

SECTION 6

Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office

Jeremy D. Goldhaber-Fiebert, PhD and Lea Prince, PhD, Stanford University
March 20, 2019

¹ Allegations fall under the state of Pennsylvania's Child Protective Service (CPS) statutes (23 Pa.C.S. § 6303) or General Protective Service (GPS) statutes (23 Pa.C.S. § 6334). 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 well-being 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 2017, 21 percent of DHS referrals were CPS referrals and 79 percent were GPS referrals.

EXECUTIVE SUMMARY

Goal: The impact evaluation assesses how implementation of the screening score (Allegheny Family Screening Tool or AFST) within Allegheny County's child welfare office helped to:

- Improve accuracy of referrals by call screeners (increasing the fraction of children who screen-in with further action taken upon investigation and the fraction of children who screen-out with no re-referrals within 2 months)
- Maintain reasonable workload in terms of the rate of screen-ins (and subsequent investigations)
- Reduce disparities in terms of the above outcomes for similar children from age and race/ethnic subgroups
- Promote consistency in terms of the above outcomes across the call screeners

Approach: The evaluation uses a set of methodologically strong, quasi-experimental techniques (e.g., interrupted time series analyses, generalized linear models) to achieve the impact evaluation's goals. The study primarily compares outcomes for children involved in general protective service (GPS)¹ referrals in the 15 - 17 months after the full implementation of the AFST (December 1, 2016 through May 31, 2018) (~34,000 children) to outcomes for children involved in GPS referrals in the period before implementation, primarily January 1, 2015 through July 31, 2016 (~31,000 children). Further details of the approach appear in the **Methods** sections below.

Findings and Interpretation:

1. **Accuracy:** Implementation of the AFST and associated policies increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out.
 - Implementation of the AFST increased the proportion of children who screened-in for investigation and upon investigation either had further action taken or else were re-referred within 60 days. The larger initial effect appeared to partially attenuate over time.

Implementation of the AFST and associated policies may have slightly decreased accuracy in terms of the proportion of children screened-out who had no re-referrals within 60 days, with the majority of this small effect in children aged 4 to 6 years.

2. *Workload:* Implementation of the AFST and associated policies halted the downward trend in the rate of children screened-in for investigation.
3. *Disparities:*
 - *Accuracy:*
 - Further action taken or re-referral within 60 days after being screened-in for investigation: There were larger increases in accuracy of being screened-in and/or less attenuation of the effect over time for white children and children aged < 4 years. In contrast, the initial improvement in accuracy attenuated more rapidly for Black/African American children.
 - Re-referral within 60 days after being screened-out: There were greater losses of accuracy of being screened-out for children ages 4 to 6 years though the overall size of effect even in this age-group was relatively small. This may be due to concurrent changes in the mandatory in-home assessment (field screening) policy in terms of the maximum age being reduced from under 7 to under 4 years of age.
 - *Workload:* The effect of the AFST and associated policies of halting the downward trend in the rate of children screened-in for investigation was larger for older children (e.g., ages 13 to 17) and for Black/African American children. The effect was smallest for children ages 4 to 6 years where a screen-in for investigation may have replaced field screening which was no longer required for this age-group.
4. *Consistency:* For the subgroup of 11 call screeners handling a substantial volume of referrals in both the Pre-AFST and Post-AFST implementation periods, the AFST and associated policies did not significantly alter the consistency of outcomes relating to accuracy or workload across call screeners. Likewise, the AFST did not significantly alter age group-specific or race group-specific consistency for any of these outcomes. Of note, particularly for call screener consistency outcomes by age-group or race group, there was likely insufficient sample size (power) to detect changes.

CONTENTS

Methods	10
Overview	10
Implementation of the AFST	10
Other Changes Occurring with the Implementation of the AFST	10
Outcomes	11
Accuracy Outcomes	11
Workload Outcomes	12
Consistency Outcomes	13
Disparities Outcomes	13
Hypothesized Effects of the Implementation of the AFST Score and Related Policies	14
Data	15
Analytic Approach	15
Overview	15
Description of Specific Methods	16
Covariates and Standard Errors	18
Description of Trends in the Study Population and Changes to Case Mix over Time	19
Results	19
Accuracy Outcomes and Disparities in Accuracy Outcomes	19
How did the AFST change accuracy for referrals screened-in?	19
How did the AFST change accuracy for referrals screened-in related to children in different age-groups?	20
How did the AFST change accuracy for referrals screened-in related to children in different race groups?	20
How did the AFST change accuracy for referrals screened-out?	20
How did the AFST change accuracy for referrals screened-out related to children in different age-groups?	21
How did the AFST change accuracy for referrals screened-out related to children in different race groups?	21

Workload Outcomes and Disparities in Workload Outcomes	21
How did the AFST change workload as measured by the fraction of referrals screened-in for further investigation?	21
How did the AFST change workload related to children in different age-groups?	22
How did the AFST change workload related to children in different race groups?	22
Consistency Outcomes and Disparities in Consistency Outcomes	22
How did the AFST change the consistency of accuracy for referrals screened-in by call screener?	22
How did the AFST change the consistency of accuracy for referrals screened-in by call screener by age-groups?	23
How did the AFST change the consistency of accuracy for referrals screened-in by call screener by race groups?	23
How did the AFST change the consistency of accuracy for referrals screened-out by call screener?	23
How did the AFST change the consistency of accuracy for referrals screened-out by call screener by age-groups?	23
How did the AFST change the consistency of accuracy for referrals screened-out by call screener by race groups?	24
How did the AFST change workload differentially by call screener?	24
How did the AFST change workload differentially by call screener by age-groups?	24
How did the AFST change workload differentially by call screener by race groups?	25
Discussion, Conclusions, and Implications	25
TABLES	28
TABLE 1A: Summary Statistics, child characteristics	28
TABLE 1B: Summary Statistics, household characteristics	29
TABLE 2: Means of outcomes	30
TABLE 3A: Accuracy of screen-in, ITSA analysis, all children	31
TABLE 3B: Accuracy of screen-in, ITSA analysis, < 4 years old	31
TABLE 3C: Accuracy of screen-in, ITSA analysis, 4 to 6 years old	32
TABLE 3D: Accuracy of screen-in, ITSA analysis, 7 to 12 years old	32
TABLE 3E: Accuracy of screen-in, ITSA analysis, 13 to 17 years old	33
TABLE 3F: Accuracy of screen-in, ITSA analysis, White	33
TABLE 3G: Accuracy of screen-in, ITSA analysis, Black/African American	34
TABLE 4A: Accuracy of screen-in, adjusted analysis, all children	34

TABLE 4B: Accuracy of screen-in, adjusted analysis, by age group	34
TABLE 4C: Accuracy of screen-in, adjusted analysis, by race	35
TABLE 5A: Accuracy of screen-out, ITSA analysis, all children	35
TABLE 5B: Accuracy of screen-out, ITSA analysis, < 4 years old	36
TABLE 5C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old	36
TABLE 5D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old	37
TABLE 5E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old	37
TABLE 5F: Accuracy of screen-out, ITSA analysis, White	38
TABLE 5G: Accuracy of screen-out, ITSA analysis, Black/African American	38
TABLE 6A: Accuracy of screen-out, adjusted analysis, all children	39
TABLE 6B: Accuracy of screen-out, adjusted analysis, by age group	39
TABLE 6C: Accuracy of screen-out, adjusted analysis, by race	40
TABLE 7A: Workload, ITSA analysis, all children	40
TABLE 7B: Workload, ITSA analysis, < 4 years old	41
TABLE 7C: Workload, ITSA analysis, 4 to 6 years old	41
TABLE 7D: Workload, ITSA analysis, 7 to 12 years old	42
TABLE 7E: Workload, ITSA analysis, 13 to 17 years old	42
TABLE 7F: Workload, ITSA analysis, White	43
TABLE 7G: Workload, ITSA analysis, Black/African American	43
TABLE 8A: Workload, adjusted analysis, all children	43
TABLE 8B: Workload, adjusted analysis, by age-group	44
TABLE 8C: Workload, adjusted analysis, by race	44
TABLE 9: Means of Outcomes (1) - (3) for call screeners included and excluded from Outcome 4/Consistency analyses	45
TABLE 10A: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners	45
TABLE 10B: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by age-group	46
TABLE 10C: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by race	47
TABLE 11A: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, for 11 included screeners	48
TABLE 11B: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by age-group	48

TABLE 11C: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by race 49

TABLE 12A: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener 49

TABLE 12B: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener, by age-group 50

TABLE 12C: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener, by race 51

TABLE 13A: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, for 11 included screener 52

TABLE 13B: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by age-group 52

TABLE 13C: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by race 53

TABLE 14A: Consistency in workload, adjusted analysis, for 11 included call screener 53

TABLE 14B: Consistency in workload, adjusted analysis, for 11 included call screeners, by age-group 54

TABLE 14C: Consistency in workload, adjusted analysis, for 11 included call screener, by race 55

TABLE 15A: Means and variance of screener's predicted workload, adjusted analysis, for 11 included screeners 56

TABLE 15B: Means and variance of screener's predicted workload, adjusted analysis, by age-group 56

TABLE 15C: Means and variance of screener's predicted workload, adjusted analysis, by race 57

TABLE 16A: Estimated magnitude of monthly impact of AFST on accuracy of screen-in 57

