SECTION 7

Frequently-Asked Questions

by the Allegheny County Department of Human Services

CONTENTS

INTRODUCTION

In August 2016, the Allegheny County Department of Human Services (DHS) implemented the Allegheny Family Screening Tool (AFST), a predictive risk modeling tool designed to improve child welfare call screening decisions. The AFST was the result of a two-year process of exploration about how existing data could be used more effectively to improve decision-making at the time of a child welfare referral. The original model (Version 1) was utilized from August 2016 through November 2018. An updated model (Version 2) is now being used. For more information about the AFST, see here.

The process began in 2014 with a Request for Proposals and selection of a team from Auckland University of Technology led by Rhema Vaithianathan and including Emily Putnam-Hornstein from University of Southern California, Irene de Haan from the University of Auckland, Marianne Bitler from University of California – Irvine and Tim Maloney and Nan Jiang from Auckland University of Technology. Prior to implementation, the model was subjected to an ethical review by Tim Dare of the University of Auckland and Eileen Gambrill of the University of California-Berkeley. Upon the conclusion of this review, to which DHS prepared a response, the County proceeded with implementation. Concurrent with this process was the issuance of a second Request for Proposals, at the end of 2015, for an impact and process evaluation of the model. Contracts were awarded to Stanford University (impact evaluation) and Hornby Zeller Associates (process evaluation).

A report on the development of the AFST,[1] prepared by Rhema Vaithianathan, PhD; Nan Jiang, PhD; Tim Maloney, PhD; Parma Nand, PhD; and Emily Putnam-Hornstein, PhD, was published in April 2017 and a report on the development of the AFST Version 2 was published in April 2019. The following Frequently-Asked Questions are presented as a quick reference for those interested in highlights from these publications as well as the evaluations and should be considered within the context of the full publications. Page numbers are provided throughout the document, indicating where the reader may find more detailed information.

1 Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening

BACKGROUND

What is the Allegheny Family Screening Tool (AFST) and how does it work?

The AFST was developed to support one key decision in the child welfare process: whether or not to screen-in a referral for investigation.

To generate the AFST scores, the AFST uses more than 100 predictive factors for each child on the referral. In V1 of the AFST, these factors were then weighted through a logistic regression model to calculate two AFST scores (ranging from 1–20) for each child: the risk of placement within two years if the referral is screened-in and the risk of re-referral within two years if the referral is screened-out.[2] Call screeners and supervisors see the maximum AFST score from the referral. For example, if there are two children on the referral and one has a maximum risk score of 12 and the other has a maximum risk score of 16, the call screener will see a score of 16.

2 This methodology was altered in V2 of the AFST; see page 18 of this FAQs document for more information about V2.

It should be noted that while in some settings machines have been used to make decisions that were previously made by humans, this is not the case for the AFST. It was never intended or suggested that the algorithm would replace human decision-making. Rather, the AFST should help to inform, train and improve the decisions made by the child welfare staff.

Who are the key partners and how were they selected?

The Allegheny County Department of Human Services (DHS) issued a Request for Proposals (RFP) in 2014, to design and implement a system of decision-support tools and predictive analytics for human services.[3]

3 Decision Support Tools and Predictive Analytics in Human Services RFP

We received 15 proposals in response to the RFP. After review by an evaluation committee, researchers from Auckland University of Technology (AUT), University of Southern California (USC), University of California-Berkeley and University of Auckland were awarded the contract and conducted the work. The research team was led by Rhema Vaithianathan (AUT).

Has the local community been involved in the decision to use the AFST?

Community engagement has been a priority for the County throughout the project. The County sought input from the community through various meetings, including six project-specific meetings. Three were held at early stages of the project to collect feedback from key external stakeholders and funders. DHS then held three open community meetings where over 30 stakeholder groups (including the Courts and the ACLU) were invited to discuss the work to date, implementation timeline and results. Additionally, DHS shared project updates with existing community networks including the Children's Cabinet and the Children, Youth and Families Advisory Board, and through the DHS Speaker Series. Feedback from these community meetings has influenced the project throughout its development.

How will the AFST be evaluated?

An RFP for two independent evaluations of the AFST (process and impact) was issued in 2015.[4] Hornby Zeller Associates was selected to conduct an process evaluation and Stanford University was selected to conduct an impact evaluation. The process evaluation is available here. The impact evaluation focused on whether the AFST increased the accuracy of decisions, reduced unwarranted variation in decision-making and reduced disparities, and also examined overall referral rates and workload. A summary of the impact evaluation can be viewed in Section 5 and the full impact evaluation can be viewed in Section 6.

4 Evaluation of a Predictive Risk Modeling Tool for Improving the Decisions of Child Welfare Workers RFP

AFST VERSION 1

What was the total cost of developing the AFST?

The total cost was $1,185,424, as detailed below:

VENDOR	SERVICE	TOTAL
Auckland University of Technology	Methodology and Model Design	$500,000
Deloitte	Technology	$280,000
Stanford University	Impact Evaluation	$310,000
Hornby Zeller Associates	Process Evaluation	$95,424
TOTAL		**$1,185,424**

What data does the AFST use?

The AFST uses information from DHS's integrated data system that links administrative data from 21 sources including child protective services, publicly funded mental health and drug and alcohol services, and bookings in the County jail. Please **see page 11** of the methodology and implementation report for additional information on the data used. See the section about AFST Version 2 for information about changes that have been made to data sources since implementation.

Doesn't the AFST just predict child welfare system decision-making?

A challenge is to identify outcomes to predict that are truly independent of the system and not too rare to be predicted.

The first adverse outcome predicted by the AFST is placement within two years of screen-in. Because placements are determined by a judge, and all parties (parents, children and County) are represented by attorneys, a placement outcome is reasonably independent of the County child welfare system.

The second adverse outcome that the AFST predicts — re-referral after an initial referral has been screened-out — is independent of the County child welfare system because referrals come from the community. In AFST Version 2, we eliminated the second outcome. See the FAQs section related to Version 2 for information about this change.

Does the AFST use race as a factor?

No. The County made the decision not to include race as a factor in the AFST because including race does not improve the accuracy of the score. This doesn't mean, however, that other variables in the tool aren't correlated with race. There are other predictors that are correlated with race due to potentially institutionalized racial bias (e.g., criminal justice history) that would imply that race is still a factor. For this reason, continued monitoring of application of the model with regard to racial disparities should be undertaken.

Please **see page 29** of the methodology and implementation report for additional information on the impact of race as a predictor and *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County.*

Does the AFST use prior allegations of maltreatment as a factor?

Yes, because historical data tell us that previous reports of maltreatment, substantiated or not, have predictive power (there is no factor included in the model that does not have significant predictive power). However, Title 23 Sec. 6337 of the PA Consolidated Statutes and the Pennsylvania Department of Human Services provide guidance as to the length of time that allegation reports remain in KIDS (the child welfare case management system), one of the sources queried by the algorithm. Once a report is expunged, the algorithm is no longer able to access it and it is therefore not included in the algorithm. Expungement timelines range from one year and 120 days (for unfounded reports) to five years and 120 days after receipt of the report or closure of services (or until the subject child is 23) for founded reports.

