Impact Evaluation of the Allegheny Family Screening Tool [Phase 2]

ALLEGHENY COUNTY SUMMARY

This report provides an overview of the Phase 2 evaluation of the Allegheny Family Screening Tool (AFST). This work was led by Professor Jeremy Goldhaber-Fiebert and Dr. Lea Prince of Stanford University and reflects a continuation of their earlier evaluation of the AFST as part of front-end child maltreatment screening decisions made by the Allegheny County [PA] Department of Human Services (DHS) child welfare intake office. Analyses were conducted to assess how the implementation of the AFST tool and associated policies impacted investigation decisions. The evaluators also examined downstream outcomes for children.

In the sections that follow we highlight the key findings that emerged. We also provide a summary of the methods used by our evaluation partners as an Appendix. The full inventory of analyses submitted to DHS as part of this evaluation effort have been posted alongside other reports related to the AFST at www.alleghenycountyanalytics.us.

Important to understanding the data is the statutory context in Pennsylvania. Specifically, allegations fall under either the state's Child Protective Service (CPS) statutes (23 Pa.C.S. § 6303) or General Protective Service (GPS) statutes (23 Pa.C.S. § 6334).[1] CPS referrals must all be investigated and so were unaffected by the introduction of the AFST. In contrast, county screening staff are afforded discretion as to whether a GPS allegation/referral is investigated or screened out. As such, some analyses used CPS allegation trends across policy periods (e.g., pre-AFST vs. post-AFST) as the counterfactual for assessing the impact of the AFST, in addition to methods that compared referrals pre- and-post-AFST implementation while holding referral characteristics constant.

EVALUATION FINDINGS: SCREENING DECISIONS

1 CPS referrals include those made for child abuse, including physical and sexual abuse. CPS referrals must be investigated and require more urgent response times, often overlap with law enforcement and medical investigations, and lead to a determination of whether abuse occurred (that may result in perpetrators being registered in the state's ChildLine registry). GPS referrals include referrals made when there is a risk of harm. For example, neglect, truancy and substance use by parents would all fall under GPS referrals. GPS referrals may be investigated or screened out without further assessment, at the discretion of call screening staff. GPS investigations assess for risk and safety to ensure wellbeing of children and provide families with any supports they may need. GPS investigations cannot result in registry with the state's ChildLine registry. Both CPS and GPS referrals can result in a family having a case opened at the end of an investigation for ongoing services and supports. In the pre-AFST period, approximately 21 percent of referrals were CPS allegations and 79 percent were GPS. In post-AFST period, approximately 18 percent of DHS referrals were CPS referrals and 82 percent were GPS referrals.

Near-term, the introduction of the AFST and surrounding policy and practice changes led to the following:

- **A slight increase in the probability that a referral was screened in.** Across policy periods, the probability that a GPS referral was screened-in for investigation increased by 2 percentage points overall (3.8% increase). No CPS comparison could be made given that state law mandates investigation for those referrals.
- **Notable changes in the probability of investigation corresponding to risk of future system involvement.** The pattern of change across children of different risk scores as classified by the AFST was consistent: post-implementation, GPS allegations involving *children who were at lower risk were less likely to be investigated* while *those who were higher risk were substantially more likely to be investigated*.
- **A decline in the probability that investigated children had a case opened for services.** Notwithstanding an increase in investigations for children in high-risk GPS referrals, there was an 8 percentage point decline (20% decrease) in the probability that investigated children had a case opened for services. This decrease in case openings was likely due to other County policy and practice changes: CPS referrals (where the AFST had no impact due to mandated investigation) also experienced a similar decline in case opening rates (18.4% decrease).