TABLE 16B: Estimated magnitude of impact of AFST on accuracy of screen-out 58

TABLE 16C: Estimated magnitude of impact of AFST on workload 58

FIGURES 59

FIGURE 1. Example of the AFST Score 59

FIGURE 2A: Accuracy of Screen-In, ITSA analysis 60

FIGURE 2B: Accuracy of Screen-In, by age-group 60

FIGURE 2C: Accuracy of Screen-In, by race 61

FIGURE 3A: Accuracy of Screen-In, adjusted analysis 61

FIGURE 3B: Accuracy of Screen-In, adjusted analysis, by age-groups 62

FIGURE 3C: Accuracy of Screen-In, adjusted analysis, by race 62

FIGURE 4A: Accuracy of Screen-Out, ITSA analysis	63
FIGURE 4B: Accuracy of Screen-Out, ITSA analysis, by age-group	63
FIGURE 4C: Accuracy of Screen-Out, ITSA analysis, by race	64
FIGURE 5A: Accuracy of Screen-Out, adjusted analysis	64
FIGURE 5B: Accuracy of Screen-Out, adjusted analysis, by age-group	65
FIGURE 5C: Accuracy of Screen-Out, adjusted analysis, by race	65
FIGURE 6A: Workload, ITSA Analysis	66
FIGURE 6B: Workload, ITSA Analysis, by age-group	66
FIGURE 6C: Workload, ITSA Analysis, by race	67
FIGURE 7A: Workload, adjusted analysis	67
FIGURE 7B: Workload, adjusted analysis, by age-group	68
FIGURE 7C: Workload, adjusted analysis, by race	68
FIGURE 8A: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis	69
FIGURE 8B: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by age-group	69
FIGURE 8C: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by race	70
FIGURE 9A: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis	70
FIGURE 9B: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by age-group	71
FIGURE 9C: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by race	71
FIGURE 10A: Consistency of workload for 11 call screeners, adjusted analysis	72
FIGURE 10B: Consistency of workload for 11 call screeners, adjusted analysis, by age-group	72
FIGURE 10C: Consistency of workload for 11 call screeners, adjusted analysis, by race	73

APPENDIX 74

APPENDIX A1: Analytic dataset and variable construction	74
APPENDIX A2: Notes on interrupted time-series (ITSA)	76

Appendix Tables 77

TABLE A1A: Accuracy of Screen-in, ITSA analysis, all children, 6 month re-referral window	77
TABLE A1B: Accuracy of Screen-in, ITSA analysis, < 4 years old, 6 month re-referral window	77

TABLE A1C: Accuracy of Screen-in, ITSA analysis, 4 to 6 years old, 6 month re-referral window 78

TABLE A1D: Accuracy of Screen-in, ITSA analysis, 7 to 12 years old, 6 month re-referral window 78

TABLE A1E: Accuracy of Screen-in, ITSA analysis, 13 to 17 years old, 6 month re-referral window 79

TABLE A1F: Accuracy of Screen-in, ITSA analysis, White, 6 month re-referral window 79

TABLE A1G: Accuracy of Screen-in, ITSA analysis, Black/African American, 6 month re-referral window 80

TABLE A2A : Accuracy of screen-in, adjusted analysis, all children, 6 month re-referral window 80

TABLE A2B: Accuracy of screen-in, adjusted analysis, by age group, 6 month re-referral window 80

TABLE A2C: Accuracy of screen-in, adjusted analysis, by race, 6 month re-referral window 81

TABLE A3A: Accuracy of screen-out, ITSA analysis, all children, 6 month re-referral window 81

TABLE A3B: Accuracy of screen-out, ITSA analysis, < 4 years old, 6 month re-referral window 82

TABLE A3C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old, 6 month re-referral window 82

TABLE A3D: Accuracy of screen-out, , ITSA analysis, 7 to 12 years old, 6 month re-referral window 83

TABLE A3E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old, 6 month re-referral window 83

TABLE A3F: Accuracy of screen-out, ITSA analysis, White, 6 month re-referral window 84

TABLE A3G: Accuracy of screen-out, ITSA analysis, Black/African American, 6 month re-referral window 84

TABLE A4A: Accuracy of screen-out, adjusted analysis, all children, 6 month re-referral window 84

TABLE A4B: Accuracy of screen-out, adjusted analysis, by age group, 6 month re-referral window 85

TABLE A4C: Accuracy of screen-out, adjusted analysis, by race, 6 month re-referral window 85

TABLE APPENDIX A5: Regression results for further action or no further action and re-referral within 60 days, conditional on screen-in (Outcome 1: accuracy of screen-in) 86

TABLE APPENDIX A6: Regression results for no re-referral within 60 days, conditional on screen-out (Outcome 2: accuracy of screen-out) 88

TABLE APPENDIX A7: Regression results for screen-in (Outcome 3: workload) 90

Appendix Figures 92

APPENDIX FIGURE 1A: Total children in referral calls, by month 92

APPENDIX FIGURE 1B: Total children in referral calls, by month and age-group 92

APPENDIX FIGURE 1C: Total children in referral calls, by month and race 93

APPENDIX FIGURE 2A: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis 93

APPENDIX FIGURE 2B: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by age-group 94

APPENDIX FIGURE 2C: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by race 94

APPENDIX FIGURE 3A: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis 95

APPENDIX FIGURE 3B: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by age-group 95

APPENDIX FIGURE 3C: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by race 96

APPENDIX FIGURE 4A: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis 96

APPENDIX FIGURE 4B: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by age-group 97

APPENDIX FIGURE 4C: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by race 97

METHODS

Overview

The impact evaluation analyzes how the introduction of a screening score (AFST) for use by Allegheny County Child Welfare Office's intake office as part of the decision-making process for children involved in GPS referrals combined with a set of policy and practice changes affected several important outcome measures relating to accuracy, workload, disparities in accuracy and workload, and consistency in these outcomes across call screeners. The sections below describe the AFST and its implementation, the outcome measures used, the policy changes the evaluation accounts for, the data used in the evaluation, and the analytic approaches chosen to perform the evaluation along with their rationale.

Implementation of the AFST

For each referral (involving one or more child in a household) after the implementation of the AFST, call screeners are presented with a visual which either indicates a mandatory screen-in or displays the AFST score (**Figure 1**). The latter is presented as a tool to aid the call screener in making recommendations about screening decisions regarding further investigation, along with the set of tools that the screeners used prior to implementation. Screening recommendations are made on any referral which is classified as GPS. Generation of the score is based on data related to the individual clients for each referral, which includes the victim child(ren), siblings, parents, legal guardians, perpetrators, and potentially unrelated children and adults in the home. Recommendations are made by the Allegheny County hotline staff (screeners and supervisors) and follow one of three courses: 1) Screen-out of a referral without any further evaluation or assessment, 2) Field screen of the referral to assess whether an investigation is warranted, or 3) Screen-in of a referral, which is synonymous with conducting a formal investigation. A field screen refers to an in-home assessment at the referral household.² A field screen is always followed by a decision to either screen-out or screen-in the referral.

2 A field screen is mandatory for a set of conditions; see Other Changes Occurring with the Implementation of the AFST below for details.

3 Of note, an auto screen-in does not obligate the call screener and/or supervisor to screen-in, but rather implies that this would be the default action.

At the time of the referral, each individual associated with the referral is assigned both a re-referral and a placement risk score. The AFST, which is the only score that the screener has access to, is based on the maximum score (either re-referral or placement) across all individuals associated with the referral at the time that the referral first occurs. The score has a range of 1 to 20 (where 20 is the highest "risk" and 1 is the lowest), indicating the quantile into which the AFST falls. An "auto screen-in" occurs when the AFST falls above 18 for the placement score.³

Other Changes Occurring with the Implementation of the AFST

Several other systematic changes to the call screening process accompanied the full implementation of the AFST. First, the mandatory field screen policy was updated, for households which are not already slated to screen-in and which meet a set of criteria. The maximum age for a mandatory field screen decreased from 7 (previously, households with at least one child age 0 through age 6 years required a field screen, regardless of history) to 4 years of age. In addition, the new

mandatory field screen policy adds the following three conditions: (1) all children who attend home school/cyber school receive a mandatory field screen regardless of age; (2) any family that has had 4 or more referrals in 2 years without any of the referrals being formally investigated; and (3) anything else where more information is necessary to make a final decision. Importantly, second, call screeners make a recommendation about the decision to screen-in/out to their supervisor who has the responsibility for the ultimate decision. Prior to this set of policy changes, the process of assessment and recommendation involving call screeners and supervisors differed from that described above. Prior to the implementation of the AFST, the primary role of the call screening staff was to gather information to inform supervisor decision-making. Call screeners collected data from several databases and resources, including internal DHS systems (KIDS, Client View), courts, public assistance and criminal justice. Call screeners also spoke with the individual making the report and other key contacts (e.g., schools, doctors). The information collected was given to supervisors for final decision-making. Although the process between screeners and supervisors was collaborative, following implementation of the AFST, call screeners took on a greater role in making recommendations for screening decisions and this process was incorporated into the KIDS system.