How accurate is the AFST?

Measuring the accuracy of predictive tools is not simple; however, at rollout, the accuracy of the AFST for predicting whether a child would be placed in care within two years after being referred and screened-in for investigation was 70 percent (if measured by area under the curve (AUC)[5].

The new model is better than digital mammography in asymptomatic women.

Please **see page 15** of the methodology and implementation report for additional information on model performance and AFST Version 2 for updated information on model performance.

Has the AFST been validated?

In addition to assessing the accuracy of the AFST in predicting placement and re-referral, the research team also conducted an external validation looking at the likelihood of hospital events (emergency department visits and inpatient admissions). Findings show that over a broad range of injury types there is a positive correlation between the placement scores generated by the AFST at referral and the rate of hospital events.

For example, those children with a placement risk score of 20 (the highest possible score) have a hospital event rate for self-inflicted injury or suicide of 0.65 percent compared to 0.03 percent

5 This figure is an update of a previously higher reported figure in the FAQs that over-stated the AUC because of some technical issues related to the way in which the data was split. For more technical details, please see Chouldechova, Alexandra, et al. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." Conference on Fairness, Accountability and Transparency. 2018."

for those with a placement risk score of 1 (the lowest possible score). That is, a child who scores a 20 at referral is 21 times more likely to be hospitalized for a self-inflicted injury than a child who scores a 1.

Please **see page 19** of the methodology and implementation report for additional information on the hospital validation study. Additional information is available on **page 7** of the Methodology, Version 2.

What did the research tell us about existing practice?

Prior to introduction of the AFST, call screeners could access and use historical and cross-sector administrative data related to individuals associated with a report of child abuse or neglect through Client View, a front-end application to the integrated data system. Call screeners were required to review all relevant information related to a referral and provide it to the call screening supervisor so that a screen-in/screen-out decision could be made. However, it was challenging for call screeners to efficiently access, review and make meaning of all available records. The AFST provides a consistent way to access and weight the available information to predict the risk of future adverse events for each child on the referral.

Researchers found that existing practice had screened out one in four children who the model would screen-in due to their score. For these children, who the model scored as highest risk, 9 in 10 were re-referred (if screened out) and half were placed in foster care (if screened in) within two years. Forty-eight percent of the lowest-risk cases were screened-in with only one percent of these referrals leading to placement within two years.

What happens when there is missing/duplicate information?

The AFST leverages a probabilistic matching algorithm to catch as many duplicate IDs as possible. This method, however, does not capture all duplicate IDs for the same person and, thus, it is possible for an AFST score to exclude data held on a second ID. Efforts to minimize duplicate client records are ongoing.

Is the AFST score assigned to a child/family permanently?

No, because the AFST score will change as underlying data change. The County will retain AFST scores for quality assurance and evaluation purposes.

What safeguards are in place to make sure the AFST is working appropriately?

Immediately before the AFST was put into operation, researchers validated the scores generated by the DHS Data Warehouse (for individuals in historical, de-identified data) by generating scores for the same individuals in the research environment, to ensure that the Data Warehouse was accurately running the AFST. Since implementation, County child welfare leadership has been reviewing monthly quality assurance reports to monitor the performance of the AFST.

AFST scores are securely stored and cannot be manually altered by call screeners. However, as an additional quality assurance check, DHS has added functionality to the AFST that allows workers to report feedback on scores that seem wrong/surprising to them.

The independent impact evaluation and process evaluation highlighted some issues, as did the experience of call screeners and supervisors.

Will the County improve the AFST over time?

The AFST has already been rebuilt once by the research team since it came into use in August 2016, taking learnings from practice and using those to optimize how the AFST scores are generated. In 2018, the County built Version 2 of the model, which included improvements identified by process and impact evaluations. See FAQs related to Version 2 in this document and the Methodology Version 2 report for details about the updates.

How does the AFST compare to other approaches?

The AFST has a similar purpose to other decision-support tools like the Structured Decision Making tool (SDM), but the AFST creates a score without the reliance on manual data input that is required for SDM. For the highest category of risk, the AFST outperformed the SDM model.

Please **see page 24** of the methodology and implementation report for additional information on comparing the model to SDM (including a validation study [Dankers and Johnson, 2014]), and rule-based/threshold approaches.

PRACTICE

How many referrals come into the call screening center on an annual basis?

In 2017, the call screening center received 15,768 referrals, of which 11,751 were GPS allegations.

What is the number of call screeners on staff?

As of April 2019, there were 23 call screener positions. The number of screeners working at a given time depends on the day, ranging from 4 on weekend evenings to 15 on weekday afternoons.

What is the average length of time devoted to each screening call?

A typical referral takes 30 to 60 minutes to process.

Who gets an AFST score and how?

All children involved in an allegation of maltreatment,[6] regardless of whether they are described as the victim or not, will be included in the AFST score; that is, all children living in the same household or added to the case by the call screener. When an allegation of maltreatment is received and the call screener enters details into the child welfare case management system (KIDS), a click will automatically generate the AFST score. Call screeners and call screening

6 The AFST is intended to assist in decision-making for CPS referrals; any allegation meeting CPS criteria is immediately investigated (state-mandate).

supervisors are required to generate the AFST score prior to finalizing a screening decision.

Are there some children for whom an AFST score can't be generated?

Yes, those not known to the system and those for whom not enough data are held in the Data Warehouse. The County has determined that the AFST will only be used to screen for risk when data that goes beyond demography (e.g., age, gender, address) are held for one or more person associated with the allegation. If only demographic data are held for all individuals, then the allegation will be assessed using the existing approach (no AFST score will be generated). As of April 2017, approximately 10 percent of incoming referrals were not generating an AFST score.

Who has access to the AFST score?

Only the call screener and call screening supervisor have access to the AFST score. If and when a referral moves to the investigation stage, investigations staff cannot access any AFST score. The Courts also do not have access to the AFST score. DHS is considering the value and appropriateness of changing this policy.

Please **see page 26** of the methodology and implementation report for additional information on the implementation of the AFST score.

Does a certain AFST score make screening-in mandatory?

The AFST flags some scores as "mandatory screen-ins."[7] The threshold for the mandatory screen-in was determined solely by the placement score and designed to capture as many of the children at heightened risk of abuse-related fatal or near-fatal injuries (Act 33 Events) as possible. The model includes functionality that allows call screening supervisors to override the "mandatory screen-ins" at their discretion; overrides are documented and reviewed.

Please **see page 26** of the methodology and implementation report for additional information on mandatory screen-ins.

Will caseworkers be afraid to 'defy the score?'

The only caseworkers who make screen-in/screen-out decisions are the call screening supervisors. They consider all information provided by the call screeners, including details shared during the call, by the person alleging abuse or neglect, the score generated by the AFST and recommendations from the call screener.

Screening decisions are not in any way 'dictated' by the AFST. Call screening supervisors have full discretion over call screening decisions, regardless of generated AFST scores, and call screening decisions are not required to align with the AFST score. In the AFST's first full year of operation, just 63 percent of referrals with a "mandatory screen-in" score were actually screened-in for an investigation. Conversely, even the lowest AFST scores had about a 30 percent screen-in rate.