The introduction of the AFST and surrounding policy and practice changes led to the following changes on client outcomes in the 6 months following child welfare involvement:

- **A reduction in subsequent referrals of alleged maltreatment.** The probability that there was a re-referral decreased overall and across all risk groups (except those in the highest risk). In contrast, for CPS referrals (where the AFST had no impact due to mandated investigation) there was an overall increase in re-referrals. This suggests that the introduction of the AFST and surrounding policy and practice changes led to a reduction in re-reporting.
- **Reduction in out-of-home removal rates.** Reflecting improvements in screening decisions, out-of-home removal rates following an investigation were statistically significantly lower overall and across all risk groups ($p < 0.001$). While this reduction was also observed for CPS referrals, the difference was smaller and not statistically significant ($p=0.801$), suggesting that the AFST and surrounding policy and practice changes may have contributed to a reduction in out-of-home removals.[2]
- **Decreased differences in outcomes between White and Black/African American children.** Racial differences in outcomes were reduced with the implementation of the AFST and surrounding policy and practice changes — and for removals there was suggestive evidence it was eliminated entirely (see summary table below).

2 The p-values are reported in Table 6 of Goldhaber-Fiebert et al., 2023.

ALIGNMENT OF RACIAL DISPARITY FINDINGS WITH OTHER STUDIES

Although Goldhaber-Fiebert and Prince were commissioned to undertake the overall evaluation of the AFST, other independent studies have also emerged. These include Rittenhouse et al., (2023) who used Allegheny County data with a different methodology than used here, and Mills et al., (2023) who conducted a randomized control trial of a predictive risk model similarly implemented at call screening in a Colorado county. The table below compares the estimated impact on racial disparities for these evaluations.

There is striking alignment — *all three studies documented reductions in racial disparities*.

OUTCOME	REDUCTION IN RACIAL DISPARITIES GAP RELATIVE TO INITIAL GAP
Goldhaber-Fiebert et al., 2023 (Allegheny County, PA)	
Investigation (GPS)	32% of gap eliminated
Case Opened (GPS)*	92% of gap eliminated (15% for CPS)
Removals within 6 months	100% of gap eliminated
Rittenhouse et al., 2023 (Allegheny County, PA)	
Investigation (GPS)*	46% of the gap eliminated
Case Opened*	91% of the gap eliminated
Removals within 3 months*	73% of the gap eliminated
Mills et al., 2023 (Colorado County)	
Investigation*	50% of gap eliminated
Hospitalization disparities*	56% of the gap eliminated

*Notes: * implies the reduction was statistically significant with a p-value ≤ 0.05.*

CONCLUSIONS

The AFST was designed to consistently deliver structured data and a standardized assessment of risk to call screeners. Associated policies were established such that call screeners could use the AFST to supplement the screener's assessment of risk and safety. In an effort to fulfill our ongoing commitment to using the best tools available to protect children, we contracted for multiple independent evaluations and continue to make data available to external researchers. These evaluations are part of a broader emphasis on quality improvement within child welfare, including internal DHS quality assurance and monitoring efforts.

In the evaluation summarized here, we sought to better understand how the introduction of the AFST at call screening impacted: (1) investigation and case openings, (2) re-referrals and placements at the 6-month mark and (3) racial disparities.

We are pleased to share these findings and look forward to continuing to learn how we can use data to support our workforce in their efforts to protect children and support families.

APPENDIX: METHODS

For this Phase 2 evaluation, the team from Stanford University used individual-level multivariable regressions employing generalized linear models and time-to-event models. Outcomes were compared for two groups:

1) Children named as alleged victims in GPS referrals following the implementation of the AFST vs. during the period before AFST implementation (i.e., post-AFST vs. pre-AFST).
2) Children named as alleged victims in CPS referrals during these same time periods, but for whom the AFST was not used in the investigation screening decision.

Evaluation Questions

The Phase 2 evaluation was oriented around a number of screening and subsequent 6-month outcomes, as well as analyses of racial disparities. Analyses were conducted both overall and by AFST risk level. Key questions included:

Screening Decisions

- How did the implementation of the AFST change the probability that a GPS referral was screened in?
- How did the implementation of the AFST change the probability that a screened-in GPS referral led to a case opened for services?
- During this period, were any changes observed in the probabilities for CPS referrals?

