Summarizing Recent Research on Predictive Risk Models in Child Welfare

April 2024

INTRODUCTION

Since 2016, the Allegheny County Department of Human Services (DHS) in Pennsylvania has utilized the Allegheny Family Screening Tool (AFST), an algorithm designed to assist child welfare call screening caseworkers in their assessment of general protective service (GPS) referrals regarding potential child maltreatment. Read more about the AFST here. In compliance with Pennsylvania law, GPS referrals can either be screened in for investigation or dismissed at the discretion of the county. The AFST employs administrative data trained to estimate the probability of safety issues so significant that a judge will order a child to be removed from their home within 2 years of the report of suspected abuse or neglect. This information is used alongside other information in the referral process to support call screening case workers in their decision-making process.

The primary objectives of adopting the AFST are to improve the quality of decision making among call screeners by: (1) reducing the frequency of failure to detect conditions that may contribute to future maltreatment; (2) reducing intrusive investigations where there is low likelihood that maltreatment exists; and (3) increasing the consistency of call screeners in their use of data in deciding which GPS reports to investigate.

Although DHS's child welfare call screeners always maintain the discretion to use the AFST risk score as part of a holistic consideration of risk, the inclusion of a data-driven model in decision making has been controversial. It has also attracted extensive research on the tool's effectiveness and how it is perceived by stakeholders. In this paper, we review the research evidence on algorithms in child welfare with a focus on causal analyses of the AFST and similar predictive risk models (PRMs). We first discuss the tools' effects on child welfare decisions. Next, we consider perceived effects, which are often quite different from the actual effects. Finally, we address the need to bridge the gap between perception and reality to reduce concerns that may be unfounded and optimize the utility of the AFST and related tools.

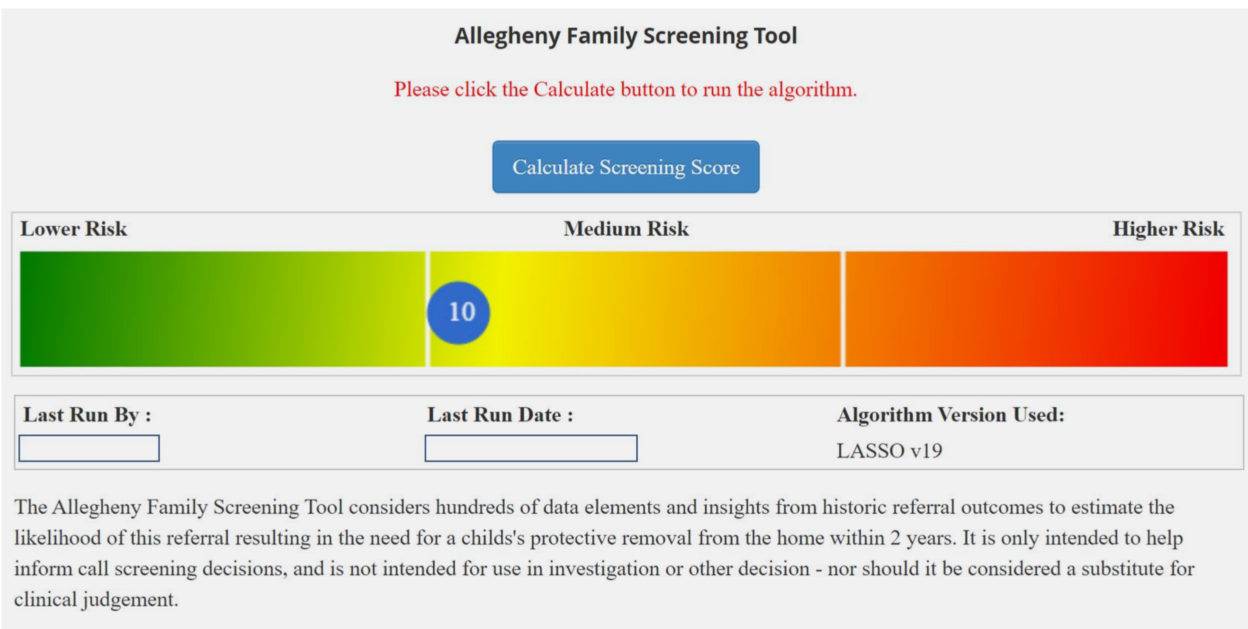
HOW THE AFST WORKS

The initial development of the AFST relied on a PRM constructed from the analysis of nearly 77,000 referrals to child welfare from 2010 to 2014. The analysis used 800 available administrative data features (e.g., prior referrals to child welfare) available at the time referrals were screened to predict downstream outcomes for children. The first version of the algorithm identified features that were most strongly predictive of re-referral and removal; subsequent versions eliminated the re-referral outcome and predicted removal alone. The model predicting re-referrals lacked face validity because high scores could reflect outcomes other than serious abuse and neglect.