7 The term "mandatory screen-in" is enclosed in quotations to reflect the fact that call-screening supervisors may override the score.

How do the AFST and the County minimize the risk of stigma?

No system can entirely remove the chance of screening-in some of the 'wrong' children, so wrongly stigmatizing them. The ethicists suggest, however, that we must then take a comparative view: Is the proposed tool as good or better than the existing approach, when it comes to minimizing the risk of stigma? Compared to the existing system, the AFST is expected to increase accuracy and consistency of decision-making, which means wrongful stigma is expected to be reduced. The impact evaluation assesses this.

In particular, the County will work to minimize stigmatization by carefully controlling access to AFST scores and providing appropriate training that aims to reduce stigmatization and ensures that call screeners are aware of the possibility of false positives/negatives and understand the risk of confirmation bias.

Are AFST scores higher for black children?

The AFST model does not apply any weights based directly on race. However, race is associated with many of the underlying data used by the model, so it is not surprising that the tool's scores have been slightly higher for black children compared to white children. For example, up until the end of 2017, 47% of black children received a "high"-range score (15–20), compared to 39% of white children. Conversely, 18% of white children have received a "low"-range score (1–9), compared to 10% of black children. Some degree of racial disproportionality has already been identified at child welfare decision points in prior published analyses, including at call screening. Whether or not the AFST has any impact (positively or negatively) on the degree of variation associated with child race is a key focus of the impact evaluation. See Methodology, Version 2 for an update.

Can the AFST help to reduce unwarranted variation in decision making?

Whether or not the AFST reduces unwarranted variation in decision-making (such as by race/gender, or variation between individual decision-makers) is a key focus of the impact evaluation. Results are available in the impact evaluation report (Section 6 of this packet), the impact evaluation summary (Section 5), and the AFST Version 2 FAQs on **page 18** of this document.

Does involvement in services always increase the AFST score?

No. For example, for 45% of families, receiving of public benefits (e.g., SNAP, TANF) is, in fact, protective. That is, for those families, receiving those services was associated with lower scores than for similar families that did not receive those services.

It is important to note that the fact of receiving a benefit (of any kind) is not of itself associated with a positive or negative effect on the AFST score. Moreover, receiving assistance in a particular service area is not, of itself, associated with a positive or negative effect on the score. The effect depends on which individual on the referral received the service, what type of service it was, and the intensity, duration and recency of the service.

OUTCOMES

Does a "mandatory screen-in" score always mandate an investigation?

No. In fact, with AFST V1, more than one-third of children classified as highest risk by the AFST were screened out by the intake manager.

Has the AFST significantly increased the number of investigations?

In absolute terms, the percentage of calls screened in during the first year of the tool has increased by less than a percentage point. Whether this resulting screen-in rate is higher or lower than it would have otherwise been in the absence of the tool is one thing the impact evaluation hopes to more thoroughly investigate.

What are the screen-in rates by category?

For AFST V1, which was in use from August 2016 through November 2018, screen-in rates by category were as follows:

SCORE CATEGORY	PERCENT SCREENED-IN FOR INVESTIGATION
Mandatory	61%
High	47%
Medium	42%
Low	31%
No Score	23%
Total	41%

Have more families been accepted for service since implementation of the AFST?

As a percentage of new General Protective Services referrals screened-in for the investigation, the accept-for-service rate was about 39% for AFST Version 1 (in use from August 2016 through November 2018)—about a five-percentage-point rise from a comparable year of data prior to the tool's implementation. It is important to note that workers investigating a referral are not able to access the referral's score according to the AFST, and investigative practice does not vary in any way based on a referral's score.

What is the likelihood that an investigation leads to a placement?

Under Version 1 of the AFST, about 9% of GPS referrals screened in for investigation led to at least one child being removed in the following 90 days.

PROCESS EVALUATION

What data collection methods did HZA use in its process evaluation?

HZA utilized interviews, surveys and data analysis to complete the process evaluation.

Interviews were conducted prior to implementation of the AFST (in July 2016) and four months after implementation (in December 2016). The July 2016 interviews were conducted with 23 DHS administrators and staff, and were designed to learn about a) their involvement in the implementation of the AFST, b) steps taken to prepare call screening staff to use predictive risk modeling to inform their decision-making, and c) the call screening process as it existed prior to implementation of the AFST. The December 2016 interviews were conducted with DHS stakeholders (child welfare staff, staff from the DHS Office of Analysis, Technology and Planning), as well as representatives from community service providers, advocacy groups, foundations and family court. DHS staff were asked about a) their involvement in implementing the AFST, b) the training they received, and c) how the AFST informs or impacts their work. External stakeholders were asked about a) their awareness of DHS's efforts to implement predictive risk models, b) their hopes for what the AFST would accomplish, and c) the successes and challenges they expected DHS to face.

A ***web-based survey*** was administered to call screeners approximately two months post-implementation (September 2016), and a follow-up survey was administered in February 2017 to account for improvements that had been made to the AFST. Using a series of Yes/No and Likert scale questions, call screeners were asked about the training they received, the functionality of the tool, visualization of the scores and the impact of the tool on their decision making. Several open-ended questions were also asked to gather input on what could be done to improve the use of the tool and the training provided to prepare staff to use it.

Data analysis consisted of 1) quantitative analysis of summary statistics, frequency counts and percentages and 2) qualitative analysis of the common themes and items of importance from the interviews and open-ended survey questions. Using a grounded theory approach, the results of the qualitative analysis described the implementation process from the perspective of the stakeholders.

See page 5 of the HZA evaluation report for more detail on the evaluation methods.

How well did staff feel the training prepared them to use the AFST?

The survey administered to call screeners two months after implementation showed that 82% felt somewhat (38%) or very well (44%) prepared to use the AFST. Only six percent reported being "limitedly" prepared and none reported that they were not at all prepared. No opinion was expressed by 13% of responders. By the time the follow-up survey was administered, 100% of respondents reported being adequately prepared to use the tool.

What aspect of the training was found to be most helpful?

Most helpful components were Information about how predictive analytics was to be applied in Allegheny County (36%), use of case scenarios (29%), overview of predictive risk modeling (21%), and overview of changes to KIDS and policy/practice (7% each).

How well do call screeners understand the AFST?

The follow-up survey included a series of questions designed to gauge screeners' understanding of the AFST. Ninety-four percent both understand what the score is predicting and how it should inform screening decisions. Eighty-nine percent understand the content of the data sources used to produce the score.

Are call screeners confident in the AFST's ability to accurately assess the risk of a future referral or out-of-home placement?

Half of call screeners said they were confident of the AFST's ability to assess risk and 61 percent were confident in the research that went into its development. Lack of confidence in the AFST's ability to predict risk seemed to stem from its inability to take expected improvement or individual circumstances into account; for example, when families are receiving services that are improving their situation.

Have there been any technical issues related to implementation of the AFST?

Nearly three-quarters of call screeners noted that they occasionally encounter a score that seems inaccurate, with an additional 11 percent frequently encountering an inaccurate score. In response, they either notify a supervisor, review and use available data, or contact technology staff.

Two early technical issues related to missing or duplicate Master Client Index numbers, were corrected in November 2016. However, an ongoing issue is that the system is reportedly slow and sometimes times out before generating a score.