Per the evaluation team: *"AFST-related changes in screening probabilities are intended to measure how the tool's implementation impacted two features of accuracy: 1) are otherwise similar children more or less likely to be screened in for investigation and does this occur because the tool identifies those as higher (lower) risk who are then more (less) likely to screen in? 2) are otherwise similar children that screen in for investigation more likely to have further action taken and does this occur because the tool identifies and helps to screen in higher-risk children (and screen out lower-risk children)?"*

Accuracy

- How did the implementation of the AFST change the probability of future referrals after a GPS referral was screened in?
- How did the implementation of the AFST change the probability that a screened-in GPS referral had a future referral that was both screened in and led to a case opened for services?
- How did the implementation of the AFST change the probability that a screened-in GPS referral had a future referral that was both screened in and led to an out-of-home removal?
- For these same outcomes, were any changes observed in the probabilities for CPS referrals?

Per the evaluation team: *"AFST-related changes in 6-month outcomes are intended to measure how the tool's implementation impacted longer-term accuracy, i.e., are otherwise similar children that screen in for investigation who are higher (lower) risk then subsequently more (less) likely to have episodes, episodes that screen in, accept for service, or have home removal? If the AFST helps to identify children for whom investigation and services for*

the index episode can be effective, then one may expect lower rates of subsequent episodes, especially those that require another screen-in and services or result in removal. In contrast, since CPS episodes do not involve the use of the AFST, one would expect either no changes in similar outcomes or else only changes in outcomes related to other factors including some of the policy and practice changes that accompanied the implementation of the AFST."

Safety

- How did the implementation of the AFST change the 6-month probability of any future DHS interactions after a GPS referral was screened out?
- How did the implementation of the AFST change the 6-month probability of a screened-in referral (GPS or CPS) following GPS referral that was screened out?
- How did the implementation of the AFST change the 6-month probability of a screened-in referral (GPS or CPS) that resulted in a case opened for services following GPS referral that was screened out?
- How did the implementation of the AFST change the 6-month probability of a screened-in referral (GPS or CPS) that resulted in a removal following GPS referral that was screened out?

Per the evaluation team: *"AFST-related changes in these outcomes are intended to measure how AFST tool implementation impacted longer-term safety: are otherwise similar children that screen out who are higher (lower) risk then subsequently more (less) likely to have episodes, episodes that screen in, accept for service, or have home removal? If the AFST helps to identify children who do not require investigation and services for the index episode, then one may expect lower rates of subsequent episodes, especially those that require screen in, services, or home removal."*

Disparities

For the above screening, accuracy and safety questions, changes in probabilities were also examined for Black and White children. While other/undetermined race categories are controlled for in the main analyses that were conducted, the disparities analysis was limited to stratification by White and Black/African American race; these two categories comprise over 90% of children included in referrals.[3]

Per the Evaluation team: *"LASSO-related changes in these outcomes are intended to measure how LASSO tool implementation may have helped to standardize system responses which in turn may have reduced differences between outcomes for otherwise similar Black and White children."*

Differences between Phase 1 and Phase 2 Evaluations

It is important to note that the Phase 2 evaluation differed from the initial evaluation in several ways. Although both evaluations focused on outcomes related to accuracy, safety and outcomes by race, this second evaluation

3 A child was coded as "Black/African American" if his/her race was Black, African American or mixed Black or African American, at the time of the referral. For outcomes which incorporate re-referrals, race was coded based on the race recorded in the index referral.

also included longer-term, 180-day outcomes including home removal. From an analytic standpoint, the Phase 2 evaluation used individual-level multivariable analyses using generalized linear models and time-to-event models. This is in contrast to the interrupted time series analyses that were used in Phase 1.

Per the evaluation team, the Phase 2 evaluation was also substantially more challenging than Phase 1 because of two key dynamics: the timing of updates to the AFST and the onset of the COVID-19 pandemic. To address changes in the practice and policy environment, as well as other factors that may have impacted outcomes, the evaluation team employed statistical models that adjusted for multiple covariates.

