radiology* 14(10):1930–1938.

Meehl PE (1954) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. *In Proceedings of the 1955 Invitational Conference on Testing Problems*, 136–141 (University of Minnesota Press).

Moray N, Inagaki T, Itoh M (2000) Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of experimental psychology: Applied* 6(1):44.

Mosier KL, Skitka LJ, Heers S, Burdick M (1998) Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology* 8(1):47–63, URL `http://dx.doi.org/10.1207/s15327108ijap0801_3`.

Nagin DS, Sampson RJ (2019) The real gold standard: Measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology* 2(1):123–145, URL `http://dx.doi.org/10.1146/annurev-criminol-011518-024838`.

Nourani M, Kabir S, Mohseni S, Ragan ED (2019) The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 97–105.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* 61(8):1741–1759.

Raghu M, Blumer K, Corrado G, Kleinberg J, Obermeyer Z, Mullainathan S (2019) The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220* .

Sarter NB, Schroeder B (2001) Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human factors* 43(4):573–583.

Saxena D, Badillo-Urquiola K, Wisniewski PJ, Guha S (2020) A human-centered review of algorithms used within the us child welfare system. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Skeem JL, Scurich N, Monahan J (2019) Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Virginia Public Law and Legal Theory Research Paper* (2019-02).

Skitka LJ, Mosier KL, Burdick M, Rosenblatt B (2000) Automation bias and errors: are crews better than individuals? *The International journal of aviation psychology* 10(1):85–97.

Sloan C, Naufal G, Caspers H (2018) The effect of risk assessment scores on judicial behavior and defendant outcomes. IZA Discussion Paper, SSRN:3301699.

Smith VC, Lange A, Huston DR (2012) Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks* 16(3):51–61.

Stevenson M (2018) Assessing risk assessment in action. *Minn. L. Rev.* 103:303.

Stevenson MT, Doleac JL (2019) Algorithmic risk assessment in the hands of humans. *IZA Discussion Papers* (N0.12853).

Van Donselaar KH, Gaur V, Van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Science* 56(5):766–784.

Wilder B, Horvitz E, Kamar E (2020) Learning to complement humans. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 1526–1533.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2017) Making sense of recommendations. *Journal of Behavioral Decision Making* 32.

Yin M, Wortman Vaughan J, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 279 (ACM).

Yu K, Berkovsky S, Conway D, Taib R, Zhou J, Chen F (2016) Trust and reliance based on system accuracy. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 223–227 (ACM).

Yu K, Berkovsky S, Taib R, Conway D, Zhou J, Chen F (2017) User trust dynamics: An investigation driven by differences in system performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317 (ACM).

7 Additional exploratory data analysis

Table 4: Distribution of cases by assessed and shown scores and flags $(S, \tilde{S}, M, \tilde{M})$ after deployment.

	$M=0, \tilde{M}=0$	$M=0, \tilde{M}=1$	$M=1, \tilde{M}=0$	$M=1, \tilde{M}=1$
$S < \tilde{S}$	23% (2506)	0% (33)	0% (53)	1% (60)
$S = \tilde{S}$	16% (1730)	0% (7)	2% (223)	5% (498)
$S > \tilde{S}$	35% (3820)	0% (1)	15% (1682)	3% (333)

Notes: Raw numbers are reported in parentheses.

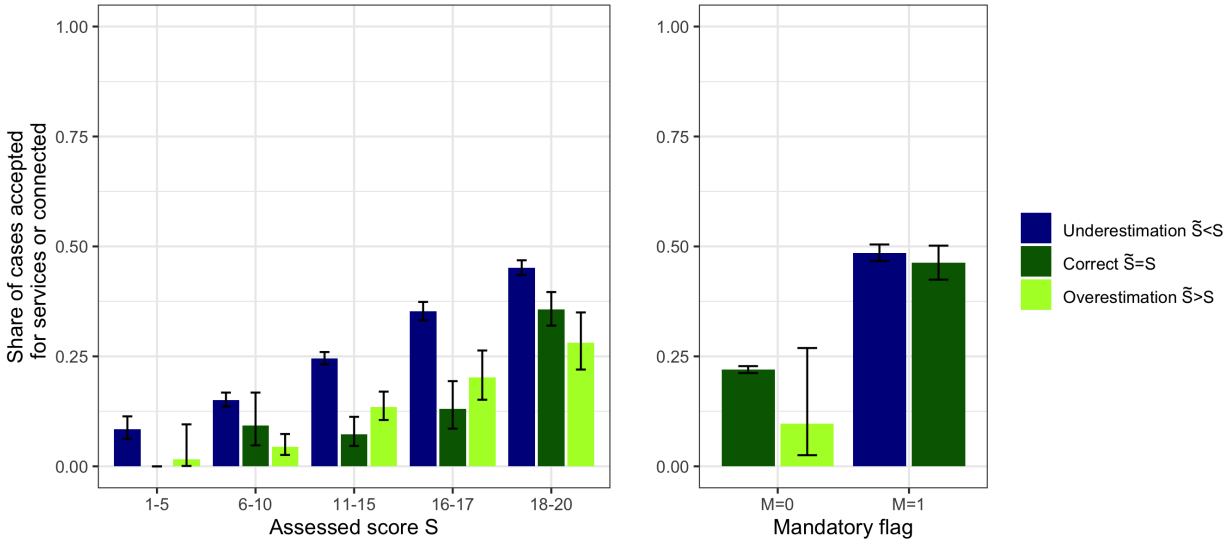

Figure 5: Analysis of rates of cases accepted for services or connected to a previously closed case across misestimated scores. In the left panel, accepted/connected rates by assessed score S and its relationship to the shown score $\tilde{S}$. In the right panel, accepted/connected rates by assessed mandatory flag M and presence of such flag $\tilde{M}$. The summary statistics are computed on data from the post-deployment period only. Error bars indicate 95% confidence intervals for the mean.