Did DHS effectively engage and communicate with external stakeholders about the development of the AFST?

External stakeholders appreciated DHS's efforts to educate and inform them about the purpose, development and implementation of the AFST. They felt positive about the tool, its potential to improve decision making, and DHS's plans for implementation. A desire for ongoing communication was noted.

How easy is it to navigate/use the AFST?

Over 60 percent of respondents found the AFST easy or very easy to use, although this response declined between the initial and follow-up surveys (from 69% to 61%). Slightly more than 30 percent of respondents to both surveys were neutral about this question while six percent of respondents to the follow-up survey found the tool difficult to use.

How useful is the graphic display of the score (in the form of a thermometer)?

Responses to this question were mixed, with 44 percent responding that the thermometer was helpful or somewhat helpful, 38 percent reporting no opinion and 19 percent reporting that it was not helpful or helpful only on a limited basis.

Do call screening staff conduct a more thorough data search (either in ClientView or in child welfare's Key Information and Demographics System) when the AFST is high?

More than 60 percent of survey respondents reported that they "rarely" or "never" conduct an additional search, with full-time screeners more likely to conduct additional searches. Most call screeners did not conduct additional searches because the AFST score is already based on those data or because they had already completed searches in the Data Warehouse earlier in the process.

What concerns do call screeners have about the AFST?

Call screener concerns related mostly to the tool's inability to incorporate human judgement into the score or to recognize information that needs to be updated, thus generating a score that inaccurately portrays a family's actual circumstances.

Do call screeners anticipate that the AFST will have an impact on practice?

Between the first and second surveys, the percentage of those who anticipated no impact decreased from 50 percent to 44 percent. The percentage of those who thought the AFST would strengthen practice remained consistent at 44 percent. There was an increase in the percentage of those who thought the tool would diminish practice (from 6% to 11%).

Is the AFST creating a more data-driven culture at DHS?

Sixty-one percent of respondents to the follow-up survey agreed that the tool is creating a data-driven culture. Considering this finding along with the impact finding (previous question) might indicate that call screeners already thought that DHS's culture was data-driven (i.e., based on good screening practices).

Are call screeners using the AFST to inform their recommendations?

By the time of the follow-up survey, 72 percent of call screeners reported using the tool at least occasionally; only 11 percent always use it, while another 28 percent almost always use it. Whereas this percentage increased slightly from the initial survey (at 69%), the percentage of those who always use the tool decreased and the percentage of those using it occasionally or almost always both increased.

What recommendations emerged from the process evaluation?

HZA made the following recommendations in response to the evaluation results:

1. Maintain transparent communication with internal and external stakeholders.
2. Increase user buy-in.
3. Continue to resolve technical issues as they arise, documenting solutions.
4. Develop implementation benchmarks to foster buy-in and promote use of the tool for decision-making.

See page 19 of the HZA evaluation report for more detail about the recommendations.

IMPACT EVALUATION

Where can I find the full evaluation report?

To read the full technical report, please see: Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office.

Where can I find a summary of the evaluation?

To read, please see: Impact Evaluation Summary.

Who conducted the impact evaluation?

Stanford University was awarded the contract for the impact evaluation through a competitive process (Evaluation of a Predictive Risk Modeling Tool for Improving the Decisions of Child Welfare Workers RFP). The Request for Proposals was issued in December 2015; we received seven proposals and the County made its selection in early 2016. A report describing the results of the impact evaluation was finalized in March 2019. Two peer reviewers provided critical feedback on drafts of the report.

What evaluation methods were used?

Stanford University used a set of methodologically strong, quasi-experimental methods (e.g., interrupted time series analyses, generalized linear models). Quasi-experimental methods refer to a type of evaluation approach used when it is not possible or desirable to implement a randomized controlled trial (RCT). While less robust than a gold-standard RCT, carefully designed quasi-experimental methods are considered the next-best approach to testing program impact. The County decided not to pursue an RCT primarily for practical reasons.[8]

More specifically, evaluators used the following tests:

- **Unadjusted Population Means.** The simplest comparison performed was a comparison of unadjusted means for the Pre- and Post-AFST periods, testing whether they are statistically different from one another using a two-sided t-test of equality of means.
- **Interrupted Time Series Analysis (ITSA).** Changes in the level and trend of monthly rates of each outcome during the Pre- and Post-AFST periods were assessed using an Interrupted Time Series Analysis. In this evaluation, the ITSA measures changes in both the level and slope of each outcome in the Post-AFST months in relation to the Pre-AFST months.

8 The State of Colorado has contracted with Cornell University to conduct an RCT of their implementation of a similar predictive risk model implemented at the child welfare hotline in Douglass County, Colorado.

The ITSA approach captures population-level changes in outcomes and trends after a policy change (in this case, the implementation of the AFST) in comparison to the levels and trends prior to that change.

- **Child-Level Multivariate Regression Analysis.** Finally, the evaluators used multivariate individual-level regression analyses to assess the impact of the AFST on the predicted level of each outcome Pre- and Post-AFST, while adjusting for child and household characteristics. These analyses focus on estimates of the average effect of the AFST, adjusting for evolving case mix over time. The predictive margins presented in the evaluation can be interpreted as the average outcome if all children in the sample were in either the Pre-AFST or the Post-AFST time-frame, holding all other control variables constant.

What time period does the evaluation cover?

The evaluation consists of outcome comparisons for two groups of children: (1) the approximately 31,000 children who were referred for alleged maltreatment during the 18-month period before the AFST was implemented (Pre-AFST: January 1, 2015 through July 31, 2016) and (2) the approximately 34,000 children reported after the AFST was fully implemented (Post-AFST: December 1, 2016 through May 31, 2018). Outcomes for both groups (Pre-AFST vs. Post-AFST) were examined for 15 to 17 months after the initial maltreatment report was received.

What were the main evaluation findings?

1. **Overall, the AFST did not lead to increases in the rate of children screened-in for investigation.** Use of the tool appears to have resulted in a different pool of children screened-in for investigation (including more children who needed intervention supports, see finding 2 below). But from a workload perspective, there was no significant increase in the number or proportion of children investigated among all children referred for maltreatment.
2. **Implementation of the AFST increased the identification of children determined to be in need of further child welfare intervention.** Use of the tool led to an increase in screen-in rates for "higher-risk" children who needed intervention supports. Specifically, there was a statistically significant increase in the proportion of children screened-in who then had a child welfare case opened or, if no case was opened, were re-referred within 60 days. (Please note that investigators and supervisors making case opening decisions remained blind to the score.)
3. **Use of the AFST did not lead to decreases in re-referral rates for children screened-out without investigation.** Re-referral rates among children screened-out stayed the same for children overall, with the exception of children who were 4–6 years of age. This age group was directly affected by County changes to mandatory field screening protocols, which changed mandatory field screenings for referred families with a child under 7 years of age to families with a child under 4 years of age. Unfortunately, for the 4–6 age group there was a slight but statistically significant increase in the likelihood of being re-referred.