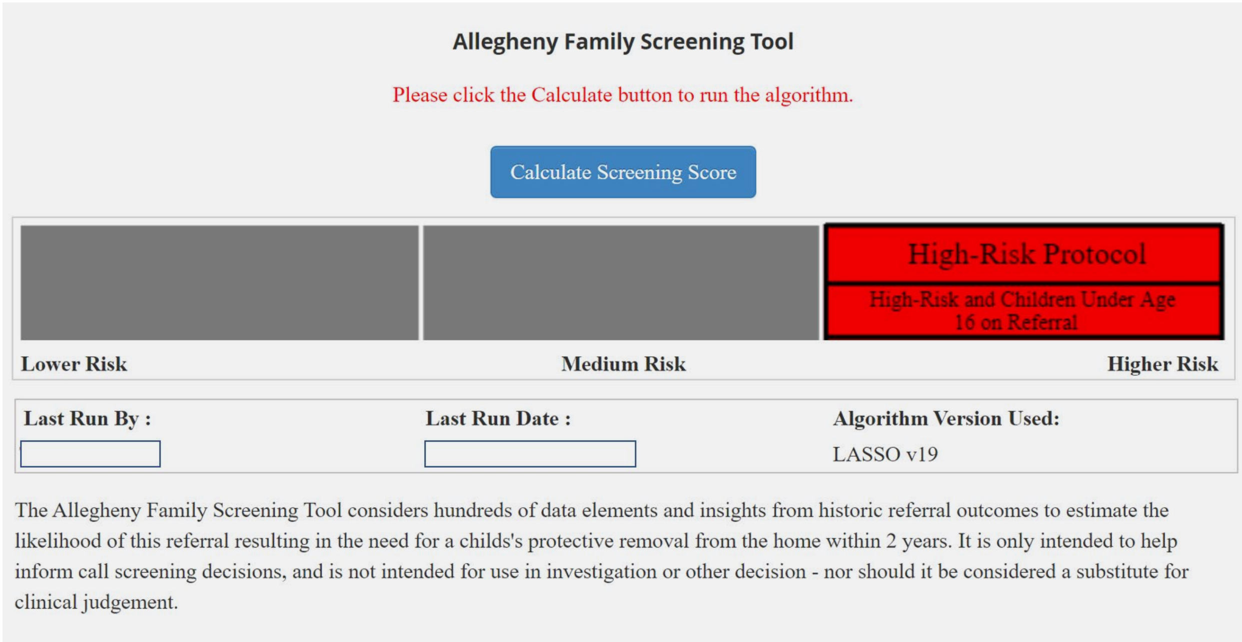
Child maltreatment referrals start with a report of suspected abuse or neglect from a source detailing information about the incident and allegations, the alleged victim child(ren), alleged perpetrator(s) and associated individuals. The call screening unit validates this information with the reporting source and against official data sources and adds any additional relevant individuals in the household. Names are cross-checked against DHS's Data Warehouse, and, once the referral has complete data on clients, risk, safety and allegations, call screeners generate a referral risk score from 1 to 20 via the AFST.

Accompanying policies relevant to using the AFST and how it affects the screening decision have been developed and implemented. These policies have changed over time, driven by leadership, but also reflecting input from frontline staff and our ongoing analyses of data. Currently, when the AFST score is 18–20 and at least one child younger than 16 years of age is present in the household, the call screener is alerted that the report has been designated as "high-risk protocol" rather than receiving a numeric score. Likewise, when the AFST score is 1 to 12 and all children are aged 7 years or older, the call screener is notified that the report has been classified as "low-risk protocol." These protocols default the screening decision in the application to screen-in (if high risk) or screen-out (if low risk), although intake supervisors have discretion to override the defaulted selections and do so regularly. **Figure 1** displays 3 call screener AFST experiences: no protocol, high-risk protocol, and low-risk protocol.