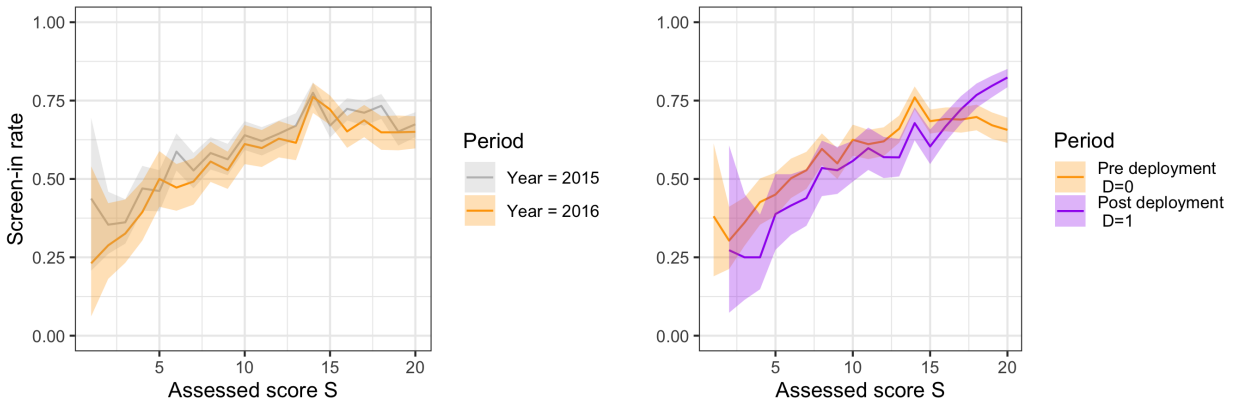

Figure 6: Screen-in rates across values of the assessed score *S*. The lines indicate screen-in rates. The bands indicate 95% confidence intervals for the mean. The left panel shows screen-in rates for two periods preceding the tool's deployment: 2015 and the pre-deployment period of 2016. The right panel shows screen-in rates computed solely on data of call workers that handled at least 500 cases in both pre- and post-deployment period.

8 Robustness tests

In this section, we present the results of several tests we conduct to assess the robustness of the findings presented in the main body of the paper. Those findings are based on the model specification in Equation (1), estimated through logistic regression. The tests that we discuss here use alternative estimation strategies and modeling specifications.

8.1 Alternative link function

We first compare the coefficients estimated through the logistic regression with those obtained using probit regression and linear regression via ordinary least squares (OLS, a.k.a. linear probability model) to estimate model specification (1). The coefficients estimates produced by the three models are displayed in Table 5 and 6 for decisions and outcomes, respectively. We observe that the signs and statistical significance of the estimates of the regression coefficients produced via OLS, the probit regression, and the logistic regression match for all terms. Thus, the probit regression and OLS deliver a similar set of results to those of our main regression model, which we have discussed in Section 5. The coefficients estimates for the regressions targeting racial disparities and socioeconomic in screen-in decisions are reported in Tables 7 and 8 respectively. The direction of the estimated effects is similar across all models.

8.2 Estimation on data from narrower time windows

In this section, we assess whether the estimates obtained by fitting the logistic regression to estimate the coefficients in model (1) remain the same when the same model and estimation strategy are employed on data of referrals either from the post-deployment period alone or occurring in a

Table 5: Comparison of regression results for the assessment of the impact of the tool's recommendations on call workers' screen-in decisions via alternative estimation strategies.

	Dependent variable: call worker's decision		
	Estimation Strategy:		
	logistic	*probit*	*OLS*
	(1)	(2)	(3)
S	0.061***	0.037***	0.013***
	(0.005)	(0.003)	(0.001)
D	-0.725***	-0.439***	-0.107***
	(0.185)	(0.111)	(0.037)
SD	0.003	0.002	-0.001
	(0.012)	(0.007)	(0.002)
M	-0.210***	-0.125***	-0.047***
	(0.049)	(0.030)	(0.009)
MD	1.160***	0.682***	0.198***
	(0.099)	(0.058)	(0.017)
$\mathbb{1}(\tilde{S} > S)D$	0.185***	0.111***	0.024**
	(0.052)	(0.032)	(0.011)
$\mathbb{1}(\tilde{S} < S)D$	0.536***	0.325***	0.093***
	(0.056)	(0.033)	(0.012)
$\tilde{M}(1-M)D$	0.719*	0.435**	0.133**
	(0.378)	(0.219)	(0.062)
$(1-\tilde{M})MD$	-0.231***	-0.141***	-0.045***
	(0.068)	(0.039)	(0.010)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,795	23,795	23,795
R^2			0.233

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table.

*p<0.1; **p<0.05; ***p<0.01

Table 6: Comparison of regression results for the assessment of the impact of the tool's recommendations on outcomes of screened-in cases via alternative estimation strategies.

	Dependent variable: Case outcome		
	Estimation Strategy:		
	logistic	*probit*	*OLS*
	(1)	(2)	(3)
S	0.084***	0.048***	0.012***
	(0.008)	(0.005)	(0.001)
D	-2.041***	-1.148***	-0.242***
	(0.195)	(0.107)	(0.023)
SD	0.084***	0.047***	0.008***
	(0.013)	(0.007)	(0.002)
M	0.052	0.037	0.029
	(0.113)	(0.068)	(0.022)
MD	0.334**	0.211**	0.093***
	(0.150)	(0.091)	(0.029)
$\mathbb{1}(\tilde{S} > S)D$	0.204*	0.131*	0.048***
	(0.122)	(0.070)	(0.017)
$\mathbb{1}(\tilde{S} < S)D$	1.015***	0.597***	0.173***
	(0.085)	(0.051)	(0.016)
$\tilde{M}(1-M)D$	-1.449**	-0.819***	-0.191***
	(0.563)	(0.308)	(0.047)
$(1-\tilde{M})MD$	0.128	0.089	0.050**
	(0.106)	(0.064)	(0.023)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	14,399	14,399	14,399
R^2			0.152

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table. The case outcome is coded as "1" if the case is accepted for services or connected to a closed case and an investigation is opened, as "0" if the case is not accepted for services.

*p<0.1; **p<0.05; ***p<0.01

Table 7: Comparison of regression results for the assessment of racial disparities in call workers' screen-in decisions via alternative estimation strategies.

	Dependent variable: call worker's decision		
	Estimation Strategy:		
	logistic	*probit*	*OLS*
	(1)	(2)	(3)
A	0.426***	0.262***	0.098***
	(0.105)	(0.065)	(0.021)
S	0.055***	0.033***	0.012***
	(0.005)	(0.003)	(0.001)
AS	-0.0004	-0.001	-0.002
	(0.010)	(0.006)	(0.002)
D	-0.675***	-0.399***	-0.085*
	(0.239)	(0.144)	(0.045)
AD	-0.035	-0.027	-0.033
	(0.214)	(0.131)	(0.038)
SD	0.005	0.002	-0.001
	(0.015)	(0.009)	(0.003)
ASD	-0.007	-0.004	0.0002
	(0.015)	(0.009)	(0.003)
M	-0.169**	-0.101**	-0.043***
	(0.083)	(0.051)	(0.016)
AM	-0.097	-0.056	-0.007
	(0.122)	(0.073)	(0.023)
MD	1.397***	0.820***	0.230***
	(0.169)	(0.097)	(0.027)
AMD	-0.311	-0.180	-0.045
	(0.219)	(0.125)	(0.034)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,626	23,626	23,626
R^2			0.238

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table.