4. **The AFST led to reductions in overall case opening disparities between black and white children.** During the Post-AFST period, increases in the identification of higher-risk white children, coupled with slight declines in the rate at which black children were screened-in for investigation, led to reductions in racial disparities. Specifically, there was an increase in the number of white children who had cases opened for services, reducing Pre-AFST case disparities between black and white children.
5. **There was no evidence that the AFST resulted in greater screening consistency within individual call screeners.** Specifically, for the subgroup of 11 call screeners who handled a substantial volume of both Pre-AFST and Post-AFST referrals, attempts were made to assess whether the AFST led to more "within-screener" consistency. Likewise, changes in screening consistency by children's age group and racial group were also assessed. No changes were detected, although it should be noted that there was likely insufficient power to identify anything other than very large shifts.

Will the County continue to fund an independent evaluation?

Yes, Stanford University will continue to follow the outcomes of the AFST in practice, extending the results in time, observing AFST Version 2, and expanding the outcomes reviewed to look at home removals.

AFST VERSION 2

Why were changes made to the original model, operating from August 2016 to November 2018?

DHS was always committed to continuing to improve the model, and we expected to make changes once we had implementation and outcome data. Specifically, the changes were motivated by a number of factors, including:

- Some of the variables (data sources) were unsteady, meaning that they either changed significantly while the model was live and/or they changed from the time period the researchers used to construct the model.
- The re-referral model (which predicted whether a child would be a re-referred within two years) was not as strongly linked to the primary outcome of concern, serious abuse and neglect. Additionally, initial incoming referral rates also represent the most racially disproportionate step of the referral pathway, and so a model predicting future referrals figures to overrepresent black children relative to white. Finally, the nature and characteristics of calls with higher scores using the re-referral model were resonating less strongly with screening staff as cases appropriate for investigation.
- LASSO, the machine learning approach used in the second version, performs better than the logistic regression model used in the original model.

What changes were made to the methodology for AFST Version 2?

Changes were made to the **target outcome**, to the **data sources used in the algorithm**, and to the **policies regarding high-risk and low-risk (and how they are displayed on the visualization)**. Specifically:

- **Target Outcome:** AFST Version 1 (V1) was designed to predict: 1) the likelihood a child would experience abuse or neglect serious enough to be placed in an out-of-home setting within two years of the initial call if the call were screened-in for investigation and 2) the likelihood there would be a re-referral to the hotline within two years if the call were screened-out. Based on feedback from staff and external validation of the model using hospitalization data, we determined that the scores from the re-referral model were not as strongly related to the key outcome of concern, serious abuse and neglect. AFST Version 2 (V2) therefore only predicts the likelihood of out-of-home placement within two years.
- **Data Sources:** In V2, public benefits data were excluded, as were a majority of behavioral health records. Birth records – which Allegheny County began to receive after the building of V1 – were added to the model. Public benefits data were excluded as the current data feeds no longer align to the historic data used to develop V1. Some behavioral health records were eliminated because of temporal variability. In addition, variables regarding the current allegations on the referral were added at the request of call-screening staff.

  Data sources used in V1 of the AFST and continued in V2 include child welfare, jail, and juvenile probation records.

  A complete listing of the variables used in V2 can be found in **Appendix B** of the Methodology V2 report.
- **High-Risk and Low-Risk Policies/Visualization:** The visualization was changed to reflect new high- and low-risk protocols and to provide a visual cue to remind staff that this is a new version (the new visualization can be seen in **Appendix C** of the Methodology V2 report).

In V2, if the maximum referral score is greater than 17 and any child on the referral is younger than 16, the referral is designated to be screened-in for investigation (although supervisory discretion allows an override to screen-out the referral, requiring supporting documentation), and the visualization displays the text "High-Risk Protocol, High Risk and Children Under Age 16 on Referral." If the maximum referral score is under 11 and all children are at least 12, the visualization displays the text "Low-Risk Protocol, Low-Risk and All Children Age 12+ on Referral." The default for referrals identified as low risk is a screen-out unless otherwise deemed necessary; low-risk referrals have to be overridden to be screened in. All other scores are displayed in the visualization and staff has full screening discretion.

What modeling methodology is used in AFST V2?

A number of methodologies were explored for V2, including LASSO, XG-BOOST, Random Forest and SVM logistic regression. To determine which methodology would be used, researchers considered 1) overall performance and accuracy for the high-risk groups; 2) accuracy for black

children versus non-black children; 3) ease of implementation and quality checking; and 4) whether the model showed a positive correlation between the score generated and the probability that the child would be involved in a fatality or near-fatality 50 days or more after the score was generated.

Based on these factors, we chose the LASSO model. A discussion of the performance and external validation of LASSO appears in the Methodology V2 report.

Was the model validated?

An external validation of the model was conducted using Children's Hospital of Pittsburgh data. Encounters were examined (by cause) using four approaches (highest risk score and an injury encounter, randomly selected risk score and an injury encounter, highest risk score before an injury encounter, and randomly selected risk score before an injury encounter). We found a positive correlation between the risk scores and medical encounters for injury, abusive injuries and suicide, showing that the model accurately identifies the children most at risk for relevant hospital events.

Were there accuracy improvements in AFST V2?

There are a number of metrics that can provide information about the accuracy of the algorithm (e.g., area under the receiver operator curve [AUC], outcome plots, mortality regressions, how well the algorithm distinguishes high- and low-risk children, accuracy for black vs. non-black children). Among the methodologies tested, LASSO provided the best balance in increased accuracy, with an overall AUC of 76 percent (74.42% for black children and 77.35% for non-black children) and ability to implement and perform quality assurance checks.

IMPLEMENTATION LESSONS

Do the process and impact evaluations cover everything you've learned since you started building and using the AFST?

No, even the best evaluations can't cover everything. The following FAQs provide additional information and lessons learned during implementation, covering technical, practice and policy reflections.

What are some of the technical lessons learned during AFST implementation?

The technical lessons fall into three categories: efficiency and auditability of variable calculations, unforeseen changes in data availability or content, and complexity of database structures and "real-time" calculations.

1. **Efficiency and auditability of variable calculations**
   The broad array of variables built into the tool, and the initial design for variable calculation and storage made thorough testing and the discovery of all possible calculation defects challenging. It also made the creation of long-term research datasets a burdensome

undertaking; for example, to compile a research modeling dataset with full variables for approximately six years of historic child welfare referrals required almost two weeks of continuous runtime. Initially, the tool called for hundreds (approximately 1000) of distinct variables to be constructed, without sufficient consideration for how they would be utilized by the eventual algorithm(s) producing a score, how they would be used for rebuilding the model (generating a multi-year research data set), and the quality assurance requirements. Tool processing time and design were secondary considerations and created many challenges throughout the initial implementation.

In the interest of enhancing real-time AFST processing speed from the worker perspective, recent efforts have been undertaken to backtrack and decommission obsolete or unused variables from the initial design that are calculated and stored when the tool runs but not actively weighted in any actual algorithms. Being more strategic and selective with initial variable creation in the design stage may have resulted in a leaner and more manageable product structure. A couple of examples of lessons related to efficiency of the model include:

- Many of the variables are repetitive with just slight variations in the data being summarized, however each variable is an independent script in the implementation. In cases where we identify an issue in the calculation, we have to identify all of the variables impacted by the issue and update each one independently. Ideally, there would be more shared/referenced code so that the update would need to be made only once and the changes would be consistent across affected variables.
- The initial design did not capture enough of the intermediate data and calculations. For quality assurance purposes, staff should be able to walk through the logic to get from the source data to the final variable calculations in a clear and transparent way. Investigating potential calculation errors took a significant amount of time due to the way the data model was designed. This process could be much more efficient if the data model was designed with strong consideration for ease of quality assurance.