Outcomes

We selected outcomes to measure underlying effects of the AFST implementation in terms of accuracy of the call screening process for children involved in referrals, workload entering the system, and disparities across children's age and race/ethnicity in terms of accuracy and workload. We also examined the consistency of these outcomes across call screeners. Specifically, the analysis examines how the implementation of the AFST may have impacted these multiple outcomes, including:

ACCURACY OUTCOMES

Outcome 1 — Likelihood of a screen-in with action taken upon investigation or no further action taken and a re-referral within 2 months: A child is considered to be in this category if a referral that includes the child is screened-in and upon investigation the disposition is further action taken or there is no further action taken and a re-referral occurs within the 2-month time window starting from when the call was referred. "Further action" is defined by the referral service decision and occurs when a referral accepts for service or connects to either an open case or connects to a closed case and is re-opened for service. Because of the re-referral period, the last 2 months (April and May 2018) of calls in the data are not allowed as "index events" since there will not be complete follow-up in the data for re-referrals. These last 2 months of data are only used to determine whether re-referrals occurred or not for calls in the prior months. The rate of such screen-ins was defined as the number of children falling into this category divided by the total number of children screened-in for investigation, computed for referrals falling in each calendar month. We repeat the Outcome 1 analysis, using a 6-month re-referral window, as a robustness check and include the results in the Appendix. This outcome is computed for all children. *AFST-related changes in Outcome 1 are intended to measure how AFST implementation*

impacted one feature of accuracy: do those who are screened-in for investigation have further action taken or if not do re-referrals within several months indicate ongoing issues that the initial screen-in may have been sensitive to?

Outcome 2 — Likelihood of a screen-out with no re-referrals within a 2-month time window:

A child is considered screened-out if the report that includes the child is not referred by the call screener for further investigation. If another referral occurs within the time window (i.e., within 2 months of the referral call), then the child is considered to have been screened-out with a re-referral. Because of the re-referral period, the last 2 months (April and May 2018) of calls in the data are not allowed as “index events” since there will not be complete follow-up in the data for re-referrals. These last 2 months of data are only used to determine whether re-referrals occurred or not for calls in the prior months. Of note, a child can have more than one screen-out and re-referral over time, but the “index event” of a call that can be considered a screen-out is only assessed for any subsequent calls outside of the re-referral time window of the previous “index event”. We define the rate of screen-outs with no re-referrals within a 2-month time window as the number of children in reports who were not referred by the call screener for further investigation and were subsequently not re-referred within a given number of months divided by the total number of children in reports, computed for referrals falling in each calendar month. This outcome is computed for all children. We repeat the Outcome 2 analysis, using a 6-month re-referral window, as a robustness check and include the results in the appendix. AFST-related changes in Outcome 2 are intended to measure how AFST implementation impacted one feature of accuracy: do those who are screened-out remain unassociated with subsequent referrals?

An analogy can be made between these accuracy outcomes for both screen-ins and screen-outs and the more general concepts and language of screening test assessment. In general screening test terminology, test accuracy is measured based on sensitivity (i.e., true positive fractions) which is the percentage of those subjects with the underlying condition who test positive and specificity (i.e., the true negative fraction) which is the percentage of those subjects without the underlying condition who test negative. Ultimately, the ideal is to have a test with a high positive predictive value, the fraction of test positives that have the underlying condition and the fraction of test negatives who do not have the underlying condition. In our context, accuracy for screen-ins (i.e., test positives) is the fraction of children who screen-in that have further action taken upon investigation (an indicator of the underlying condition). Likewise, accuracy for screen-outs (i.e., test negatives) is the fraction of children who screen-out that have no re-referrals for 2 months (an indicator of the absence of the underlying condition). The hypothesized effect of the AFST score would be to increase both of these.

WORKLOAD OUTCOMES

Outcome 3 — Rates of calls screened-in for investigation: A child is considered screened-in for investigation if the referral (household) that includes the child is referred by the call screener for further investigation. Therefore, we define the rate of “screened-in for investigation” as equal to

the number of children in referrals assigned to further investigation divided by the total number of children in referrals, computed for calls falling in each time interval (i.e., calendar month). This outcome is computed for all children. AFST-related changes in Outcomes 3 are intended to measure how AFST implementation impacted the workload entering the investigative system.

CONSISTENCY OUTCOMES

Outcome 4 — Consistency in call screener actions, as related to Outcomes 1–3: We estimate Outcomes 1–3 by call screener. This outcome is computed for calls screeners associated with at least 350 referrals before and after the implementation of the AFST to enable reasonably stable call screener specific estimates. AFST-related changes in these call screener-specific outcomes are intended to measure how AFST implementation impacted consistency in accuracy and in workload (as measured by Outcomes 1–3).

DISPARITIES OUTCOMES

Across Outcomes 1–4: We estimate Outcomes 1-4 for age and race/ethnic subgroups to examine how AFST-related changes in them differed across these subgroups. Such AFST-related changes are intended to measure how AFST implementation impacted disparities.

Stratification of outcomes by age-group included four age-groups: <4 years, 4 to 6 years, 7 to 12 years, 13 to 17 years.⁴ While other/undetermined race categories are controlled for in the main analysis for Outcomes 1–3, the disparities analysis is limited to stratification by white and Black/African only, which include over 90% of children included in referrals.⁵

The entire period of analysis spans August 1, 2013 through May 31, 2018 and focuses on how outcome measures changed as a result of the full implementation of the AFST (December 1, 2016). Hence, for the analyses we divide the data into multiple periods. The period after the full implementation period is termed the “Post-AFST Period” and its outcomes are compared to outcomes in periods prior to this. The time prior to the implementation of the AFST spans August 1, 2013 through July 31, 2016 and is divided in two parts. It is divided into the period August 1, 2013 through December 31, 2014 and the period January 1, 2015 through July 31, 2016. The decision to divide the of pre-AFST period is based on a set of amendments to the State of Pennsylvania’s existing Child Protective Services Law, which became effective on December 31, 2014 and had the effect of altering a number of features of referrals to the call center.⁶ We focus on this second, later pre-AFST period as the point of comparison in our analysis, and refer to this as the “Pre-AFST” from here. Notably, data for the period between August 1, 2016 and November 30, 2016 are omitted from all analyses. When the AFST was launched, an initial policy decision sought to restrict score generation to only individuals and families who could be substantively identified in prior county data (preventing scores from being displayed that were solely constructed from basic referral/geographic information when the family is otherwise unknown to DHS). The first iteration of this policy initially restricted scores to only situations where a scored child on the call needed to be positively identified with a prior county identifier. Many children, most notably newborns and young babies who are most at risk, often do not have

4 In the individual-level analyses described further below, the analytic models adjust for household composition, including the number of children within a household who are under 1 years of age to control for unobserved effects of very young children on Outcomes 1–3. Due to strong interest in outcomes related to the youngest children, future planned analyses will disaggregate the current age-specific subgroups such that children <1 years and children 1–4 years old will be included in two separate categories when sufficient subgroup-specific sample size has accrued.

5 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.

6 Child Protective Services Act, P.L.1240, No.206, 23 PA §6301-6386 (2015)

system involvement, but their parents or caregivers may have significant current and prior system involvement. The initial design led to situations where known information about adults on the referral could not be used in generating a score if none of the children were recognized, and this was quickly deemed too restrictive. After November 30, 2016, scores could be generated on a call if any individual on the call had a past county identifier. Given the reason for the differential missing-ness of the AFST score, analyses of these data cannot rely on missing at random assumptions, as selection mechanisms could operate to make outcomes higher or lower than both the pre-implementation and post-implementation levels that would not be genuinely attributable to the AFST score implementation per se.

Hypothesized Effects of the Implementation of the AFST Score and Related Policies

Accuracy Outcomes (Outcomes 1 and 2): We hypothesized that if implementation performed as expected, both accuracy outcomes would increase (i.e., a higher proportion of screen-ins with have further action taken upon investigation and a higher proportion of screen-outs would have no re-referrals in the subsequent 2 months). The rationale for this hypothesis is that the score is intended to quantitatively integrate a great deal of available information that is predictive of placement and re-referral probabilities, with such information presumed relevant for screen-in/out decisions.

Workload Outcome (Outcome 3): We hypothesized that there would be no substantial change in workload if implementation performed as expected, though we acknowledged that this hypothesis was weaker than for some of our other outcomes. We believed that the mix of which referrals were screened-in and which were screened-out might change with the hypothesized improvements in accuracy but had no reason to believe that there would be a higher (lower) proportion of total calls screening-in due to the implementation.

Consistency Outcome (Outcome 4 [Outcomes 1–3 by Call Screener]): We hypothesized that if implementation performed as expected, consistency in both accuracy and workload outcomes across call screeners would increase (i.e., variation in outcomes across call screeners would decrease). The rationale for this hypothesis is that the AFST score will be the same for a referral regardless of to which call screener it is displayed, and hence, it will provide a regularized/standardized input/measure to help to inform the screen-in/out decisions.

Disparity Outcomes (Across Outcomes 1–4): We hypothesized that if implementation performed as expected, disparities in outcomes across age and race groups would diminish, though we acknowledged that this hypothesis than for some of our other outcomes. For example, if in the pre-implementation period children in two age groups had similar rates of screen-ins that upon investigation led to further action, but the AFST was better able to predict for one age group than another, then in the post-implementation one age group's accuracy would increase differentially from the other's and the disparity between groups could widen.

Data

All analyses use de-identified data relating to those involved in referrals to Allegheny County's call center. The data consist of information about individual household members including race, legal sex, and age. Additionally, the data enumerate the call screeners and supervisors associated with the referral and track previous referrals and investigations with child welfare and other child-serving systems from August 1, 2013 through May 31, 2018.