Finally, the Phase 1 evaluation treated each referral as a separate unit of analysis. In Phase 2, the Stanford evaluation team constructed "episodes" as the unit of analysis. This decision reflected a practice reality in which multiple referrals could occur within close proximity to one another – and an awareness that the types of referrals and decisions about whether or not to screen in a referral for investigation may be related to protocols concerning other referrals. The evaluation team defined an episode as a referral or cluster of referrals within 14 days of one another.

Data

All analyses used de-identified data provided by the County to the evaluation team. Data consisted of demographics, allegation, call-screening decisions, investigations, case openings and removal decisions. Records also include all previous referrals and investigations in Allegheny County from August 1, 2013 through March 31, 2021.

The analytic dataset included records for children under the age of 18 at the time of the first referral in an episode. Children over 17 years-of-age at the time of the first referral were excluded (although we account for 18-year-old children in subsequent episodes).

Unit of Analysis

Episodes were categorized into one of three mutually exclusive groups:

1) **CPS EPISODE:** One or more referral(s) within the episode in which a CPS referral was received (note: CPS is always screened in)
2) **GPS SCREEN IN:** One or more GPS referral(s) screened in within the episode and no CPS referrals
3) **GPS SCREEN OUT:** Only screened-out GPS referral(s) and no CPS referrals

In both the pre- and post-implementation periods, more than 95% of episodes contained only one referral. The vast majority of episodes with more than one referral consisted of two referrals. There was no substantial change in the mix of referral types across the implementation periods.

Evaluation Window

Outcomes were compared for children involved in GPS referral episodes in the 17 months after full implementation of the AFST ("post-implementation"; October 27, 2019 through March 31, 2021) (~20,000 children) to outcomes for children involved in GPS referral episodes in the period before implementation of any tool ("pre-implementation"; primarily January 1, 2015 through July 31, 2016) (~31,000 children). As a

second point of comparison, changes in outcomes across these periods for children involved in CPS referral episodes for whom the AFST is not used as all referrals are screened in by the state.

Analytic Approach

For all outcomes examined, the evaluation team examined unadjusted population means for the pre-implementation period (January 1, 2015 through July 31, 2016) and the post-implementation period (October 27, 2019 through March 31, 2021).

For screening decisions, the evaluation team performed individual-level multivariable regression analyses to estimate the impact of the AFST implementation and surrounding policy and practice changes on the predicted level of each outcome (e.g., probabilities of screening in). Results from the multivariable analyses are reported as the predicted probability that an outcome will occur. Predicted probabilities are averaged predicted values for each child in the dataset with all variables held at their observed value.

For analyses of accuracy and safety, the evaluation team used multivariable time-to-event (survival) analyses to estimate the rates of an outcome occurring (e.g., subsequent episodes of various types and home removal). The model was a generalized gamma survival model, chosen for its flexibility in underlying assumptions regarding changes in the baseline hazard over follow-up time. Children were considered "at-risk" while they were 18 or younger and not removed from their home (as determined by dates of removal/return in the data set). The cumulative risk for each outcome was estimated by measuring the area under the average adjusted survival curve at 6 months for each policy period.

The evaluation team also examined how outcomes varied across policy periods (e.g., pre-AFST vs. post-AFST) within each risk score category. Predicted probabilities were generated for all children as described above, while holding score category at a fixed value and all other independent variables at their observed values.

Finally, for all outcomes, to examine the AFST impact for children of different races (Black or White),[4] the evaluators estimated the predicted probability of an outcome for all children by holding race category at a fixed value and all other independent variables at their observed values.

Covariates and Standard Errors

Multivariate adjustments included child characteristics at the time of the first referral in an episode (i.e., child age, race, sex) and risk score category based on the maximum score across all referrals within a given episode (using the AFST algorithm to create risk scores for children in the pre-AFST period). The evaluation team also included covariates for month and other dimensions of time, as well as referrals that were attached to an already opened case for services and/or that involved truancy allegations. Standard errors were clustered at the child-level.

---

4 The race of child was coded as noted above. For outcomes that incorporate re-referrals, race was coded based on the race recorded in the index referral.