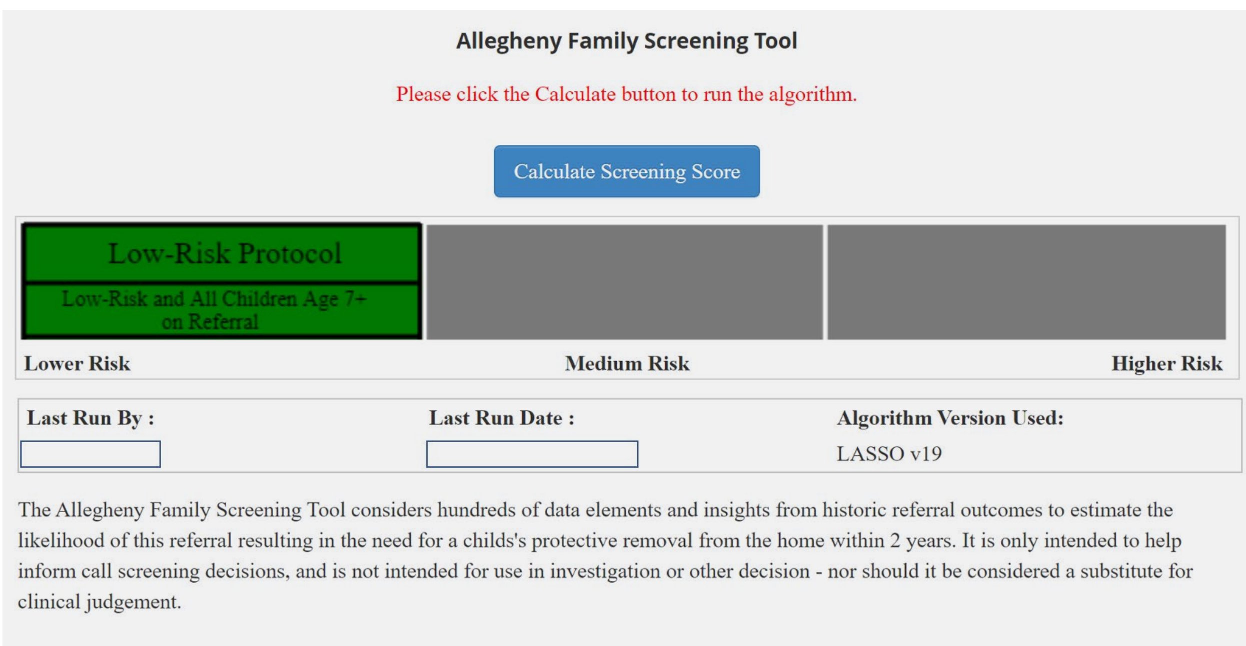
FIGURE 1. **Screener Display of AFST Scores and Protocols**
(a) No Protocol



(b) High-Risk Protocol



(c) Low-Risk Protocol



In all cases, screeners and their supervisors can use the information as guidance but have decision-making discretion. After a decision to investigate has been made, neither the assigned investigator nor any caseworker ever sees the AFST score, so it does not affect ultimate assessments of maltreatment or decisions to open a case or remove a child from their home. The AFST is designed to complement, not replace, human decision-making at intake where there are more than 80 decisions a day.

Since 2016, when Allegheny County implemented the AFST to support child welfare screening decisions, similar systems have been introduced in various jurisdictions across the country. Today, decision aids are being used in several U.S. states and counties, including Los Angeles County, California and Larimer, Arapahoe and Douglas counties, Colorado.

ALTERNATIVES TO PRMS IN CHILD WELFARE AGENCIES

To understand the impacts of predictive risk algorithms in child welfare, it is crucial to consider other methods by which maltreatment screening decisions are made. In the research findings, these other methods represent the status quo. The AFST falls into a class of tools conceptualized as "risk assessment tools." Most existing risk assessment tools used by child welfare agencies are "operator driven" or manual in nature, including the Structured Decision Making (SDM) risk tool—a version of which is implemented in approximately two thirds of states. For states that have not implemented SDM, most have developed structured checklists for assessing risk. Manual tools require information to be gathered and entered by case workers; these tools have several drawbacks, including that they: (1) must be brief enough for workers to complete quickly, so they may fail to incorporate the full range of factors relevant to understanding differences in risk; (2) frequently contain subjective data elements that can lead to biased assessments; and (3) are prone to errors in how they are filled out and completed. In addition, almost none of these tools are locally validated, which means they don't account for the differences in context and needs that exist in a specific location (i.e., Allegheny County).