The analytic dataset focuses on outcomes (described above) for children below age 18 years at the time of the referral. The analytic dataset also retains data, specifically relating to re-referrals, for children who turned 18 between the time of the initial referral call and the end of a 2-month window. The analytic dataset focuses on children with a referral type of GPS and excludes children with other referral types since GPS referrals allow discretion regarding the call screen decision relative to CPS-type referrals, which are mandated as screen-in.

The analytic dataset also contains several variables used in the analysis to control for child (age, legal sex, race) and household (household counts and composition, socioeconomic status and maximum risk scores) characteristics. For child race, we used a categorical variable which included the category "Unable to Determine", when race was not coded as white, Black/African American or other. Other control variables had complete data. See Appendix A1 for detail on the construction of variables.

Analytic Approach

Overview

For Outcomes 1 to 3, we report three main types of analyses, described in detail below. The first is a comparison of unadjusted population means for the Pre-AFST Period (January 1, 2015 through July 31, 2016) and the Post-AFST Period (December 1, 2016 through May 31, 2018). The second is an analysis of changes in the level and trend of monthly rates of the outcomes in these periods using an Interrupted Time Series Analysis (ITSA). The third is an individual-level multivariate regression analyses to estimate the impact of the AFST on the predicted level of each outcome both Pre- and Post-AFST. For each of these analyses, we consider the effect of the policy for the overall population of children with GPS referrals during the analytic period. For Outcome 4, we focus on the individual-level multivariate regression analyses and how they differed across call screeners and how these differences across call screeners changed in the Post-AFST period.⁷ To determine how outcomes for individual call screeners changed after the AFST, we examine the Pre-Post-AFST difference in the predicted outcomes. To determine how the AFST may have regularized and made outcomes more consistent for similar children across call screeners as a group, we compare the Pre- versus Post-AFST variance of predicted outcomes. Of note, the evaluation has the least power to detect differences in call screener consistency. Finally, we consider the effects of the policy implementation on Outcomes 1-4 on subgroups of children defined by their age-group and race/ethnicity characterization. This enables us to consider potential heterogeneity and disparities in the policy's effects across these subgroups. We perform all analyses using Stata (v14) software.

7 To determine consistency between and across call screeners, it is necessary to adjust for differences in case mixes, and therefore, the individual-level multivariate analysis is the appropriate one to focus on. Furthermore, the frequency of referrals per month, by screener, is not consistently high enough to provide statistically meaningful interpretation of call screener-specific ITSA analysis results.

The rationale for performing multiple analyses for Outcomes 1 to 3 is that findings from them are complementary and help to highlight various features of the AFST's effects. Comparisons of unadjusted means and the ITSA analyses describe levels and changes in outcomes within the system overall and for age and race/ethnic subgroups given existing trends in the Pre-AFST period. Individual-level regression analyses focus on changes in levels after adjustment for changes in referral case-mix over time. See "Description of Trends in the Study Population and Changes to Case Mix over Time" below for further details.

Description of Specific Methods

First, the simplest comparison we perform is the comparison of unadjusted means in the Pre-AFST and Post-AFST period, testing whether they are statistically different from one another using a two-sided t-test of equality of means.

Second, Interrupted Time Series Analysis is a proven quasi-experimental research design which is particularly useful in the evaluation of a program change when a randomized trial is infeasible and/or unethical. In the evaluation of the AFST, we perform ITSA on a series of monthly rates of each outcome divided into policy periods as described above. 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. The ITSA approach captures population-level changes in outcomes and trends after a policy change in comparison to the levels and trends prior to that change. For our application, the biggest strength of the ITSA approach is the ability to test changes in trends because we observe clear time trends in outcomes in the Pre-AFST data. Making causal inference with ITSA relies on several assumptions with the most important being that the rates of change in outcomes from other causes (secular trends) are much slower than changes due to the abrupt implementation of the policy of interest (see **Appendix A2**).⁸

For our outcomes, we made ITSA model estimates of the form:

$$Outcome_t = \beta_0 + \beta_1 time_t + \beta_2 policy_{1t} + \beta_3 policy_{1t} \times time_t + \beta_4 policy_{2t} + \beta_5 policy_{2t} \times time_t + e_t$$

where $Outcome_t$ is the outcome variable (monthly mean), $time_t$ is the time since the start of the data series, $policy_{1t}$ and $policy_{2t}$ are binary indicators (0 prior to policy implementation, otherwise 1) for: 1) December 31, 2014 amendments to the State of Pennsylvania's existing Child Protective Services Law; 2) December 1, 2016 full implementation of the AFST. The coefficient β_0 represents the intercept, or starting level of the outcome; β_1 represents the slope (or trend) in the outcome prior to any of the policies considered, β_2 and β_4 represent the change in intercept (the immediate impact on levels caused by each policy) and β_3 and β_5 capture the difference in the trends after each policy, respectively; and e_t represents stochastic error. In the ITSA analysis, outcomes are modeled as rates calculated by month (t). The multi-period/multi-policy cumulative changes to level/intercept and trend can be estimated as an extension to this ITSA model and tested for statistical significance. We use the *ITSA* command in Stata.⁹

8 Within the context of ITSA, it is possible to use either an Ordinary Least Squares (OLS) or autoregressive integrated moving-average model (ARIMA) approach. The default *itsa* command in Stata uses an OLS regression model instead of an ARIMA model because OLS tends to be more flexible and interpretable in an interrupted time-series setting than an ARIMA (Box and Jenkins 1976; Velicer and Harrop 1983) (see Also **Appendix A2**).

9 The *itsa* command allows the user to specify the number of lags to control for autocorrelation. We test for the correctness of our specification using *actest* which performs the Cumby-Huizinga general specification test of serial correlation.

We estimate similar *ITSA* models for age-specific subgroups of children and for race/ethnic subgroups.

Third, we used multivariate individual-level regression analyses to assess outcomes while adjusting for child and household characteristics. Specifically, we estimate Generalized Linear Models (glm in Stata) with a logit link enabling greater flexibility in the distribution of the error term than a standard logit model. We run our analysis at the level of any child involved in any call (i.e., not only the child for which the call was made):

$$Outcome_{it} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + e_{it}$$

Where $Outcome_{it}$ is a binary indicator equaling 1 if the child is either screened-in with further action taken upon investigation or re-referral within 2 months (Outcome 1), screened out with no re-referral within 2 months (Outcome 2), or screened in (Outcome 3). X_i is a vector of child and household characteristics and $policy_t$ is a binary indicator for calls after the AFST was implemented and e_{it} is a stochastic error term. The multivariate analysis is at the child level (i) with time (t) represented as continuous days across the Pre- and Post- AFST periods. The policy period is coded as 0 if the referral took place between January 1, 2015 and July 31, 2016 and coded as 1 if the referral call took place after December 1, 2016.

We also examine how race and age-groups are differentially impacted by the AFST, with the following model:

$$Outcome_{it} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 policy_t \times Var_i + e_{it}$$

Where Var_i is either race- or age-specific subgroup. Note, race- or age-specific subgroups are also included in the full vector of child/household characteristics, X_i .¹⁰

We use the margins command in Stata to compute the predicted level of each outcome, both Pre-and Post-AFST implementation. These analyses do not evaluate Pre-AFST or Post-AFST time trends as in the *ITSA* analyses but rather, they focus on estimates of the average effect of the AFST adjusting for evolving case mix over time. The predictive margins presented in tables and figures of the results can be interpreted as the average outcome if all children in the sample were in either the Pre-AFST or the Post-AFST period, holding all other control variables as they happen to be.

¹⁰ Because there is specific interest in outcomes related to black/African American children versus white children, the model testing differential impacts by race excludes children coded as "other" or "undetermined" (~10%). In all other models, all children were included in the analytic sample.

To examine consistency of call screener actions, related to Outcomes 1–3, we limit our sample to those call screeners with ≥ 350 calls in each of the Pre- and Post-AFST and perform a similar set of multivariate analyses using the model:

$$Outcome_{ijt} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 Screener_j + \beta_4 policy_t \times Screener_j + e_{ijt}$$

Where $Screener_j$ is a screener fixed-effects variable, which defines outcomes based for each individual call screener (j).

For the consistency outcomes, we are interested in how much the screener effects varied in the Pre-AFST period (i.e., variance of the predicted means/margins) and how much they varied in the Post-AFST period and testing for whether changes in variance were statistically significant.

Again, we examine how race- and age-groups are differentially impacted by the AFST by call screener with the following model:

$$Outcome_{ijt} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 Screener_j + \beta_4 policy_t \times Screener_j + \beta_5 policy_t \times Var_i + \beta_6 Screener_j \times Var_i + \beta_7 policy_t \times Screener_j \times Var_i + e_{ijt}$$

As above, we estimate call screener variance and test for changes in call screener variance in performance on the outcomes within each group.