*p<0.1; **p<0.05; ***p<0.01

Table 8: Comparison of regression results for the assessment of socioeconomic disparities in call workers' screen-in decisions via alternative estimation strategies

	Dependent variable: call worker's decision		
	Estimation Strategy:		
	logistic	*probit*	*OLS*
	(1)	(2)	(3)
A	-0.320*	-0.189*	-0.053
	(0.168)	(0.102)	(0.033)
S	0.033***	0.012***	
	(0.005)	(0.003)	(0.001)
AS	0.031**	0.018**	0.005**
	(0.013)	(0.008)	(0.002)
D	-0.607***	-0.368***	-0.097***
	(0.185)	(0.111)	(0.037)
AD	-0.020	0.012	0.042
	(0.220)	(0.139)	(0.041)
SD	0.002	0.001	-0.001
	(0.013)	(0.008)	(0.003)
ASD	-0.012	-0.008	-0.004
	(0.016)	(0.010)	(0.003)
M	-0.203***	-0.122***	-0.048***
	(0.057)	(0.036)	(0.011)
AM	-0.118	-0.069	-0.015
	(0.113)	(0.067)	(0.021)
MD	1.205***	0.712***	0.209***
	(0.128)	(0.074)	(0.019)
AMD	-0.003	-0.0005	-0.003
	(0.221)	(0.129)	(0.034)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,626	23,626	23,626

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table.

*p<0.1; **p<0.05; ***p<0.01

narrower time window around the tool's deployment. Unfortunately, due to initial limitations in the implementation, the risk scores of newborns in the first four months after deployment were shown as 0. This was fixed at the end of November 2016. Since our analysis only considers cases for which a score was estimated, the characteristics of our sample right before and after deployment will differ. Consequently, the usual regression discontinuity design cannot be applied to data from consecutive months. We circumvent the issue by considering referrals that occurred in the three months prior to deployment and the three months following the resolution of the issue, i.e., the pre-deployment months that we consider are May, June, and July 2016 and the post-deployment months are December 2016, January and February 2017.

This analysis is important for several reasons. First, referrals occurring in the period close to the deployment likely share similar characteristics. Thus, by considering a narrow time window, it seems possible that our coefficients estimates would be less affected by variations in the cases characteristics that are associated to both the score and to call workers' decisions, but that are not captured by our model. Second, while there were no major changes to the counties' policies during the post-deployment time that we consider, it is possible that minor changes regarding how to handle the various types of allegations may have taken place. Finally, it is also possible that there were drifts in call workers' behavior over time. Through the regression results presented in this section, we wish to test whether any of these factors, if present, affected our interpretation of the main regression results.

We run two separate regressions with decisions and outcomes as dependent variables, respectively. Tables 9 and 10 show the results of these regression analyses. The two additional regression models where call workers' decisions represent the dependent variable deliver a set of coefficients estimates whose signs are identical to those in our main model. The size of the estimated effects are also similar. For the models where the outcomes of screen-in decisions are the dependent variable, we observe only one marginal difference between the alternative regressions and our main model: While the estimate of the coefficient of $\mathbb{1}(\tilde{S} > S)D$ is positive and statistically significant in the main regression model, its estimate is close to 0 for the regression fitted on data around the period of deployment. This finding does not affect our conclusions on the absence of automation bias in the call workers' interactions with the score. All other coefficients estimates are similar across the three models. The regression results for the assessment of racial and socioeconomic disparities in screen-in decisions, which are reported in Tables 11 and 12 respectively, are consistent with our discussion in the main body of the paper as well.

Table 9: Comparison of regression results for logistic regression estimating the impact of the score on screen-in decisions on data relative to different time periods

	Dependent variable: call worker's decision		
		Time period considered:	
	Entire period	May, Jun, Jul 2016 and Dic 2016, Jan, Feb 2017	Post deployment
	(1)	(2)	(3)
S	0.061***	0.079***	0.067***
	(0.005)	(0.009)	(0.011)
D	-0.725***	-0.586	
	(0.185)	(0.394)	
SD	0.003	-0.014	
	(0.012)	(0.029)	
M	-0.210***	-0.390***	0.971***
	(0.049)	(0.144)	(0.085)
MD	1.160***	1.528***	
	(0.099)	(0.408)	
$\mathbb{1}(\tilde{S} > S)D$	0.185***	0.279	0.182***
	(0.052)	(0.171)	(0.051)
$\mathbb{1}(\tilde{S} < S)D$	0.536***	0.668***	0.554***
	(0.056)	(0.202)	(0.054)
$\tilde{M}(1-M)D$	0.719*	0.315	0.715*
	(0.378)	(1.065)	(0.380)
$(1-\tilde{M})MD$	-0.231***	-0.506	-0.263***
	(0.068)	(0.351)	(0.068)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,795	3,850	10,946

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table. Note that the regression coefficients reported in the middle column are based on a logistic regression model estimated on data from May, June, July 2016 (pre-deployment) and December 2016, January and February 2017 (post-deployment period).

*p<0.1; **p<0.05; ***p<0.01

Table 10: Comparison of regression results for the assessment of the impact of the tool's recommendations on outcomes of screen-in decisions on data relative to different time periods

	Dependent variable: case outcome		
	Time period considered:		
	Entire period	May, Jun, Jul 2016 and Dic 2016, Jan, Feb 2017	Post deployment
	(1)	(2)	(3)
S	0.084***	0.075***	0.171***
	(0.005)	(0.014)	(0.013)
D	-2.041***	-2.523***	
	(0.185)	(0.445)	
SD	0.084***	0.119***	
	(0.012)	(0.023)	
M	0.052	-0.189	0.411***
	(0.049)	(0.260)	(0.114)
MD	0.334***	0.501	
	(0.099)	(0.352)	
$\mathbb{1}(\tilde{S} > S)D$	0.204***	-0.016	0.199
	(0.052)	(0.296)	(0.122)
$\mathbb{1}(\tilde{S} < S)D$	1.015***	0.979***	1.065***
	(0.056)	(0.153)	(0.086)
$\tilde{M}(1-M)D$	-1.449***	-1.700	-1.416**
	(0.378)	(1.586)	(0.565)
$(1-\tilde{M})MD$	0.128*	0.140	0.126
	(0.068)	(0.241)	(0.109)
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,795	2,281	6,255

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table. Note that the regression coefficients reported in the middle column are based on a logistic regression model estimated on data from May, June, July 2016 (pre-deployment) and December 2016, January and February 2017 (post-deployment period). The case outcome is coded as "1" if the case is accepted for services or connected to a closed case and an investigation is opened, as "0" if the case is not accepted for services.