Prior to implementation of the AFST, Allegheny County did not use any assessment tool. Call screening case workers could access and use historical and cross-sector administrative data on individuals associated with a report of child abuse or neglect. These case workers were required to review relevant information before a screen-in/screen-out decision was made. However, call screeners could not efficiently access, review and make meaning of all available records. In addition, there was inconsistency in decisions by different call screeners. As a result, call screening case workers screened-in almost half of the lowest-risk-cases and screened out one in four of the highest risk ones.

In 2016, the Children's Data Network at the University of Southern California received a 2-year research grant to develop a proof-of-concept PRM for California modeled closely on the work in Allegheny County. The goal was to assess whether a PRM built exclusively from child protection records could advance the state's ongoing efforts to ensure consistent, equitable and sound decision-making at entry to the system.

As part of this research project, attempts were made to not only document the classification accuracy possible through a PRM such as the AFST, but also assess how a PRM might improve child welfare practice and supervision beyond what exists today. This assessment included comparing the PRM score with the SDM family risk assessment tool already in use in California. This was critical to the project because it allowed us to document when PRMs could generate additional (or even different) insights about children's risk trajectories. Importantly, it also helped illustrate how those insights from an automated model, on the margins, might ultimately help, harm, or make no difference at all.

Findings indicated that the PRM was more accurate than the SDM risk tool in identifying children who would have chronic or intensive involvement with the child protection system. In California, researchers compared the accuracy of risk classification decisions from a SDM family risk assessment with PRM results. They found that the top 7% of children identified by the algorithm as "very high risk" were more likely to be placed in foster care than an equivalent-sized group identified as very high risk through the SDM risk assessment tool (74.9% vs. 50.3%), more likely to have 3 additional referrals than the SDM-identified highest risk group (29.5% vs. 12.7%), and more likely to have a substantiated allegation of abuse (33.0% vs. 12.8%). These results suggest that the riskiest referrals flagged by the PRM had higher rates of adverse outcomes than the riskiest referrals identified through the SDM (Putnam-Hornstein et al., 2018). Other white papers, audits and research analyses have also documented serious shortcomings of manual risk assessment tools in terms of both accuracy and implementation (e.g., Bosk, 2018; California State Auditor, 2019; Gillingham, 2011). Reinforcing this point, a recent scoping review found a very thin evidence base for the use of most risk assessment tools (McNellan et al., 2022).

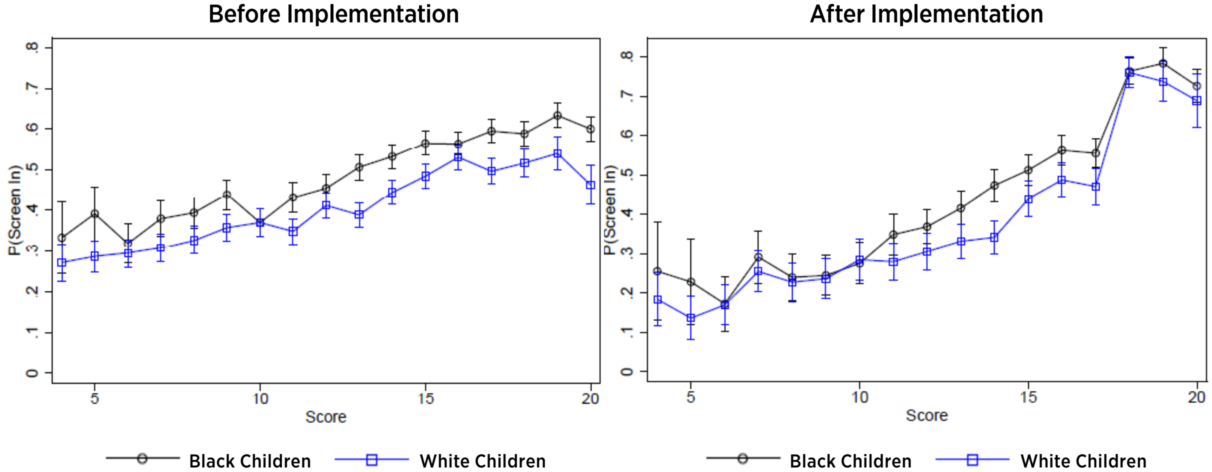
HOW THE AFST IS BEING USED AND ITS IMPACT

The following sections detail the main conclusions derived from recent research on the AFST and other predictive risk models implemented by child welfare agencies.

The AFST changed the composition of investigated referrals.