Covariates and Standard Errors

For the models involving multivariate adjustments, definitions of the included covariates are as follows. Child characteristics at the time of the referral include the child's age at referral, race, and legal sex. Household characteristics include a risk score category (low, medium, high, mandatory) based on the household maximum of either the computer generated referral or the placement score, after cutoffs are applied (using the Pre-AFST algorithm to create risk score for Post-AFST children), the household composition, including total persons in the household at the time of the call in these categories: less than 1 year, 1 to 5 years, 6 to 12 years, 13 to 17 years, parents of victim, other adults; a binary variable indicating if the mean age of adults in the household is: 18 to 29 years, 30 to 49 years, 50 to 65 years, or 66+ years, or "no adult age is listed"; and a set of binary indicators for which of five poverty categories the household's zip code belongs to or if the household does not have data for zip code. When covariate information is not complete, variables indicating unknown or unable to determine are used as described above and in the Appendices. Because the outcome is the same for all children in a referral (household), and because we expect some correlation in outcomes among individual call screeners, we cluster the standard errors at the level of the call screener identification number for Outcomes 1–3. Because the consistency outcome and related disparity outcomes are at the are at the call screener level, we cluster standard errors at the referral level for these analyses.

Description of Trends in the Study Population and Changes to Case Mix over Time

The importance of evaluating changes in trends in outcomes is highlighted by trends in the study population over the periods prior to and after the implementation of the AFST (January 1, 2015 through May 31, 2016 [the Pre-AFST Period] and December 1, 2016 through May 31, 2018 [the Post-AFST Period]). In principle, outcomes like investigational workload or call screener accuracy could be related to overall volume of referrals. If there is a limit on the number of investigations that can be conducted in a month then for days/months with high levels of referrals this limit may force a lower screen-in rate. If there is a limit on the amount of time that call screeners have, then for days/months with high levels of referrals, this will tend to reduce the time the screeners can spend on each call which could decrease accuracy of triaging calls. In fact, the monthly volume of children involved in GPS calls increased between the Pre-AFST period (~1,600 children per month) to (~1,900 children per month) with bigger increases in call volumes involving older children than younger children and with bigger increases for Black/African American children than white children. Hence, trends and changes in trends are important to consider (**Appendix Figures 1a–1c**).

The importance of evaluating outcomes adjusting for case mix is highlighted by changes in the case mix of the study population over time. In principle, outcomes like accuracy could be related to case mix. For example, if the system is more accurate for some subgroups of children then increases in the prevalence of that group could lead to an estimated increase in the unadjusted outcome which might otherwise be attributed to AFST. We examine individual child characteristics in the Pre-AFST and Post-AFST periods (**Table 1a**) as well as household characteristics in these periods (**Table 1b**) for changes in case mix. In general, most individual child and household measures stay similar though there is an increase in the prevalence of Black/African American children in referrals.

RESULTS

Accuracy Outcomes and Disparities in Accuracy Outcomes

In this section, we describe how the implementation of the AFST and related policies changed accuracy in two ways. First, we focus on accuracy for children screening in as measured by the proportion of children with referrals which screened-in for investigation that had further action taken, or if not, had a re-referral within 2 months. Second, we focus on accuracy for children screened out as measured by the proportion of children with referrals which screened-out who had no re-referral calls within 2 months. We also examine differential impacts of the AFST on accuracy by age-specific subgroups and race-specific subgroups to assess disparities in accuracy outcomes.

How did the AFST change accuracy for referrals screened-in?

The AFST increased the accuracy for referrals that screened-in as measured by an increase in the percentage of children screened-in for investigation who had further action taken or, if not had a

re-referral within 2 months. While this improvement in accuracy of being screened-in remained higher throughout the Post-AFST period, the initial improvement effect of implementing the AFST did attenuate somewhat over time. With a re-referral window of 6 months, the direction of the result was the same and there was somewhat less attenuation over time. See **Tables 2, 3a, 4a, Figures 2a, 3a, Appendix Table A1a, Appendix Table A2a and Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-in related to children in different age-groups?

The AFST increased the accuracy for referrals that screened-in for children over age 4, as measured by an increase in the percentage of children in each age-group screened-in for investigation who had further action taken or, if not, had a re-referral within 2 months. While this age subgroup-specific improvement in accuracy of being screened-in remained higher throughout the Post-AFST period, the initial improvement effect of implementing the AFST in each subgroup did attenuate somewhat over time. With a re-referral window of 6 months, the direction of the result was the same and there was somewhat less attenuation over time. See **Tables 2, 3b-e, 4b, Figures 2b, 3b, Appendix Tables A1b-A1e, Appendix Table A2b and Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-in related to children in different race groups?

The AFST had an immediate upward effect on the accuracy for referrals that screened-in for both white and Black/African American children, as measured by an increase in the percentage of children in each race subgroup screened-in for investigation who had further action taken or, if not, had a re-referral within 2 months. For Black/African American children, the initial improvement effect of implementing the AFST attenuated over time, such that there was no significant overall increase in accuracy for Black/African American children when compared to White children. With a re-referral window of 6 months, there was a slightly larger increase in the accuracy of screening-in for Black/African American children, although the increase was not significantly different than zero. See **Tables 2, 3f-g, 4c, Figures 2c, 3c, Appendix Tables A1f-A1g, Appendix Table A2c and Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-out?

The AFST had little effect on accuracy for referrals that screened-out as measured by a decrease in the percentage of children screened-out who had a re-referral within 2 months. The multiple analyses showed small decreases in the accuracy of screening out but only sometimes found this decrease to be statistically significant. Prior to the implementation of the AFST, the accuracy for referrals screened-out was increasing slightly, a trend that the AFST largely halted. The results were similar when we used a re-referral window of 6 months. See **Tables 2, 5a, 6a, Figures 4a, 5a, Appendix Table A3a, Appendix Table A4a and Appendix Table A6** for numerical details.

How did the AFST change accuracy for referrals screened-out related to children in different age-groups?

In breaking down the AFST's effect on accuracy of being screened-out by age-group, the largest decrease in accuracy occurred in children aged 4 to 6 years. While there were small, non-significant reductions in the accuracy of being screened-out for all age-groups, the larger effect in 4 to 6 year-olds may be due to changes in the policy regarding the maximum age for mandatory field screening which was reduced from under 7 years if age to under 4 years of age in the Post-AFST period, where previously the field screening in this age-group helped to identify more children in this age-group for whom being screened-in for investigation was appropriate. The results were similar when we used a re-referral window of 6 months. See **Tables 2, 5b–e, 6b, Figures 4b, 5b, Appendix Table A3b–A3e, Appendix Table A4b and Appendix Table A6** for numerical details.

How did the AFST change accuracy for referrals screened-out related to children in different race groups?

The AFST had little effect on accuracy for referrals that screened-out for both white and Black/African American children, as measured by a decrease in the percentage of children screened-out who had a re-referral within 2 months. The multiple analyses showed small decreases for both race subgroups, which were not significant for white children and only occasionally significant for Black/African American children. The results were similar when we used a re-referral window of 6 months. See **Tables 2, 5f–g, 6c, Figures 4c, 5c, Appendix Table A3f–A3g, Appendix Table A4c and Appendix Table A6** for numerical details.

Workload Outcomes and Disparities in Workload Outcomes

In this section, we describe how the implementation of the AFST and related policies changed workload in terms of the fraction of children who screened in and hence had investigations conducted. We also examine differential impacts of the AFST on workload by age-specific subgroups and race-specific subgroups to assess disparities in workload outcomes.

How did the AFST change workload as measured by the fraction of referrals screened-in for further investigation?

Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining. The AFST largely halted this decline. Hence, even though the average level in the Pre-AFST period was higher than in the Post-AFST period, it may well be the case that had the AFST not been implemented, screen-in rates could have continued to decline and been lower than were observed in the Post-AFST period. See **Tables 2, 7a, 8a, Figures 6a, 7a and Appendix Table A7** for numerical details.

How did the AFST change workload related to children in different age-groups?

Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining for children in all subgroups except for age 4 to 6, with larger declines observed in the oldest age-group (13 to 17). The AFST largely halted these age-specific declines, most noticeably for children aged 7 years and older. Therefore, despite age subgroup-specific declines in levels from the Pre-AFST and Post-AFST periods, it may well have been the case that the subgroup-specific levels in the Post-AFST period could have been even lower without the implementation of the AFST. See **Tables 2, 7b–e, 8b, Figures 6b, 7b** and **Appendix Table A7** for numerical details.

How did the AFST change workload related to children in different race groups?

Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining for children in all race groups with larger declines observed in Black / African American children compared to those in white children. The AFST largely halted these race-specific declines, most noticeably for Black / African American children. Therefore, despite race subgroup-specific declines in levels from the Pre-AFST and Post-AFST periods, it may well have been the case that the subgroup-specific levels in the Post-AFST period could have been even lower without the implementation of the AFST. See **Tables 2, 7f–g, 8c, Figures 6c, 7c** and **Appendix Table A7** for numerical details.

Consistency Outcomes and Disparities in Consistency Outcomes

In this section, we examine the consistency of the AFST's effects on accuracy and workload across call screeners. We examine whether the magnitudes of AFST's effects differed across call screeners and specifically whether AFST decreased their variation in outcomes in the Post-AFST period relative to variation the Pre-AFST period. Finally, we examine whether the AFST's variation in outcomes and change in variation in outcomes differed for referrals involving children of age- or race-specific subgroups. Of note, this analysis is restricted to the 11 call screeners with at least 350 referrals in both the Pre-AFST and Post-AFST periods.¹¹ Hence, outcomes are not directly comparable to the outcomes for all call screeners reported to this point. See **Table 9**, which shows comparison of the unadjusted means for call screeners included and those excluded from this analysis. While most outcomes are quite similar between the two groups of call screeners, the Post-AFST value for Outcome 3 (fraction of referrals screened-in) are substantially higher for the group included in this analysis than they are for call screeners not included in this analysis, highlighting that this analysis is relevant only for examination of the consistency of Outcomes 1–3 (and not the outcomes themselves), and across only the part of the call screener workforce that has been stable over time.