*p<0.1; **p<0.05; ***p<0.01

Table 11: Comparison of regression results for the assessment of racial disparities in screen-in decisions on data relative to different time periods

	Dependent variable: call worker's decision		
	Time period considered:		
	Entire period	May, Jun, Jul 2016 and Dic 2016, Jan, Feb 2017	Post deployment
	(1)	(2)	(3)
A	0.426***	0.177	0.433*
	(0.105)	(0.340)	(0.234)
S	0.055***	0.065***	0.062***
	(0.005)	(0.012)	(0.014)
AS	-0.0004	0.014	-0.010
	(0.010)	(0.026)	(0.014)
D	-0.675***	-0.338	
	(0.239)	(0.612)	
AD	-0.035	-0.630	
	(0.214)	(0.870)	
SD	0.005	-0.044	
	(0.015)	(0.043)	
ASD	-0.007	0.060	
	(0.015)	(0.063)	
M	-0.169**	-0.436**	1.208***
	(0.083)	(0.189)	(0.142)
AM	-0.097	0.092	-0.348*
	(0.122)	(0.245)	(0.199)
MD	1.397***	1.621***	
	(0.169)	(0.489)	
AMD	-0.311	-0.134	
	(0.219)	(0.624)	
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,626	3,850	10,946

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table. Note that the regression coefficients reported in the middle column are based on a logistic regression model estimated on data from May, June, July 2016 (pre-deployment) and December 2016, January and February 2017 (post-deployment period).

*p<0.1; **p<0.05; ***p<0.01

Table 12: Comparison of regression results for the assessment of socioeconomic disparities in screen-in decisions on data relative to different time periods

	Dependent variable: call worker's decision		
	Time period considered:		
	Entire period	May, Jun, Jul 2016 and Dic 2016, Jan, Feb 2017	Post deployment
	(1)	(2)	(3)
A	-0.320*	-0.931**	-0.288
	(0.168)	(0.392)	(0.237)
S	0.054***	0.057***	0.059***
	(0.005)	(0.014)	(0.014)
AS	0.031**	0.077**	0.017
	(0.013)	(0.030)	(0.016)
D	-0.607***	-0.812*	
	(0.185)	(0.444)	
AD	-0.020	0.819	
	(0.220)	(0.831)	
SD	0.002	-0.005	
	(0.013)	(0.033)	
ASD	-0.012	-0.046	
	(0.016)	(0.057)	
M	-0.203***	-0.417**	1.006***
	(0.057)	(0.212)	(0.110)
AM	-0.118	-0.160	-0.088
	(0.113)	(0.350)	(0.150)
MD	1.205***	1.769***	
	(0.128)	(0.548)	
AMD	-0.003	-0.282	
	(0.221)	(0.685)	
Month FE	Yes	Yes	Yes
Call worker FE	Yes	Yes	Yes
Allegation type FE	Yes	Yes	Yes
Observations	23,626	3,850	10,946

Notes: Month of referral, call workers', and allegation types fixed effects (FE) included. Standard errors are clustered at the call workers' level. Intercepts coefficients are omitted from the table. Note that the regression coefficients reported in the middle column are based on a logistic regression model estimated on data from May, June, July 2016 (pre-deployment) and December 2016, January and February 2017 (post-deployment period).

*p<0.1; **p<0.05; ***p<0.01

8.3 Focal reweighting variable model diagnostics

In this section, we present the model diagnostics proposed by Buja et al. (2019) to detect and analyze the consequences of modeling misspecification. The diagnostics leverage the fact that, if the model is well specified, then its coefficients values do not change under arbitrary reweighting of the regressors distribution (see proposition 4.1 in Buja et al. (2019)). The type of diagnostic that we employ in this work is named by the authors "focal reweighting variable". Through this graphical device, we aim to understand how the estimates of the coefficients of interest may vary when we change the distribution of the regressors along a chosen variable, the so-called "reweighting variable". In particular, we wish to assess whether, for certain configurations of the regressors distribution, our conclusions based on the regression analysis would change, i.e., whether the direction of the estimated effect would change. Note that our analysis of Section 8.2, where we fit a regression only on data from the post-deployment period, represents one specific example of these diagnostics. In terms of reweighting variables, we consider the assessed score *S*, the identity of the call worker, and the month in which the referral occurred. To obtain the estimates of the coefficients of model (1) under the reweighting, we perform the following steps. First, we define a grid of "reweighting centers". For the assessed score *S*, we let the individual scores (1–20) be the reweighting centers and assign to the observations weights that are proportional to a Gaussian kernel with mean equal to the value of the reweighting center and standard deviation equal to 2. For the two categorical regressors, we consider all the values for which the dataset contains at least 100 observations. Then, for each reweighting center we obtain 100 sets of coefficients estimates via empirical bootstrap. In case of the assessed score, we resample from the data with probabilities given by the sampling weights described above and normalized to 1. The focal reweighting variable model diagnostics for month of referral, call worker's identity, and assessed score *S* are displayed in Figures 7, 8, and 9 respectively. We focus solely on the analysis of call workers' decisions as dependent variable.

We first focus on the coefficients estimates obtained with the month of the year as the reweighting variable. We estimate that the positive association of screen-in rates with the assessed score *S* is stronger during the summer months (June, July, August) compared to the rest of the year. By contrast, the estimates of the coefficient relative to *SD* are close to zero during the summer months but positive for the first six months of the year. It is possible that our model, despite the inclusion of allegation types, does not account for the different nature of the cases and of the reporters that are reported in the two periods of the year. For example, throughout the academic year many cases are referred by teachers, who are mandatory reporters. We observe that the screen-in rates for these months are 5–10% higher than in the other months of the year, but the cases show similar risk levels. The estimates of the coefficients for the other terms are stable across the months of the year, with the only notable exception being the month of January. For the referrals occurring

during this month, the estimated coefficients for $\mathbb{1}(\tilde{S} > S)D$ and $\mathbb{1}(\tilde{S} < S)D$ are large.

We then turn to the diagnostics in Figure 8 where the identity of the call worker represents the reweighting variable. We observe that the models fitted on three of the call workers' data, which are denoted as "CW1", "CW2", and "CW6", show larger coefficients estimates for S. One interpretation of this result is that before deployment these two workers relied more heavily on the same information that is used by the risk assessment tool than on the information that is communicated in the call. This hypothesis also aligns with the fact that the same set of workers are characterized by coefficients estimates for SD that are close to zero, which suggests that changes in the assessed scores did not impact their decisions differently in the two periods. For the other workers, the estimates of the coefficients for SD are positive. We observe that the mandatory flag increased the likelihood of screen-in (i.e., $MD > 0$), and particularly for call worker 6, who was unlikely to screen-in these cases before the deployment of the tool.