Rittenhouse et al. (2023) compared screen-in (i.e., investigation) rates by risk score before and after the deployment of the AFST. Before AFST deployment, scores reflect those that would have been shown to call screeners had the tool been launched. The post-deployment period reflects decisions that were made in the context of an AFST score. **Figure 2** shows that the deployment of the AFST reduced the probability of screen-in decisions for children in the lowest risk referrals and increased the probability of screen-in decisions among children in the highest risk categories. The introduction of the AFST also reduced the racial gap in screen-in rates among higher risk referrals. Prior to implementation, Black children were screened at higher rates than White children, especially for those identified as higher risk. After implementation, the gaps in screen-in rates by race were reduced.

Figure 2. Screen-In Rates by AFST Score Before and After AFST Deployment for White and Black Children



Note: The left graph displays AFST scores prior to AFST deployment (January 2013 through June 2016). The right graph displays AFST V3 scores (July 2019 through December 2020).

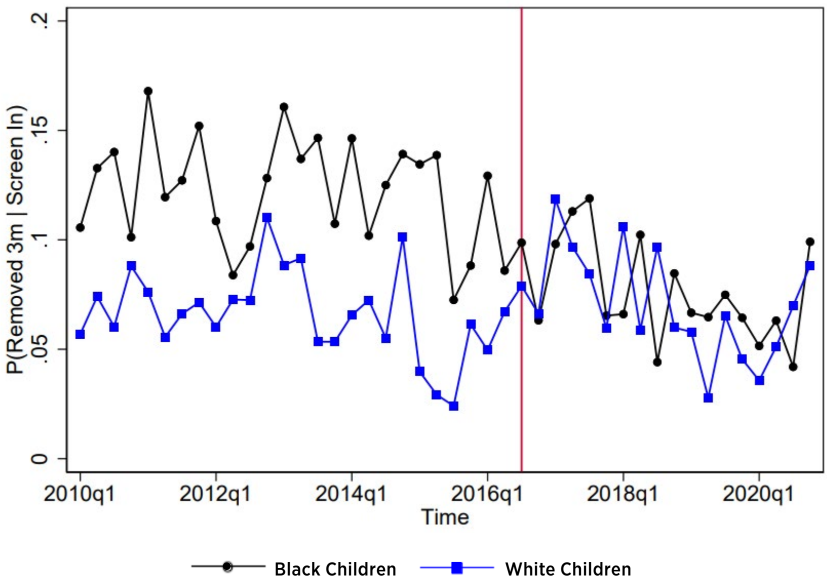
The AFST impact on case openings is inconclusive

Among referrals screened in for investigation, the percentage of case openings increased (Rittenhouse et al., 2023). The most plausible explanation for this change is that the AFST helped screeners more effectively identify referrals that required further intervention. Goldhaber-Fiebert and Prince (2023) found a 20% decrease in case-opening rates for GPS referrals; however, they also noted that CPS referrals saw a similar relative reduction in case openings, underscoring the uncertainty about interpreting their finding as being driven by the AFST rather than environmental or policy changes. Although other results across studies tended to show similar patterns, the case-opening outcomes might be more dependent on choice of analysis timeframe and methodology.

The AFST is reducing, not increasing, racial disparities.

Although **Figure 2** provides a compelling picture of the changes in screened-in referrals after the introduction of the AFST, the results cannot be interpreted as causal unless imposing strong assumptions implicit in before-and-after comparisons. To address these assumptions, Rittenhouse et al. (2023) compared child welfare-related outcomes for Black and White children before and after the algorithm's deployment, using the evolution in the Black–White gap among referrals unaffected by the AFST as a comparison. They found that use of the algorithm reduced racial screening disparities across AFST scores, although the size and precision of the reduction varied. The AFST reduced the racial disparity in screen-in rates for the highest risk referrals by 83%, from 10.6% to 1.8%. The difference based on race in case-opening rates after initial investigation also narrowed. Additionally, the researchers estimated that the AFST reduced the Black–White gap in removal rates of screened-in referrals by 73%, from 4.3% to 1.2% (**Figure 3**).

FIGURE 3. **Removal Rates among Screened-In GPS Referrals by Race Before and After AFST Deployment**



This finding was further reinforced by the evaluation of the AFST by Goldhaber-Fiebert and Prince (2022), who used DHS data but a different methodology than that of Rittenhouse et al. (2023) and Grimon and Mills (2022), who conducted a randomized controlled trial of a PRM used for call screening in Larimer County, Colorado. **Table 1** compares results across these three studies, finding similar patterns in the direction and magnitude of the results.