¹¹ Further, all analyses exclude the post-AFST months on April and May 2018 (referrals made in these months were excluded for Outcomes 1 and 2, so we exclude them for Outcome 1 to maintain consistency among included call screeners).

How did the AFST change the consistency of accuracy for referrals screened-in by call screener?

The overall increase in accuracy of a screen-in was consistent across call screeners, and variation between calls screeners in this outcome did not change significantly. Accuracy of being screened-in increased in the Post-AFST period compared to the Pre-AFST period for 10 of the 11 call

screeners (statistically significantly so in 1 call screener) and decreased in 1 call screener (not statistically significantly). The variance of call screener-specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 10a, 11a, Figures 8a, and Appendix Figure 2a** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by age-groups?

The overall increase in accuracy of a screen-in for children of different age-groups was consistent across call screeners and was generally larger for children in older age-groups, and variation between call screeners in this outcome did not change significantly for children of any age-group. Most call screeners increased in accuracy for screen-ins for children in each age-group. Very few increases were statistically significant at the call screener/age-group level given the small sample sizes. The variance of call screener specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period for children in all age-groups. See **Tables 10b, 11b, Figures 8b, and Appendix Figure 2b** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by race groups?

The overall increase in accuracy of a screen-in for both white and Black/African American children was reasonably consistent across call screeners, though average increases were smaller for Black/African American children across call screeners. Variation between call screeners in this outcome did not change significantly for referrals in either race subgroup. Accuracy increased for 10 of the 11 call screeners for white children (statistically significantly so in 2 call screeners) and decreased for 1 call screener (not statistically significantly). Accuracy increased for 7 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and decreased in 4 of the 11 call screeners (not statistically significantly). The variance of call screener outcomes increased for both race groups (not statistically significantly). See **Tables 10c, 11c, Figure 8c, and Appendix Figure 2c** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener?

The overall decrease in accuracy of screen-outs was consistent across call screeners, and variation between call screeners in this outcome did not change significantly. Accuracy of being screened-out decreased in the Post-AFST period compared to the Pre-AFST period for 11 of the 11 call screeners (statistically significantly so in 1 call screener). The variance of call screener-specific outcomes increased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 12a, 13a, Figure 9a, and Appendix Figure 3a** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by age-groups?

The decrease in accuracy of screen-outs for children of different age-groups was generally more concentrated for children in younger age-groups, particularly for those age 4-6 years,

was consistent across call screeners, and variation between call screeners in this outcome did not change significantly for children of any age-group. While most call screeners had decreases in accuracy of screen-outs for younger children especially, very few of these decreases were statistically significant (and none of the increases were). Sample size and frequency of outcome at the call screener/age-group level meant that the analyses of this outcome were underpowered. The variance of call screener specific outcomes increased (statistically significant in 7 to 12 year olds) in the Post-AFST period compared to the Pre-AFST period for children in all age-groups less than 13 years and decreased minimally for the age-group 13 to 17 years (not statistically significantly). See **Tables 12b, 13b, Figure 9b, and Appendix Figure 3b** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by race groups?

The overall decrease in accuracy of screen-outs for both white and Black/African American children was reasonably consistent across call screeners, though average decreases were somewhat larger for Black/African American children across call screeners. Variation between call screeners in this outcome did not change significantly for referrals in either race subgroup. Accuracy decreased for 8 of the 11 call screeners for white children (statistically significantly so in 1 call screener) and increased for 3 call screeners (not statistically significantly). Accuracy decreased for 8 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and increased in 3 of the 11 call screeners (not statistically significantly). The variance of call screener outcomes increased for both race groups (not statistically significantly). See **Tables 12c, 13c, Figure 9c, and Appendix Figure 3c** for numerical details.

How did the AFST change workload differentially by call screener?

The overall finding described above on workload (fraction of referrals screening in for investigation) was that the average level decreased from the Pre-AFST period to the Post-AFST period, though because of strong declining trends in workload in the Pre-AFST period, the Post-AFST levels may have been higher than what they would have been without the implementation of the AFST. There was moderate consistency in the workload outcome across call screeners. Workload increased in the Post-AFST period for 7 of the 11 call screeners (4 of these were significant increases) and decreased for 4 of the 11 call screeners (none statistically significantly). The variance of call screener-specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 14a, 15a, Figure 10a, and Appendix Figure 4a** for numerical details.

How did the AFST change workload differentially by call screener by age-groups?

The by-screener increase in workload for children of different age-groups was generally more concentrated in the middle age-groups, was consistent across call screeners, and variation between call screeners in this outcome did not change significantly for children of any age-group. Very few of the increases in workload were statistically significant, and none of the decreases in workload were statistically significant. Sample size at the call screener/age-group

level meant that the analyses of this outcome were underpowered. The variance of call screener-specific outcomes decreased for the younger age-groups (0 to 6 years) and increased of the older age-groups (not statistically significantly) between the Pre-AFST and the Post-AFST period. See **Tables 14b, 15b, Figure 10b, and Appendix Figure 4b** for numerical details.

How did the AFST change workload differentially by call screener by race groups?

The overall change in workload for calls involving white and those involving Black/African American children was reasonably consistent across call screeners. Variation between calls screeners in this outcome did not change significantly for referrals in either race subgroup. Workload increased for 7 of the 11 call screeners for white children (statistically significantly so in 3 call screeners) and decreased for 3 call screeners (statistically significantly so in 1 call screener). Workload increased for 7 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and decreased for 4 call screeners (not statistically significantly). The variance of call screener outcomes decreased for both race groups, with a larger effect apparent in the Black/African American group (not statistically significantly). See **Tables 14c, 15c, Figure 10c, and Appendix Figure 4c** for numerical details.

DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

We evaluated the impact of the AFST screening score implementation within Allegheny County's child welfare office in terms of its effect on accuracy, workload, disparity and consistency outcomes for children involved in GPS referrals. Overall, our analyses showed that the AFST and associated policies increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out. Improvements in accuracy attenuated somewhat over time post-implementation. The AFST and associated policies also stopped the downward trend in the rate of children screened-in for investigation. Age- and race-specific subgroup analyses showed that screen-in accuracy improvements were largest and/or had less attenuation over time for white children and children aged < 4 years. Loss of accuracy in screening-out was concentrated most in children ages 4 to 6 years though the overall size of effect even in this age-group was relatively small. This may be due to concurrent changes in the mandatory in-home assessment (field screening) policy in terms of the maximum age being reduced from under 7 to under 4 years of age. Effects were generally consistent across call screeners.

As with all such evaluations, methodological choices and assumptions were required. Below we discuss a number of these and the considerations that justified them as well as their potential limitations.

Accuracy-related outcomes for both screen-ins and screen-outs are defined partly based on whether subsequent referral calls are made within a given time window. In principle, additional referral calls could come shortly after the index call regarding the same incident which would potentially influence the accuracy measure. In our analyses this did not turn out to be the case. The number of index calls for which additional calls occurred within 1 day is less than 1%, within

2 days is less than 2% and within 1 week less than 3%. When we made a robustness check to exclude these calls, our main results did not change in any substantial way.

Analyses examining consistency across call screeners used a cut-off of having at least 350 calls in the pre- and post-AFST periods. While this ensures that estimates of outcomes and effects made at the call screener level tend to be more stable, the cut-off may appear arbitrary and certainly excludes call screeners with fewer calls. In fact, the distribution of calls taken by call screeners is bi-modal with a group of call screeners taking well below 200 calls in total and another taking well over 500. Hence a range of cut-off values would yield the same set of call screeners for analyses.

We analyzed outcomes stratified into age-specific groups, including a young age group of children aged 4 years and under. However, even within this age group, there could be additional heterogeneity, especially at the younger end since infants below the age of 1 year are of concern to agencies given their inherent vulnerability. While we include household composition as control variables in our multivariate analyses – specifically the count of children in the household age 1 year or below – we did not stratify our outcomes by this finer age category as the number of children in this finer age category is insufficient to provide precise effect estimates. Future analyses with more months of follow-up are planned to examine outcomes in finer age groups.

One of the outcomes we examined was the effect of the implementation of the AFST and surrounding policy changes on disparities in outcomes across race/ethnic and age-specific subgroups. It is important to note that true underlying rates of neglect and maltreatment for each of these subgroups is unknown and hence increases/decreases in a given measured system outcome (e.g., screen-ins) of one subgroup relative to another in principle could represent either a widening or a narrowing of a disparity (e.g., in terms of children experiencing actual neglect or maltreatment having the referral investigated). Given that the key assumption of the analysis is that changes in underlying conditions like rates of neglect and maltreatment are substantially slower than the change of implementing the AFST and surrounding policies, examining how outcomes changed from the pre- to the post-implementation period within groups are illustrative for exploring whether the use of the tool within the system led to bigger changes for some groups relative to others. Proper interpretation of such results critically depends on the ability of AFST to detect actual neglect or maltreatment in each group and for workers to act accordingly.