Next, we analyze the results of the diagnostics where the assessed score S represents the reweighting variable, which are displayed in Figure 9. We observe that the association between S and screen-in decisions is positive and strong for low values of S, but the corresponding coefficient estimate is close to zero for large values of S. An analogous pattern can be observed for the (unconditional) exploratory analysis presented in Figure 3. The coefficient estimates relative to SD reveal that in the post-deployment period increases in the assessed score S positively impact the likelihood of screen-in only for large values of S. This finding suggests that call workers may treat algorithmic recommendations in low and high risk cases differently. Looking at the two effects together suggests that, in the post-deployment period alone, increases in the score led to similar changes in the likelihood of screen-in across all values of the score. In addition, in light of the strong association between screen-in rates and the assessed scores for cases characterized by low values of the score in the pre-deployment period, it is perhaps not surprising that the deployment of the tool only impacted the alignment of screen-in decisions and assessed score for higher scoring cases. Turning to the diagnostics of the remaining terms, we note that the estimates of the coefficients relative to M and MD do not change under the reweighting, and they are close to those that we find in the model fitted on data without reweighting. The estimates of the coefficient relative to $\mathbb{1}(\tilde{S} > S)D$ are largest for low values the assessed score S, which suggests that the overestimation had the greatest impact on screen-in when the estimate risk was low. The positive coefficients estimates for $\mathbb{1}(\tilde{S} > S)D$ in case of high-risk cases are consistent with our finding that call workers avoid errors of omission.

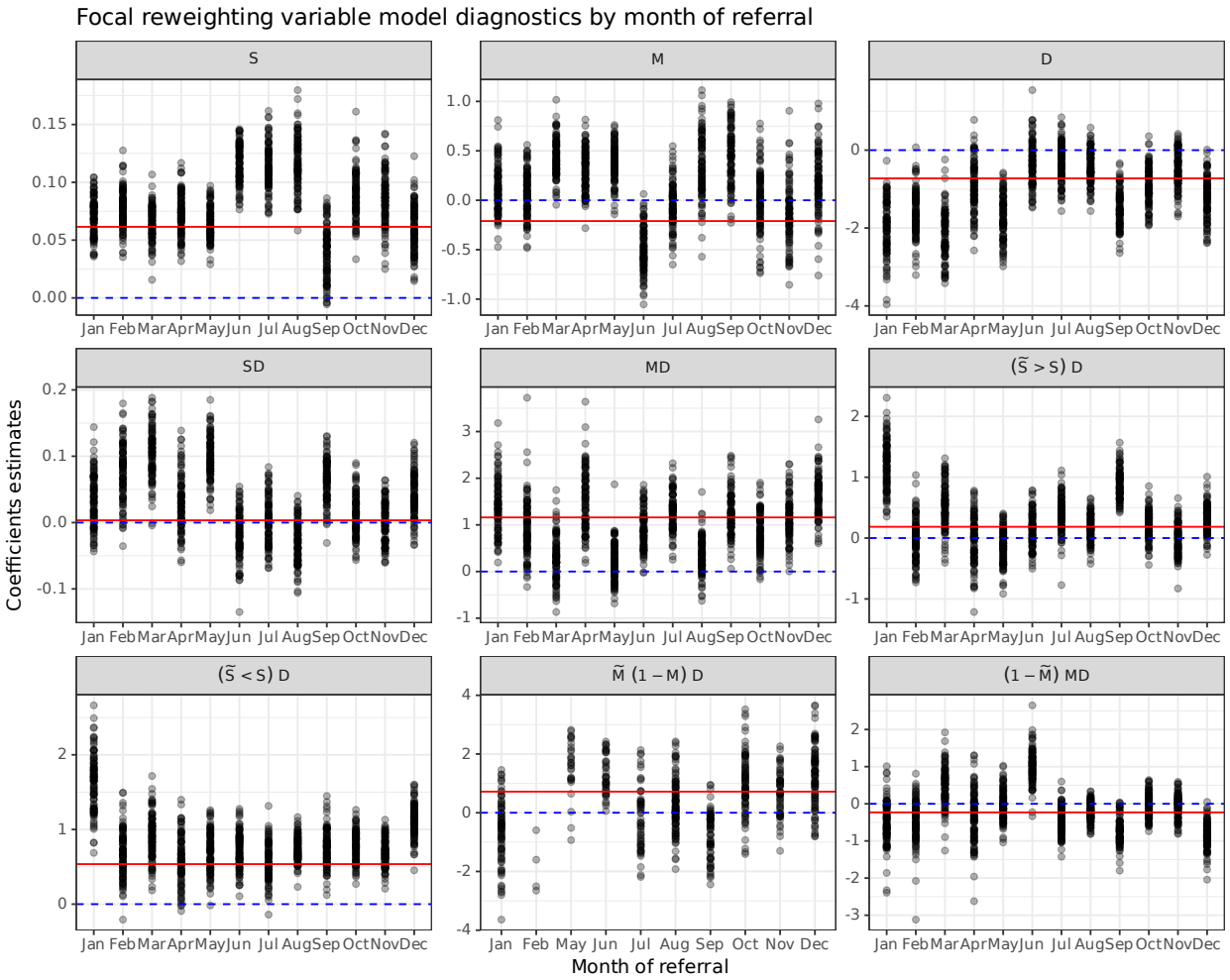

Figure 7: Focal reweighting variable model diagnostics for the coefficients of model (1) estimated through logistic regression to predict screen-in decisions, considering the month of referral as the reweighting variable. Each of the black dots corresponds to one of 100 bootstrap estimates of the regression coefficient indicated in the panel's title. The red line corresponds to the coefficients estimates presented in Table 2. The blue dashed horizontal lines are centered at 0. On the horizontal axis are displayed the values of the reweighting variable. For visualization purposes, we dropped the estimates that fell outside of the range $[-5, 5]$.