Based partly on these experiences, major improvements have occurred in how brand-new variables (such as County birth data) were implemented in V2. Performance and processing time were considered at every step in development, validation rules and outlier boundaries were carefully crafted and documented, and variables were only developed if explicitly thought to be important.

2. **Unforeseen changes in data availability or content**

Unlike many analyses conducted, a data mining approach like that used by the AFST doesn't, by definition, draw on clean data sets. In fact, it is more likely the model finds data in less used parts of the system particularly in need of quality assurance. It's probably fair to say that Allegheny County devotes more resources than is typical to monitoring data sources and data quality across its enterprise. That said, the initial quality assurance protocols established were not sufficient to properly monitor the AFST and additional quality assurance had to be developed.

For example, in recent years, structural changes occurred in how behavioral health diagnoses were defined and categorized that did not align with the historic diagnosis variables used in the initial AFST modeling. Because of this, some variables saw significant distribution increases or decreases in incoming prevalence compared to historic data that had been used to determine appropriate weights. In hindsight, the initial tool design should have included some form of automated detection for when incoming data include outliers or appear significantly different than expected, rather than requiring manual detection by analysts monitoring the tool's performance. The team is currently developing an automated quality assurance tool to monitor variable values over time and generate alerts if there are significant changes in a variable that may impact model performance.

3. **Complexity of database structure and "real-time" calculations**
   In developing a tool that aimed to access and utilize real-time, incoming data of varying quality and completeness, constructing variables that were able to properly navigate temporary data staging tables proved to be a new and challenging endeavor. Datasets produced for analyses and for AFST research and modeling inherently had the benefit of full, finalized data entry for a given historic call, without any snapshot mechanism for accurately simulating how complete a data element typically would be in the midst of the early call screening stage. Additionally, in some instances, variables were initially coded to search for data in finalized data tables (where data would eventually be stored later in the process) rather than being directed toward temporary data staging tables where the data would normally exist at the point of call screening. The lesson learned was to spend as much time as possible understanding the exact flow and completeness of data at various processing points.

What are some of the lessons learned (and still evolving) about the use of the AFST in practice?

The most significant, and probably most obvious, lesson is that practice and culture change takes time and that a new tool will have limited immediate impact on culture. As a field, we are slowly evolving from a system that focuses almost exclusively on the allegation of abuse and neglect to one that puts this input in the proper context. In the rebuild of the model, call screening staff requested that the current allegation be included in the model. We had initially decided that this variable should be excluded since the algorithm is best at assessing longer-term risk of abuse and neglect and the call screener alone could assess the current allegation alongside future risk to make a screening determination. We yielded to the requests of the call screening staff to include the variable in the model because it increased their confidence in the score. Nevertheless, our work to change the culture from an allegation-only focus to one with greater understanding of latent risk is just beginning.

Another lesson is that the AFST cannot fix, nor anticipate, other external shocks to the system that might impact practice. This means there must be either very strong communication with frontline managers and/or monitoring of the whole decision-making process. The following

example describes the type of challenge that likely occurs in systems throughout the country and which must be identified and managed if practice is to be consistent.

In late 2017, a combination of factors (staff turnover, staff on medical leave and increased call volume) led to a situation where call screening staff, overwhelmed by the call volume and reduced staff, halted their full business process and began triaging based on referral information. The staff triaged calls they thought serious in one pile, possibly serious in another pile, and likely not serious in a third pile. The problem with making these decisions so early in the process, without the benefit of the full review process or the AFST score, is that cases deemed not serious were later – sometimes a week or two later – determined to be high risk on the AFST. Because we had no monitoring in place to catch this sort of process challenge and because frontline managers did not report the problem, we were lucky to catch the issue at all. Once identified, we considered a variety of solutions and eventually DHS leadership put in place extra supports to allow call screening staff to follow the established protocol, which includes running the score. Today, we have active monitoring and have established tools that allow call screening supervisors to monitor the flow of cases through the decision-making process. The image below displays a week-by-week breakdown of data (12/31/17 through 10/20/18) showing the time that passed between initial referrals and (1) screening decisions and (2) generation of initial AFST scores.