The goal of the evaluation of the effects of the AFST and surrounding policies was to provide a set of measures that are meaningful and important. However, the evaluation makes no claim or judgement about the relative importance of one outcome related to another. Specifically, we analyzed multiple outcomes including accuracy measures for screening in and screening out, workload, consistency across call screeners, and differences in these outcomes by age and race/ethnic subgroups. The stated goal of the AFST implementation is primarily that of improving accuracy. If achieving increased accuracy also involved increases in the number of calls screening-in for investigation, this would not necessarily imply that the AFST implementation was unsuccessful.

Rather, it might imply that additional allocation of investigative resources is required to sustain improvements due to the AFST—a finding that is relevant for other systems considering implementing similar tools. We encourage interpretation of findings across outcomes in a holistic way and with reference to the stated goals and constraints of child-serving systems.

The assessment of effects of the implementation in the evaluation relies on quasi-experimental methods as direct randomization was not feasible. One strength of the evaluation is that it uses multiple quasi-experimental methods – interrupted time series analysis (ITSA) as well as multivariate regression analyses indicators for the timing of the implementation—with findings quite consistent across these methods. As noted in the methods, a key assumption is that the estimated effect (changes pre to post-implementation) can be attributed primarily to the implementation of AFST because other changes (e.g., changes in case mix over time) are much slower and less abrupt than the implementation itself. The multivariate regression adjusts for many features of case mix explicitly. Yet, for both methods, if unmeasured features change relatively abruptly, it is in principle possible that the estimated effect is not attributable entirely to the AFST implementation.

To provide some context in terms of how many children may be affected by the AFST, estimates of child-counts for Outcomes 1–3 are presented in **Tables 16a–16c**. These estimates are based on the predicted probabilities of a given outcome estimated in the adjusted analyses, the related confidence intervals and the mean monthly total counts of children in referrals, children who screen-in and children who screen-out (over both the Pre- and Post- AFST). Roughly 24 more children each month screen-in accurately after the AFST, with over half of these children in the 7 to 12 year old age range and almost all of these children in the white race group. Roughly 11 more children who screen-out are done so inaccurately each month (though this result is not statistically significant) with $\sim 2/3$ of these children falling into the Black/African American race group (although the results are not statistically significant for any breakdown of age or race). Roughly 53 fewer children included in referrals screen-in each month (not significant) with over half of these falling in the 13 to 17-year age range and $\sim 2/3$ of these children in the black race group.

In conclusion, our evaluation of the effects of implementing the AFST and surrounding policy changes shows moderate improvements in accuracy of screen-ins with small decreases in the accuracy in screen-outs, a halt in the downward trend in pre-implementation screen-ins for investigation, no large or consistent differences across race/ethnic or age-specific subgroups in these outcomes, and no large or substantial differences in consistency across call screeners. As with the initial phases of most large-scale real-world system changes, implementation challenges arose, and one can speculate as to whether the achievable effects without such challenges could have been larger. In sum, the AFST appears to have had a modest positive effect on some screening outcomes that can be determined via process measures. Ultimately, Allegheny County and other systems considering the use tools like the AFST will need to consider how such metrics relate to their core goals (e.g., safety) and how achieving these effects relate to their costs and resource constraints both in terms of implementing the tool and the downstream impacts that such a tool can have.

TABLES

TABLE 1A: Summary Statistics, child characteristics

	PRE-AFST (JANUARY 1, 2015 - JULY 31, 2016)			POST-AFST (DECEMBER 1, 20165 - MAY 31, 2018)			P-VALUE*
	MEAN	95% CI		MEAN	95% CI		
Legal sex							
Male	50.72%	50.17%	51.28%	50.36%	49.82%	50.89%	0.352
Female	48.98%	48.43%	49.54%	48.76%	48.23%	49.30%	0.574
Other	0.29%	0.23%	0.36%	0.88%	0.78%	0.98%	0.000
Race							
Black/African American	46.70%	46.15%	47.26%	50.99%	50.45%	51.52%	0.000
White	41.03%	40.48%	41.58%	41.97%	41.44%	42.49%	0.015
Other	12.27%	11.90%	12.63%	7.05%	6.78%	7.32%	0.000
Age-group							
< 4 years	22.88%	22.42%	23.35%	21.95%	21.51%	22.39%	0.004
4–6 years	17.83%	17.40%	18.25%	16.64%	16.24%	17.04%	0.000
7–12 years	34.60%	34.07%	35.13%	36.07%	35.56%	36.58%	0.000
13–17 years	24.69%	24.21%	25.17%	25.34%	24.88%	25.80%	0.055

Sample sizes are 31,190 (Pre-AFST) and 33,966 (Post-AFST). The child is considered "Black or African American" if their race is coded as "Black or African American" or "Black or African American" mixed with another race. *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.

TABLE 1B: Summary Statistics, household characteristics

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)			POST-AFST (DECEMBER 1, 2016–MAY 31, 2018)			P-VALUE*
	MEAN	STANDARD ERROR	95% CI	MEAN	STANDARD ERROR	95% CI	
Risk score category							
Mandatory	22.60%		22.14% 23.07%	24.84%		24.38% 25.30%	0.000
High	34.06%		33.53% 34.58%	35.55%		35.04% 36.06%	0.000
Medium	24.14%		23.67% 24.62%	22.45%		22.01% 22.90%	0.000
Low	18.41%		17.98% 18.84%	16.92%		16.52% 17.32%	0.000
No score	0.79%		0.69% 0.88%	0.24%		0.18% 0.29%	0.000
Household poverty category (zip code)							
Wealthiest	24.89%		24.41% 25.37%	25.52%		25.06% 25.99%	0.063
Wealthier	20.90%		20.45% 21.36%	19.73%		19.30% 20.15%	0.000
Middle	10.31%		9.97% 10.65%	9.89%		9.57% 10.21%	0.074
Poor	25.60%		25.11% 26.08%	24.30%		23.84% 24.75%	0.000
Poorest	15.33%		14.93% 15.73%	15.44%		15.05% 15.82%	0.705
No zip code information	2.97%		2.78% 3.15%	5.13%		4.89% 5.36%	0.000
Mean age of household adults							
18 - 29 years	21.69%		21.23% 22.15%	20.09%		19.66% 20.52%	0.000
30 - 49 years	69.40%		68.89% 69.91%	71.14%		70.65% 71.62%	0.000
50 - 65 years	4.98%		4.74% 5.22%	5.47%		5.23% 5.72%	0.005
66 years–max	0.38%		0.31% 0.45%	0.28%		0.23% 0.34%	0.028
No adult age information	3.55%		3.35% 3.76%	3.02%		2.84% 3.20%	0.000
Household composition (counts)							
# parents	1.327	0.006		1.282	0.006		0.000
# other adults	1.488	0.005		1.589	0.005		0.000
# age 13–17	0.697	0.005		0.726	0.005		0.000
# age 6–12	1.278	0.006		1.301	0.006		0.007
# age–5	0.805	0.005		0.780	0.005		0.001
# age < 1	0.176	0.002		0.177	0.002		0.849

All means are for entire sample of all referred children. Sample sizes are 31,190 (Pre-AFST) and 33,966 (Post-AFST). Risk scores categories are based on the maximum risk score within a given referral (household) of either the referral or the placement risk score. Risk bins were calculated using raw risk scores, and bin cutoffs were provided by Allegheny. Individual households have their zip codes categorized into poverty categories based on the American Community Survey (2008–2012) and its determination of the percentage of all households living below the poverty line, as follows: Poorest (>= 25%); Poor (20% to <25%); Mid (15% to <20%); Wealthier (10% to <15%); and Wealthiest (0% to <10%). *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.

TABLE 2: Means of outcomes

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)			POST-AFST (DECEMBER 1, 2016–MAY 31, 2018)			P-VALUE*		
	MEAN	N	95% CI	MEAN	N	95% CI			
Outcome (1) Accuracy of screen-in: Screen-in with further action taken or re-referral within 60 days									
All children	42.85%	15,016	42.06%	43.64%	46.61%	14,599	45.80%	47.42%	0.000
< 4 years	44.01%	3,947	42.46%	45.56%	45.18%	3,805	43.60%	46.76%	0.301
4 to 6 years	42.68%	2,570	40.77%	44.60%	45.97%	2,482	44.01%	47.93%	0.019
7 to 12 years	40.96%	4,997	39.60%	42.33%	45.93%	5,034	44.55%	47.30%	0.000
13 to 17years	44.37%	3,502	42.73%	46.02%	49.82%	3,278	48.10%	51.53%	0.000
White	39.26%	5,589	37.98%	40.54%	46.35%	5,685	45.05%	47.65%	0.000
Black/ African American	47.28%	7,715	46.17%	48.40%	47.47%	8,091	46.38%	48.56%	0.813
Outcome (2) Accuracy of screen-out: Screen-out with no re-referral within 60 days									
All children	85.02%	14,676	84.45%	85.60%	84.25%	16,433	83.69%	84.80%	0.094
< 4 years	85.49%	2,861	84.20%	86.79%	84.89%	2,979	83.61%	86.18%	0.519
4 to 6 years	85.13%	2,696	83.78%	86.47%	83.71%	2,683	82.31%	85.11%	0.153
7 to 12 years	84.78%	5,269	83.81%	85.75%	84.11%	6,143	83.20%	85.03%	0.327
13 to 17years	84.73%	3,850	83.59%	85.86%	84.45%	4,629	83.40%	85.49%	0.721
White	84.05%	6,488	83.16%	84.94%	83.79%	7,247	82.94%	84.64%	0.678
Black/ African American	84.42%	6,221	83.52%	85.33%	82.99%	7,726	82.15%	83.83%	0.023
Outcome (3) Workload: Screen-in									
All children	48.23%	31,176	47.67%	48.78%	46.19%	33,524	45.65%	46.72%	0.000
< 4 years	55.39%	7,133	54.24%	56.54%	55.22%	7,296	54.08%	56.36%	0.839
4 to 6 years	46.32%	5,559	45.01%	47.63%	47.07%	5,573	45.76%	48.38%	0.431
7 to 12 years	46.38%	10,789	45.44%	47.32%	44.23%	12,119	43.34%	45.11%	0.001
13 to 17years	45.55%	7,695	44.44%	46.66%	40.67%	8,536	39.63%	41.72%	0.000
White	43.69%	12,794	42.83%	44.55%	42.92%	14,067	42.10%	43.73%	0.200
Black/ African American	53.11%	14,559	52.30%	53.93%	50.28%	17,082	49.53%	51.03%	0.000