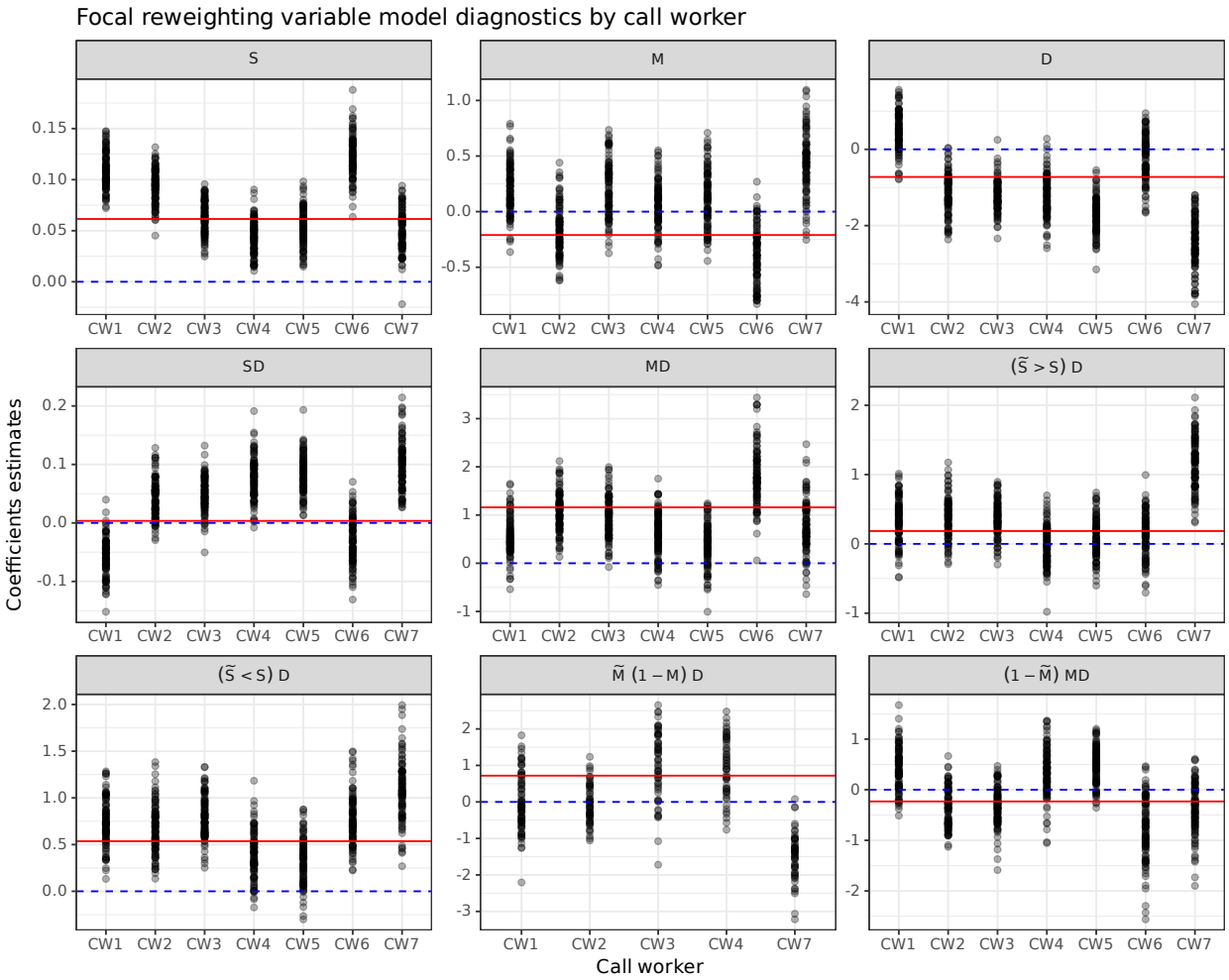

Figure 8: Focal reweighting variable model diagnostics for the coefficients of model (1) estimated through logistic regression to predict screen-in decisions, considering the call worker's identity as the reweighting variable. We only consider call workers that have handled at least 500 cases in both the pre- and the post-deployment periods. Each of the black dots corresponds to one of 100 bootstrap estimates of the regression coefficient indicated in the panel's title. The red line corresponds to the coefficients estimates presented in Table 2. The blue dashed horizontal lines are centered at 0. On the horizontal axis are displayed the values of the reweighting variable. For visualization purposes, we dropped the estimates that fell outside of the range $[-5, 5]$.

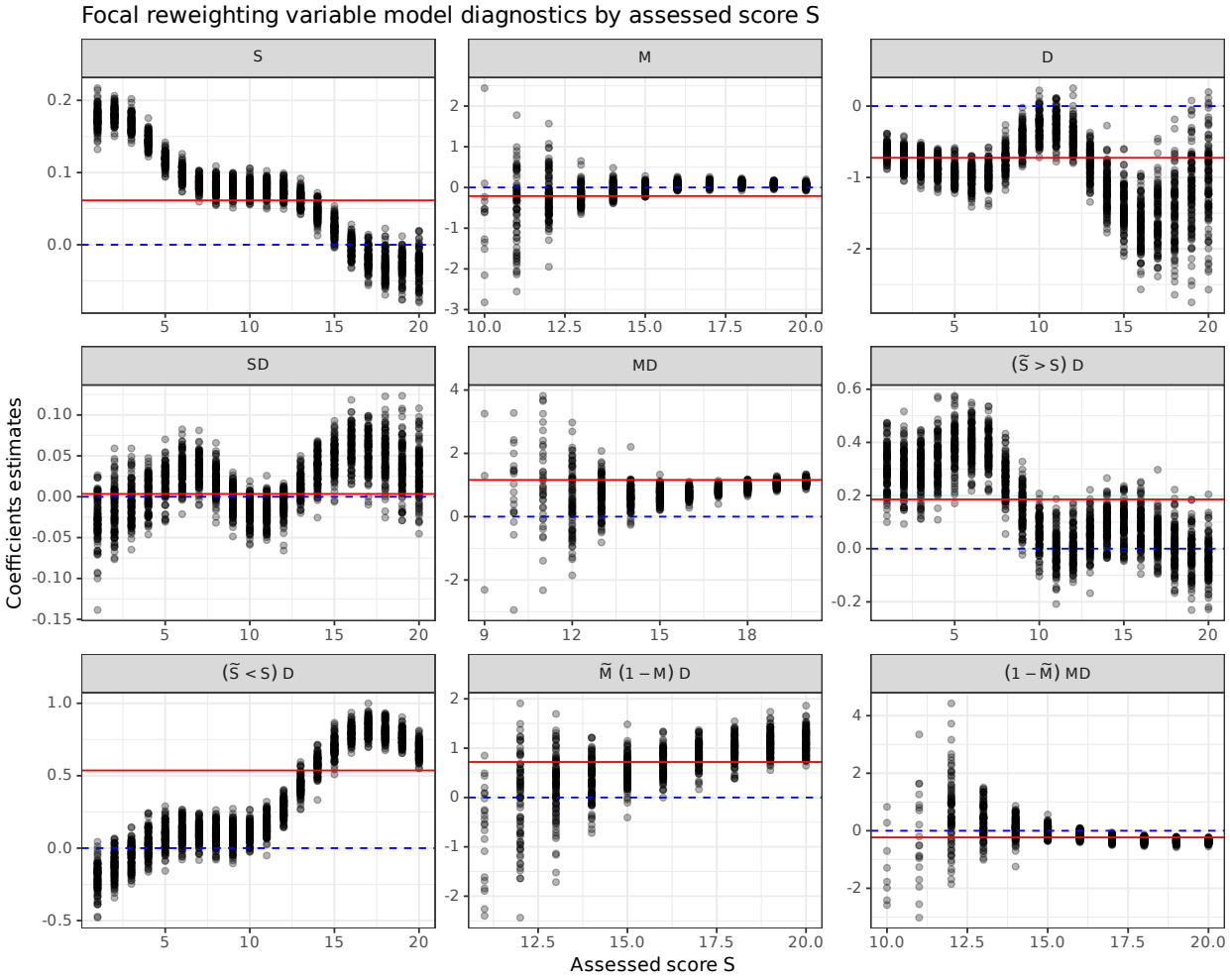

Figure 9: Focal reweighting variable model diagnostics for the coefficients of model (1) estimated through logistic regression to predict screen-in decisions, considering the assessed score as the reweighting variable. Each of the black dots corresponds to one of 100 bootstrap estimates of the regression coefficient indicated in the panel's title. The red line corresponds to the coefficients estimates presented in Table 2. The blue dashed line is an horizontal line centered at 0. On the horizontal axis are displayed the values of the reweighting variable. For visualization purposes, we dropped the estimates that fell outside of the range $[-5, 5]$.