					Time to Screening Decision Approval						Time to Initial ASFT Score					
Week	Start	End	Incoming Referrals Processed	Scr %	Median Days	Average Days	Pct. Over 2 Days	Pct. SOs Over 2 Days	Pct. Over 7 Days	Pct. SOs Over 7 Days	Median Days	Average Days	Pct. Over 2 Days	Pct. SOs Over 2 Days	Pct. Over 7 Days	Pct. SOs Over 7 Days
1	12/31/2017	1/6/2018	249	100%	2.0	3.8	60.6%	85%	16.1%	36%	0.4	1.4	14.5%	20%	4.0%	8%
2	1/7/2018	1/13/2018	312	99%	1.0	3.0	47.4%	74%	11.9%	23%	0.3	0.7	7.4%	14%	0.3%	1%
3	1/14/2018	1/20/2018	292	100%	1.0	3.6	46.9%	66%	15.4%	30%	0.2	1.0	13.7%	19%	0.7%	2%
4	1/21/2018	1/27/2018	356	99%	2.0	5.9	58.1%	87%	18.3%	36%	0.3	2.9	12.4%	24%	1.7%	3%
5	1/28/2018	2/3/2018	349	99%	2.0	4.7	60.5%	92%	18.3%	40%	0.6	2.1	17.5%	36%	2.3%	3%
6	2/4/2018	2/10/2018	307	99%	2.0	3.9	54.4%	87%	13.4%	29%	0.3	1.1	8.1%	17%	2.6%	7%
7	2/11/2018	2/17/2018	331	99%	2.0	6.1	55.3%	78%	26.6%	54%	0.4	3.5	21.8%	31%	12.4%	25%
8	2/18/2018	2/24/2018	308	100%	3.0	6.4	71.4%	94%	32.5%	70%	0.9	3.0	25.0%	39%	13.6%	30%
9	2/25/2018	3/3/2018	337	99%	2.0	4.5	56.7%	87%	19.6%	46%	0.3	1.7	11.3%	22%	6.8%	17%
10	3/4/2018	3/10/2018	377	99%	3.0	5.4	57.3%	81%	29.2%	57%	0.5	2.3	15.6%	29%	13.0%	26%
11	3/11/2018	3/17/2018	344	100%	2.5	5.6	59.3%	93%	28.8%	57%	0.4	2.8	14.8%	29%	12.8%	26%
12	3/18/2018	3/24/2018	323	100%	2.0	3.9	57.3%	78%	21.7%	49%	0.3	1.9	17.3%	37%	12.7%	29%
13	3/25/2018	3/31/2018	286	98%	2.0	4.2	53.5%	82%	18.9%	35%	0.3	1.4	13.3%	23%	3.8%	7%
14	4/1/2018	4/7/2018	330	100%	3.0	4.6	64.5%	84%	27.9%	55%	0.9	2.1	26.4%	37%	10.6%	22%
15	4/8/2018	4/14/2018	307	100%	1.0	5.8	48.9%	84%	24.1%	54%	0.3	3.0	14.3%	27%	9.8%	22%
16	4/15/2018	4/21/2018	349	99%	2.0	5.6	53.9%	83%	27.8%	58%	0.7	3.0	16.9%	32%	11.2%	24%
17	4/22/2018	4/28/2018	385	99%	2.0	7.7	61.3%	83%	29.6%	63%	0.9	4.9	27.0%	42%	15.6%	35%
18	4/29/2018	5/5/2018	367	100%	3.0	5.7	64.9%	89%	31.6%	65%	0.7	2.1	26.2%	37%	14.2%	30%
19	5/6/2018	5/12/2018	348	100%	3.0	1.5	71.8%	88%	30.5%	63%	1.1	4.3	28.7%	43%	15.2%	33%
20	5/13/2018	5/19/2018	334	100%	3.0	3.4	76.6%	96%	32.3%	65%	0.9	4.2	30.8%	48%	14.4%	30%
21	5/20/2018	5/26/2018	369	100%	2.0	1.4	65.3%	87%	25.2%	47%	0.9	-1.9	25.7%	36%	10.6%	20%
22	5/27/2018	6/2/2018	293	100%	2.0	-3.2	64.8%	84%	10.9%	23%	0.4	1.5	22.2%	32%	1.7%	4%
23	6/3/2018	6/9/2018	304	99%	2.0	3.4	51.0%	79%	14.8%	32%	0.4	1.0	11.8%	19%	1.6%	3%
24	6/10/2018	6/16/2018	234	100%	1.0	2.7	40.6%	64%	11.1%	25%	0.2	0.5	6.8%	9%	0.4%	1%
25	6/17/2018	6/23/2018	266	99%	1.0	2.8	46.6%	71%	10.9%	24%	0.2	0.7	9.8%	9%	0.4%	1%
26	6/24/2018	6/30/2018	266	100%	1.0	2.1	39.5%	59%	8.3%	21%	0.2	0.4	8.6%	9%	1.1%	2%
27	7/1/2018	7/7/2018	203	100%	2.0	3.3	62.6%	87%	11.8%	23%	0.2	0.8	13.3%	14%	0.5%	1%
28	7/8/2018	7/14/2018	270	99%	1.0	3.4	43.0%	63%	14.8%	30%	0.2	0.9	10.7%	12%	1.9%	2%
29	7/15/2018	7/21/2018	282	100%	2.0	3.6	59.6%	79%	13.8%	26%	0.3	1.1	17.0%	19%	0.7%	1%
30	7/22/2018	7/28/2018	285	99%	2.0	3.2	53.3%	79%	11.9%	23%	0.5	1.0	9.8%	14%	1.4%	3%
31	7/29/2018	8/4/2018	271	100%	1.0	1.8	34.7%	50%	10.3%	20%	0.2	-0.1	5.5%	5%	1.1%	2%
32	8/5/2018	8/11/2018	256	99%	1.0	1.1	45.7%	68%	16.8%	36%	0.3	0.8	5.9%	9%	1.6%	3%
33	8/12/2018	8/18/2018	264	99%	1.0	0.8	44.7%	58%	13.3%	20%	0.2	0.7	7.6%	8%	0.8%	1%
34	8/19/2018	8/25/2018	226	100%	1.0	2.7	41.2%	57%	10.6%	22%	0.2	0.6	9.7%	12%	0.0%	0%
35	8/26/2018	9/1/2018	339	100%	1.0	3.3	48.1%	64%	13.0%	26%	0.8	1.3	15.0%	18%	1.5%	3%
36	9/2/2018	9/8/2018	310	98%	3.0	4.4	65.8%	90%	18.4%	33%	0.9	1.9	28.7%	44%	1.0%	1%
37	9/9/2018	9/15/2018	304	100%	2.5	4.3	62.5%	91%	16.1%	34%	0.8	1.6	25.0%	34%	1.0%	2%
38	9/16/2018	9/22/2018	368	99%	2.0	3.7	62.8%	75%	16.0%	34%	0.9	1.6	21.7%	22%	1.1%	1%
39	9/23/2018	9/29/2018	344	95%	2.0	3.5	48.5%	78%	12.8%	30%	0.8	2.2	18.6%	29%	5.8%	9%
40	9/30/2018	10/6/2018	391	95%	2.0	3.2	56.0%	79%	14.8%	31%	1.0	2.0	22.8%	34%	4.1%	6%
41	10/7/2018	10/13/2018	311	84%	2.0	3.3	57.9%	90%	12.2%	40%	1.0	2.3	24.7%	47%	8.4%	18%
42	10/14/2018	10/20/2018	252	73%	1.0	1.3	31.4%	62%	6.0%	0%	0.8	1.1	17.9%	9%	0.4%	0%

What are some reflections around the policies associated with the AFST?

Two policy reflections jump to the fore: (1) whether Allegheny County made the right decision to limit the score to call screeners/supervisors and whether this is still the right decision and (2) whether high- and low-risk protocols are sufficient.

(1) Allegheny County leadership took a conservative approach to the use of the AFST, determining that the score was only to be used by call screeners and call screening supervisors, with no exceptions. We've been successful in applying this approach and think it was the right decision. However, now that we are more than two years into the process, we see improvements in call screening decision-making, but the established process still leaves far too many high-risk cases that are either not accepted for services or not triaged properly once accepted for services. In recognition of this reality, beginning in spring of 2019, DHS will explore how the score might be used elsewhere in the child welfare process. We will do this work, as we have in the past, thoughtfully and with engagement with experts and community leaders. As part of this exploration we will consider:

- whether the AFST should be provided to the clinical manager overseeing investigations to help him/her determine the response time and staffing.
- whether to use the score as one additional way to identify cases that undergo our quality assurance reviews (compliance and/or quality reviews).
- whether the score should be available to investigative supervisors to help them ensure due diligence on high-risk cases.
- whether the score could/should replace the state required risk assessment.

Any additional use of the AFST must be weighed carefully to assess the value of its ability to help us protect children and support families versus the risk of providing undue weight to one approach or reinforcing our own system behavior. As in the past, we will have to consider the way in which the system currently makes these determinations and whether the AFST can help improve that process (and the outcomes), acknowledging that such a model will never be perfect.

(2) High- and low-risk protocols: Because of concern that the score would have too much power in decision-making, we implemented "nudges," which defaulted the highest-risk cases to be screened in and required supervisors to explicitly override the decision with written justification if they felt it should not be investigated; a similar default-based nudge with override capability was later added to the lowest-risk cases. These nudges have led to only minimal additional concurrence with the model. We are looking at whether we should take a stronger approach to achieve more concurrence on very high-risk and very low-risk cases (acknowledging that the low-risk protocol has only been in place since November 2018). One particular reason that the high-risk protocol is only followed in about sixty-one percent of GPS cases is because many of these children are older and have allegation reasons that do not feel like abuse/neglect to the call screening staff. The model views

them as high-risk because of the considerable child protective services history and history of other supports, and the validation (using hospital data) confirms that these children face elevated risk of serious injuries, including self-inflicted injuries. Given this information, we should consider how DHS, as an integrated human services department, can divert these youth from the child welfare system (within the child protective services law) into a set of supports better aligned to meet their ongoing service needs. This is an ongoing challenge that requires additional work.