TABLE 1. **Impact of Child Welfare Algorithms on Racial Gaps**

		REDUCTION IN BLACK–WHITE GAP			
STUDY	LOCATION	SCREENED IN	ACCEPTED FOR SERVICE	HOME REMOVAL	HOSPITALIZATION
Goldhaber-Fiebert and Prince (2023)	Allegheny County	32%	92%	100%	N/A
Rittenhouse et al. (2023)	Allegheny County	46%	91%	73%	N/A
Grimon and Mills (2022)	Larimer County	50%	N/A	N/A	56%

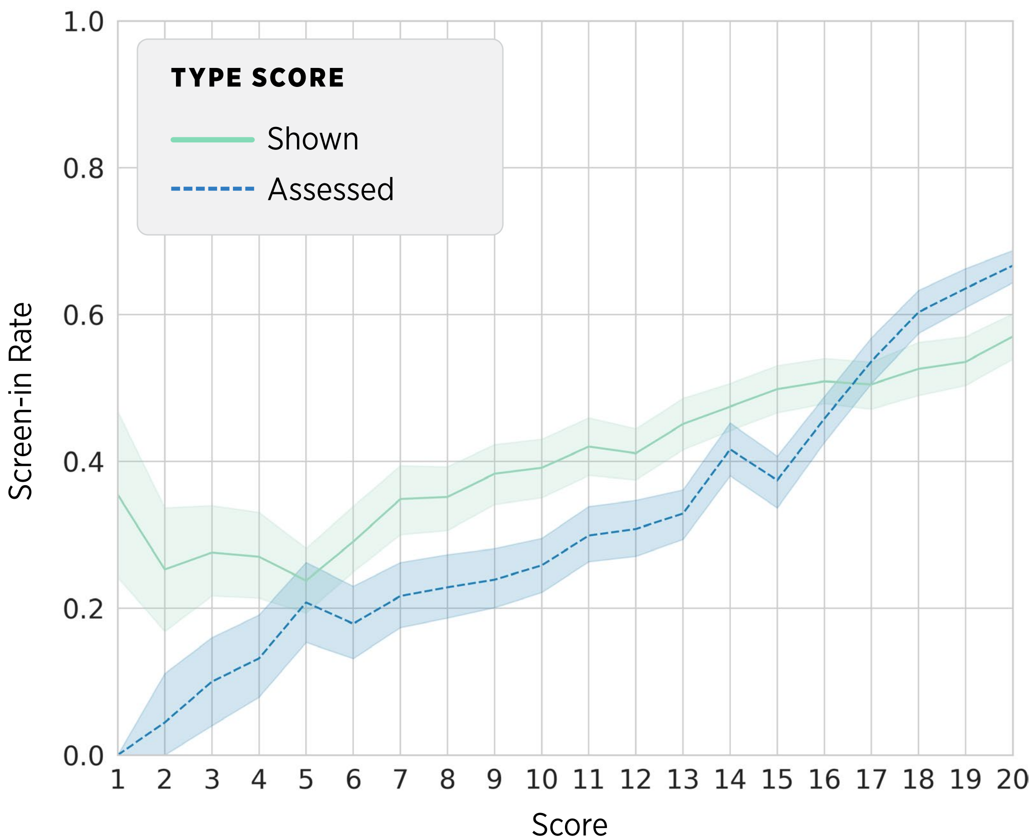
Note: All figures are relative to the baseline Black–White racial gap for each outcome. Screen-in and service-acceptance rates for Allegheny County studies refer to GPS referrals. Removals for Goldhaber-Fiebert and Prince include removals up to 180 days after referral; removals in Rittenhouse et al. include removals within 3 months of the case opening conditional on screening in. Hospitalization refers to child injury hospitalization.

Screeners use the algorithm but with caution.

De-Arteaga et al. (2020) compared data from 2016 and 2017, when the AFST was newly implemented, to data from 2015–2016 before its introduction. They found that screeners' decisions were more closely aligned with 2-year risk of removal than they were prior to the AFST's introduction. In other words, call screening case workers revised their decisions consistent with the algorithm's recommendations.