8.4 Alternative modeling specifications

We assess whether our results are affected by an alternative modeling specification. Instead of model (1), we assume that the decision to screen in a case i made by call worker c in month t is described by the following equation,

$$\begin{aligned} \textit{screen-in}_{itc} \sim \beta_0 + \beta_1 f_S(S_i) + \beta_2 M_i + [\beta_3 + \beta_4 f_{SD}(S_i) + \beta_5 M_i] * D_i \\ + [f_{(\tilde{S}-S)D}(\tilde{S}_i - S_i) + \beta_8 \tilde{M}_i * (1 - M_i) + \beta_9 (1 - \tilde{M}_i) * M] * D_i + \psi_t + \mu_c + \theta_i \quad (2) \end{aligned}$$

where f_S, f_{SD}, and $f_{\tilde{S}-S}$ are smooth functions. Note that, compared to model (1), here we make weaker assumptions on the dependence of screen-in rates on the assessed score S, i.e., we only assume that the functions f are smooth rather than linear. We fit model specification (2) via a generalized additive model (GAM) with logit link and penalized regression splines for the estimation of f. We employ the same model specification to predict the likelihood that screened-in cases will be accepted for service, and for the assessment of disparities in screen-in decisions. The coefficients estimates produced by the first two models are reported in Table 13, while those relative to the models containing the interaction terms with the sensitive attributes are reported in Table 14. The estimates of the smooth functions are shown in Figure 10 and 11.

Let's first focus on the association between screen-in rates and the assessed score S. On the one hand, in the top left panel of Figure 10 we observe that the probability that a case will be screened in increases with the assessed score even before deployment, quite substantially for low values of the score. This is consistent with our findings from the reweighting diagnostic plots of the assessed score in Figure 9, which revealed an analogous pattern. The middle panel shows that, after the deployment of the tool, increases in the values of the scores have a positive impact on the likelihood of screen-in for high-risk cases, compared to the pre-deployment period (*SD*). Again, this is consistent with our results from the diagnostic plots. As in our main regression model, the coefficients estimates in Table 13 for M and MD show that the presence of the mandatory flag had a positive and large impact on the likelihood of the case being screened in. We finally turn to the misestimation of the assessed scores. The top right panel of Figure 10 reveals that, as the assessed and shown scores diverge, the probability that the case will be screened in increases. Overestimation appears to increase significantly the likelihood of screen-in, but the function is not precisely estimated due to the small sample size. In case of underestimation, the likelihood of screen-in appears to increase as well for $\tilde{S} - S \in [-5, 0]$, and then becomes constant. As we have discussed in the main body of the paper, this should not be surprising because the glitch was non-random and S may be an imperfect and non-comprehensive estimate of the underlying level of risk. Lastly, the effect of requiring screen-in's for cases that are assessed as not being at high risk of out-of-home placement is positive but not statistically significant. The direction and size of the

coefficients estimates relative to the misestimation of the mandatory flag in Table 13 are similar to those in our main model. The regression results for the outcomes and the assessment of disparities are similar to those delivered by the regression for model specification (1).

Table 13: Regression results for generalized additive model (GAM) estimating the impact of the tool's recommendations on call workers' decision-making and case outcomes.

	Dependent variable:	
	Call worker's decision	Case outcome
	(1)	(2)
D	-0.592***	-2.220***
	(0.189)	(0.268)
M	0.052	0.339***
	(0.086)	(0.106)
MD	1.152***	0.411***
	(0.137)	(0.149)
$\tilde{M}(1-M)D$	0.865**	-1.283**
	(0.401)	(0.636)
$(1-\tilde{M})MD$	-0.373***	-0.048
	(0.118)	(0.113)
Month FE	Yes	Yes
Call worker FE	Yes	Yes
Allegation type FE	Yes	Yes
Observations	23,626	14,263

Note: Regression coefficients for GAM to estimate the coefficients values in model specification (2). Standard errors are reported within parentheses.

*p<0.1; **p<0.05; ***p<0.01

Table 14: Regression results for generalized additive model (GAM) for the assessment of racial and socioeconomic disparities in decision-making.

	Dependent variable: Screen-in decision	
	Demographic attribute A	
	A=1 if at least one child in referral is black	A=1 if % families living below poverty threshold $\geq 20\%$
	(1)	(2)
A	0.311**	-0.447***
	(0.137)	(0.167)
M	0.109	0.080
	(0.117)	(0.109)
AM	-0.134	-0.124
	(0.144)	(0.148)
D	-0.536**	-0.514**
	(0.243)	(0.253)
AD	-0.359	-0.147
	(0.484)	(0.465)
MD	1.366***	1.146***
	(0.216)	(0.192)
AMD	-0.340	0.068
	(0.279)	(0.274)
Month FE	Yes	Yes
Call worker FE	Yes	Yes
Allegation type FE	Yes	Yes
Observations	23,626	23,626

Note: Regression coefficients for GAM to estimate the coefficients values in model specification (2). The dependent variable in the regression is whether the case was screened in. Standard errors are reported within parentheses. Terms relative to the shown scores are also included in the model but the corresponding coefficients estimates are omitted from the table.