Did these technical, practice and other challenges impact the results of the evaluation?

It's not clear if these challenges impacted the results of the evaluation, but it's possible the results would be more robust and attenuate less absent these challenges. That's why we'll continue to improve quality assurance, monitor our work and continue independent evaluation.

Has the policy landscape around the implementation of predictive risk modeling changed since DHS began this work?

Yes, all of the fields surrounding this issue are in rapid evolution. When we started this work, there was no handbook on how to develop algorithms in the public interest and today there are numerous checklists, guidebooks and research groups established to help governments deploy predictive analytics in human services. The machine learning field is also rapidly evolving as are the official definitions of algorithmic fairness and discrimination in modeling. Allegheny County has attempted to both keep pace with these evolutions and to continue to improve our work based on these advancements. It is likely that we'll look back on the earliest models and see them for their flaws, but it is better to judge them on their improvement over previous practice and for our ability and willingness to continuously examine and improve.

You have reported outcomes for the first year of implementation —can you provide results for the full period under AFST Version 1?

Yes, for the period December 1, 2016–November 29, 2018 (and observed through 3/8/2019 data entry):

	PERCENT SCREENED-IN FOR INVESTIGATION	PERCENT OF THOSE SCREENED IN FOR INVESTIGATION THAT WERE ACCEPTED FOR SERVICE
Mandatory*	61%	45%
High	47%	41%
Medium	42%	37%
Low	31%	35%
No Score	23%	29%
Total	41.4%	39.1%
Pre AFST Comparison**	45.5%	34.3%

** Note that "mandatory" screen-ins may still be screened out at the discretion of the call screener and call screener supervisor.*

***December 1, 2014– November 29, 2015, selected as a comparison period because it is the most recent full calendar year (pre-AFST) with the same seasonal distribution of the observed AFST period above.*

See **Appendix A** for more detailed data.

APPENDIX A: AFST VERSION 1 SCREENING SCORE DATA, 12/1/16 TO 11/29/18

Dec. 1, 2016 - Nov. 29, 2018 GPS Screening Score Statistics

	Family Screening Score		GPS Referrals	Screening Outcomes											Investigation Outcomes			
			Count	Total Scrn-In	Total Scrn-Out	Total w/ Decision	Pct. Scrn-In	Tier Pct.	Field-Screen Unit Assigned	Pct. FS	Tier Pct.	FS Scrn-In	FS Scrn-Out	*Referrals on Active Family*	Investigated to Date*	Accepted for Service*	Pct. Accepted	Tier Pct.
	Mandatory	14%	3457	1402	907	2309	61%	61%	209	6%	6%	30	158	*1102*	1283	580	45%	45%
High Scores	20**		1488	471	561	1032	46%		118	8%		16	97	*440*	461	186	40%	
	19**		1093	371	473	844	44%		109	10%		14	85	*236*	364	142	39%	
	18**	36%	1112	384	525	909	42%	47%	122	11%	10%	16	103	*195*	377	145	38%	41%
	17		1728	722	755	1477	49%		188	11%		16	167	*241*	702	297	42%	
	16		1668	702	764	1466	48%		188	11%		24	152	*191*	684	295	43%	
	15		1551	700	726	1426	49%		181	12%		16	157	*115*	686	291	42%	
Med. Scores	14		1373	592	672	1264	47%		165	12%		11	147	*99*	585	222	38%	
	13		1147	453	611	1064	43%		137	12%		12	115	*69*	449	186	41%	
	12	22%	1133	425	631	1056	40%	42%	147	13%	13%	14	126	*68*	421	137	33%	37%
	11		913	356	499	855	42%		125	14%		9	109	*49*	351	118	34%	
	10		826	281	499	780	36%		114	14%		6	103	*40*	280	103	37%	
Low Scores	9		821	291	494	785	37%		106	13%		7	92	*30*	289	113	39%	
	8		690	200	464	664	30%		89	13%		3	79	*18*	197	74	38%	
	7		556	190	350	540	35%		68	12%		3	63	*14*	190	70	37%	
	6		634	193	424	617	31%		72	11%		5	64	*15*	193	69	36%	
	5	18%	540	127	402	529	24%	31%	59	11%	11%	2	54	*8*	127	37	29%	35%
	4		372	111	254	365	30%		44	12%		5	38	*5*	111	27	24%	
	3		330	89	235	324	27%		24	7%		3	20	*5*	88	29	33%	
	2		164	39	122	161	24%		11	7%		0	10	*2*	37	16	43%	
	1		124	37	87	124	30%		7	6%		0	7	*0*	37	8	22%	
	No Score	10%	2410	527	1790	2317	23%	23%	218	9%	9%	11	198	*82*	523	152	29%	29%
	Total		**24130**	**8663**	**12245**	**20908**	**41.4%**		**2501**	**10.4%**		**223**	**2144**	*3024*	**8435**	**3297**	**39.1%**	
											FS Scr. Rate:	*9.4%*	*90.6%*					

*** Scores of 18, 19, and 20 in this analysis must be driven by the re-referral model, since placement scores of 18+ would instead appear under the "Mandatory" row; this may explain differences in prevalence and screening rates*

**Investigations that end in attaching to an open case are omitted from counts; only decisions on potential "new cases" counted. As a result, "Investigated to Date" will be slightly smaller than the number screened-in for investigation*

Notes: Referral file as of 3/8/2019. Produced by ACDHS-DARE. At point of extract, a few dozen referrals from prior date ranges were stored in database "holding tables" and may not be included. This can indicate data entry lag for presumed screen-outs, but it can also include broken/aborted/accidental referrals that should be omitted anyway.
A data fix implemented on November 29, 2016 changed the tools in ways that increased the prevalence of higher (~18-20) scores; these figures span both before and after the change.
Please not that "Mandatory" scores are defined as referrals that score an 18-20 on the placement model, and thus the rows 18, 19, and 20 above are driven by the "re-referral" model (and cannot have had an 18+ on the placement model).

Dec. 1, 2014 - Nov. 29, 2015 (Pre-AFST) GPS Comparison Sample Period

	GPS Referrals	Screening Outcomes					Investigation Outcomes		
	Count	Total Scrn-In	Total Scrn-Out	Total w/ Decision	Pct. Scrn-In	*Referrals on Active Family*	*Investigated to Date**	*Accepted for Service**	*Accept %*
Total	**9742**	**4148**	**4963**	**9111**	**45.5%**	*627*	**3903**	**1340**	**34.3%**
Avg. Yr Change	2323	183.5	1159.5	1343	-4.1%	885	314.5	308.5	4.8%
Change (%)	23.8%	4.4%	23.4%	14.7%	--	141.1%	8.1%	23.0%	--