However, De-Arteaga et al. (2020) also demonstrated that screeners did not blindly follow the algorithm. During the first year of the AFST, a system glitch caused some risk scores to be reported incorrectly. Data analysis showed that screeners' decisions were more consistent with what the risk scores *should have been* than with the glitch-affected scores shown to them. This pattern is clearly displayed in **Figure 4**, which shows screen-in rates plotted by glitch-affected scores and true scores. These results indicate DHS staff's ability/tendency to treat the AFST as a helpful source of additional information without becoming overly reliant on it.

FIGURE 4. **Screen-In Rates by Shown Score (Affected by Glitch) and Assessed Score (Correct)**



Kawakami et al. (2022) conducted extensive observations of and interviews with DHS call screeners and concluded that they used qualitative details of the allegations to construct causal narratives regarding each case, which informed their decision to rely on or overrule the AFST risk score. Similarly, Fitzpatrick et al. (2023) in Colorado observed a tendency among staff members to lean less on algorithmic data when they had access to higher-quality information from other sources, highlighting a nuanced approach to the use of these tools. Despite this, Fitzpatrick et al. also estimated that the screening algorithm reduced processing time per referral by approximately 5%, suggesting an overall enhancement in the efficiency of investigation decisions with algorithmic decision-making tools.

These findings indicate that the application of the AFST has proceeded in a prudent manner, without overriding human participation, and has improved decision making and outcomes in DHS child protective services.

HOW CALL SCREENERS THINK THE AFST IS BEING USED

Why, then, is the AFST still somewhat controversial? Certainly, fear of the "black box" of mysterious algorithmic processes and risks of data misuse play a role. These concerns are understandable. In addition, however, various stakeholders have exhibited skeptical reactions ranging from misunderstanding to distrust.

Kawakami et al. (2022), in their interviews with screeners, found that those responsible for using the AFST in their decision making did not know what data points were used or how they were weighted in calculating AFST scores. Though trainings on the tool were provided to call screening case workers, staff had little understanding of how the AFST calculates its scores. The DHS staff resorted to creative approaches, such as rerunning the algorithm with slightly different data, to try to figure out the AFST's calculation process. One screener asked a researcher to explain the system to them. In the absence of better information, workers drew uninformed conclusions, not always accurate, about how the tool worked.

According to Kawakami et al. (2022), some screeners expressed feeling pressured to follow the AFST's recommendations or reluctance to put in the extra work necessary to disagree with its scores. They also felt that their feedback on the AFST was discounted. These experiences, along with their lack of detailed knowledge about the AFST, undermined their trust in the tool.

In an earlier study of the AFST, Chouldechova et al. (2018) found similar skepticism. In this study, screeners were asked to rate each case from 1 (*low*) to 3 (*high*) based on two factors: level of risk and the severity of safety threat to the potential victim. The two ratings were then multiplied to create the member's risk and safety score. Chouldechova et al. (2018) reported that screeners' decisions were more closely aligned with their ratings than with the AFST score. Moreover, supervisors overrode about a quarter of mandatory "screen-in" decisions (scores of 18 or higher at that time). The researchers theorized that lack of knowledge of why the algorithm was identifying certain cases as high risk was one factor in screeners' reluctance to trust the ratings. Screeners' reluctance must also be considered alongside the large research literature showing biases and errors in human judgment (Kahneman and Tversky, 1974; Bordalo et al., 2016; Kahneman, Sibony, and Sunstein, 2021).

One qualitative study captured perspectives on the use of PRMs from other stakeholder groups. Stapleton et al. (2022) held seven discussion groups with 35 participants, including parents, attorneys, social workers, and nonprofit agency administrators. Overall, the responses were negative; they reflected concerns about the AFST that did not appear justified. Some of the main issues Stapleton et al. reported include:

- PRMs create a perception of widening surveillance that reinforces child welfare agencies' function as punitive rather than providing support for families.
- PRMs perpetuate existing biases and racism (however, participants were not shown evidence on AFST and other algorithms reducing racial gaps and some participant comments were related to the child welfare system overall rather than PRMs specifically).
- No demographic factors or ZIP codes should be used in PRMs.
- The money spent on PRMs would be better used if given to families.
- Affected families should be more involved in decisions on how PRMs and similar technologies are used.

Regardless of whether these and other objections are applicable to Allegheny County, they highlight the challenges involved in assuring the public that algorithms will not be misused in decisions with great significance for families.

HOW TO MOVE FORWARD

The research evidence suggests that the AFST has improved the overall quality of screening decisions and reduced racial disparities, but also that the people affected—especially the screeners who interact most closely with it—may have opportunities to use predictive models more effectively. As similar tools become more widespread, users' general comfort level with them may rise. On the other hand, controversies surrounding new artificial intelligence innovations such as large language models may evoke stronger resistance.