*p<0.1; **p<0.05; ***p<0.01

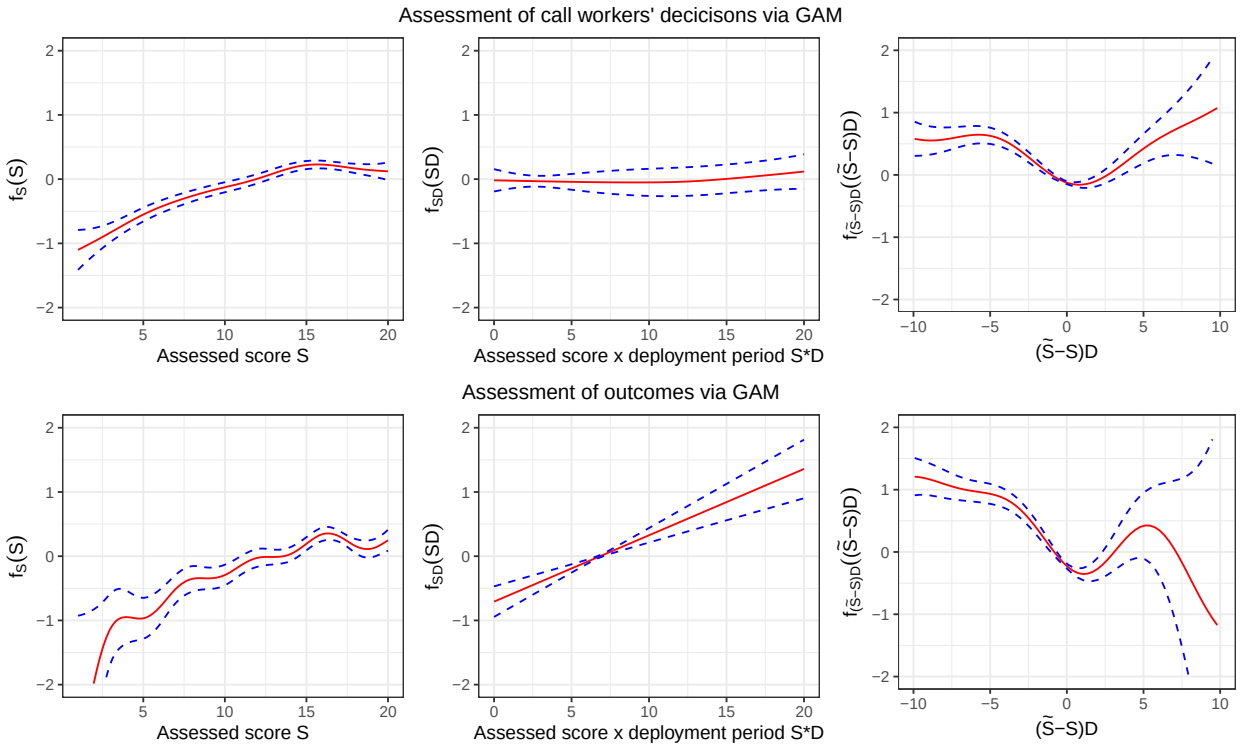

Figure 10: The plots show the estimates of the smooth functions in model specification (2) estimated using a GAM with regression splines to predict screen-in decisions (top panel) and case outcomes (bottom panel). From the left to right, the estimates of f_S, f_{SD}, and $f_{(\tilde{S}-S)D}$ are shown. Confidence bands correspond to two standard deviations.

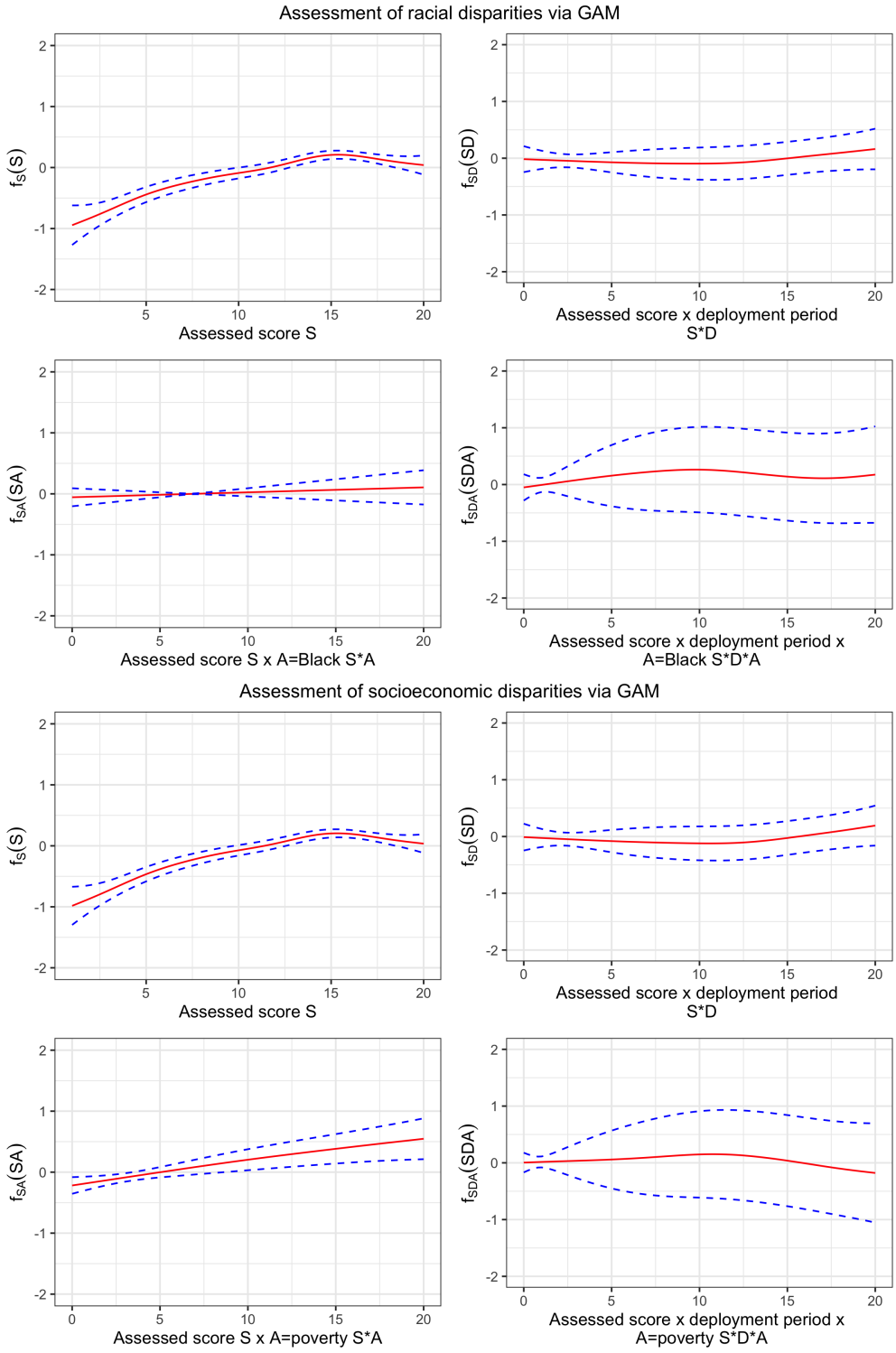

Figure 11: The plots show the estimates of the smooth functions in model specification (2) estimated using a GAM with regression splines for the assessment of racial (top panels) and socioeconomic disparities (bottom panel) in screen-in decisions. From the left to right, the estimates of f_S and f_{SD} (in the bottom, f_S and f_{SDA} respectively) are shown. Confidence bands correspond to two standard deviations.