Allegheny County will pursue further advances in shaping the human-algorithm interaction to improve decision making. Future tasks include:

- Building an onboarding training module that covers the science and implementation of the AFST.
- Providing informed, tailored feedback and guidance to call screeners on situations in which they override the algorithm in ways that may either improve or weaken performance.
- Modifying the presentation of risk scores to improve intuition and understanding of risk. One possibility is to present multiple risk scores related to various possible outcomes rather than a single score.
- Developing evaluations of actual impact, such as the effect of algorithm use on child safety and other outcomes of practical interest.

The most important principle in any technology implementation is transparency. Particularly when an algorithm is too complex to explain to affected stakeholders, maintaining public confidence in the tool's reliability—and vigilance by the people involved in preventing misuse—becomes crucial. DHS will continue to update the public on algorithms at the Department and leverage current research to shape the implementation of these tools in the field.

REFERENCES

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131 (4): 1753-1794.

Bosk, Emily Adlin. "What Counts? Quantification, Worker Judgment, and Divergence in Child Welfare Decision Making." *Human Service Organizations: Management, Leadership & Governance* 42, no. 2 (2018): 205–24. https://doi.org/10.1080/23303131.2017.1422068.

California State Auditor. *Los Angeles County Department of Children and Family Services: It Has Not Adequately Ensured the Health and Safety of All Children in Its Care*. Report 2018-126. May 2019. https://www.bsa.ca.gov/pdfs/reports/2018-126.pdf.

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, edited by Simone Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Drucker, Julie Williamson, and Koji Yatani, 1–22. New York: Association for Computing Machinery, 2022.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 134–48. New York: Association for Computing Machinery, 2018.

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. "Risk assessment and decision making in child protective services: Predictive risk modeling in context_. In *Children and Youth Services Review* 79, 291-298. https://www.datanetwork.org/wp-content/uploads/PRM-CYSR-article.pdf

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, edited by Regina Bernhaupt, Florian Mueller, David Verweij, and Josh Andres, 1–12. New York: Association for Computing Machinery, 2020.

Field, Anjalie, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risk of racial biases in NLP tools for child protective services." FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. June 2023. Pages 1479–1492. https://doi.org/10.1145/3593013.3594094

Fitzpatrick, Maria, Katharine Sadowski, and Chis Wildeman. "Do Humans Use New Information from Algorithms? A Randomized Controlled Trial of the Availability of Predictive Risk Model Information in Child Maltreatment Investigation Decisions." 2023. Working Paper.

Gillingham, Phillip. "Decision-Making Tools and the Development of Expertise in Child Protection Practitioners: Are We 'Just Breeding Workers Who are Good at Ticking Boxes?'" *Child & Family Social Work* 16, no. 4 (2011): 412–21. https://doi.org/10.1111/j.1365-2206.2011.00756.x.

Goldhaber-Fiebert, Jeremy and Lea Prince. "Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County (Phase 2)." March 26, 2023.

Grimon, Marie-Pascale, and Christopher Mills. "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial." December 17, 2022. https://www.dropbox.com/s/g9d5dqbztd41o9l/ChrisMills_JMP.pdf.

Kahneman, Daniel and Amos Tversky. "Judgment Under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131, 1974.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown and Company.

Kawakami, Anna, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, edited by Simone Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Drucker, Julie Williamson, and Koji Yatani, 1-18. New York: Association for Computing Machinery, 2022.

McNellan, Claire R., Daniel J. Gibbs, Ann S. Knobel, and Emily Putnam-Hornstein. "The Evidence Base for Risk Assessment Tools Used in U.S. Child Protection Investigations: A Systematic Scoping Review." *Child Abuse & Neglect* 134 (2022): 105887. https://doi.org/10.1016/j.chiabu.2022.105887

Putnam-Hornstein, Emily, Rhema Vaithianathan, John Prindle, Stephanie Cuccaro-Alamin, Huy Nghiem, and Tanya Gupta. "Predictive Risk Modeling: Findings from California's Proof-of-Concept." September 2018.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. "Algorithms, Humans and Racial Disparities in Child Protection Systems: Evidence from the Allegheny Family Screening Tool." June 19, 2023. https://krittenh.github.io/katherine-rittenhouse.com/Algorithms%20Humans%20Racial%20Disparities.pdf.

Stapleton, Logan, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1162-1177. 2022.