Because outcomes are not often finalized on the referral date, we censor the Post-AFST period call-outcome variable at May 31, 2018. To allow complete follow-up for the second and third outcomes (re-referral within 60 days), we only included referrals through March 31, 2018 for the Post-AFST so that April and May data could be used to verify that re-referrals had or had not occurred. Screen-ins (the first outcome) include all children (< 18 years) in all GPS referrals. For the second outcome (screen-out: no re-referrals), any referral call within the 60-day window after the index referral was considered to determine whether a re-referral had occurred. Subsequent referrals outside the window were considered new "index events" for this analysis. The third outcome (screen-in: further action) includes all children who were screened-in at index referral and had a processed "service decision". *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.

TABLE 3A: Accuracy of screen-in, ITSA analysis, all children

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	54.05%	0.000	51.14%	56.96%
Pre-2014 Policy	Trend	-0.66	0.000	-1.01	-0.32
2014 Policy	Change in level	2.42	0.413	-3.48	8.33
Post 2014 policy, Pre-AFST	Change in trend	0.47	0.025	0.06	0.88
AFST implementation	Change in level	10.19	0.000	7.19	13.19
Post-AFST	Change in trend	-0.28	0.149	-0.66	0.10
Total trend in screen-in/further action rates Pre-AFST		-0.19	0.161	-0.47	0.08
Total trend in screen-in/further action rates Post-AFST		-0.47	0.001	-0.74	-0.21

Note: change in trend is expressed in percentage points/month.

TABLE 3B: Accuracy of screen-in, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	55.88%	0.000	50.85%	60.91%
Pre-2014 Policy	Trend	-0.54	0.015	-0.97	-0.11
2014 Policy	Change in level	-2.01	0.504	-8.00	3.99
Post 2014 policy, Pre-AFST	Change in trend	0.52	0.076	-0.06	1.10
AFST implementation	Change in level	4.85	0.130	-1.48	11.19
Post-AFST	Change in trend	-0.35	0.311	-1.05	0.34
Total trend in screen-in/further action rates Pre-AFST		-0.02	0.917	-0.41	0.37
Total trend in screen-in/further action rates Post-AFST		-0.37	0.198	-0.95	0.20

Note: change in trend is expressed in percentage points/month.

TABLE 3C: Accuracy of screen-in, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	53.37%	0.000	49.37%	57.36%
Pre-2014 Policy	Trend	-0.80	0.000	-1.20	-0.39
2014 Policy	Change in level	6.02	0.068	-0.47	12.50
Post 2014 policy, Pre-AFST	Change in trend	0.55	0.043	0.02	1.07
AFST implementation	Change in level	9.16	0.000	4.83	13.48
Post-AFST	Change in trend	-0.08	0.739	-0.53	0.38
Total trend in screen-in/further action rates Pre-AFST		-0.25	0.143	-0.59	0.09
Total trend in screen-in/further action rates Post-AFST		-0.33	0.034	-0.63	-0.03

Note: change in trend is expressed in percentage points/month.

TABLE 3D: Accuracy of screen-in, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	50.69%	0.000	45.91%	55.47%
Pre-2014 Policy	Trend	-0.60	0.054	-1.21	0.01
2014 Policy	Change in level	3.74	0.359	-4.39	11.87
Post 2014 policy, Pre-AFST	Change in trend	0.30	0.362	-0.36	0.96
AFST implementation	Change in level	13.43	0.000	9.46	17.41
Post-AFST	Change in trend	-0.28	0.302	-0.83	0.26
Total trend in screen-in/further action rates Pre-AFST		-0.30	0.021	-0.55	-0.05
Total trend in screen-in/further action rates Post-AFST		-0.58	0.021	-1.07	-0.09

Note: change in trend is expressed in percentage points/month.

TABLE 3E: Accuracy of screen-in, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	56.94%	0.000	51.95%	61.94%
Pre-2014 Policy	Trend	-0.75	0.000	-1.12	-0.37
2014 Policy	Change in level	2.02	0.445	-3.25	7.28
Post 2014 policy, Pre-AFST	Change in trend	0.62	0.032	0.05	1.18
AFST implementation	Change in level	11.81	0.000	5.97	17.66
Post-AFST	Change in trend	-0.42	0.128	-0.96	0.12
Total trend in screen-in/further action rates Pre-AFST		-0.13	0.530	-0.55	0.29
Total trend in screen-in/further action rates Post-AFST		-0.55	0.003	-0.90	-0.21

Note: change in trend is expressed in percentage points/month.

TABLE 3F: Accuracy of screen-in, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	47.22%	0.000	40.98%	53.46%
Pre-2014 Policy	Trend	-0.57	0.063	-1.17	0.03
2014 Policy	Change in level	2.18	0.556	-5.22	9.58
Post 2014 policy, Pre-AFST	Change in trend	0.57	0.123	-0.16	1.31
AFST implementation	Change in level	10.02	0.005	3.13	16.91
Post-AFST	Change in trend	-0.26	0.419	-0.90	0.38
Total trend in screen-in/further action rates Pre-AFST		0.01	0.972	-0.42	0.44
Total trend in screen-in/further action rates Post-AFST		-0.25	0.292	-0.73	0.22

Note: change in trend is expressed in percentage points/month.

TABLE 3G: Accuracy of screen-in, ITSA analysis, Black/African American

		STARTING RATE (% OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	58.87%	0.000	52.83%	64.91%
Pre-2014 Policy	Trend	-0.67	0.026	-1.26	-0.08
2014 Policy	Change in level	2.82	0.483	-5.20	10.83
Post 2014 policy, Pre-AFST	Change in trend	0.41	0.260	-0.31	1.13
AFST implementation	Change in level	8.05	0.001	3.28	12.83
Post-AFST	Change in trend	-0.32	0.314	-0.94	0.31
Total trend in screen-in/further action rates Pre-AFST		-0.26	0.207	-0.68	0.15
Total trend in screen-in/further action rates Post-AFST		-0.58	0.016	-1.05	-0.11

Note: change in trend is expressed in percentage points/month.

TABLE 4A: Accuracy of screen-in, adjusted analysis, all children

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	43.68%	0.000	42.20%	45.15%
Post-AFST	46.56%	0.000	45.04%	48.08%
DIFF (Post - Pre)	2.88%	0.003	0.95%	4.81%

TABLE 4B: Accuracy of screen-in, adjusted analysis, by age group

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	43.43%	0.000	41.42%	45.44%
4-6 years	43.80%	0.000	41.51%	46.10%
7-12 years	43.82%	0.000	42.00%	45.64%
13-17 years	43.66%	0.000	41.76%	45.57%
Post-AFST				
< 4 years	43.31%	0.000	41.39%	45.22%
4-6 years	46.80%	0.000	44.46%	49.13%
7-12 years	48.46%	0.000	46.69%	50.23%
13-17 years	47.34%	0.000	45.39%	49.29%
Difference Post-Pre				
< 4 years	-0.12%	0.906	-2.19%	1.94%
4-6 years	2.99%	0.088	-0.44%	6.43%
7-12 years	4.64%	0.000	2.12%	7.16%
13-17 years	3.67%	0.019	0.60%	6.75%

TABLE 4C: Accuracy of screen-in, adjusted analysis, by race

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION*	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	43.01%	0.000	40.90%	45.13%
Black/African American	45.51%	0.000	43.62%	47.39%
Post-AFST				
White	49.49%	0.000	47.32%	51.65%
Black/African American	45.44%	0.000	43.55%	47.33%
Difference Post-Pre				
White	6.47%	0.000	3.62%	9.32%
Black/African American	-0.07%	0.961	-2.73%	2.60%
Difference Black-White				
Pre-AFST	2.49%	0.088	-0.37%	5.36%
Post-AFST	-4.05%	0.006	-6.95%	-1.14%

TABLE 5A: Accuracy of screen-out, ITSA analysis, all children

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	84.85%	0.000	82.19%	87.51%
Pre-2014 Policy	Trend	-0.07	0.596	-0.32	0.18
2014 Policy	Change in level	0.20	0.876	-2.34	2.73
Post 2014 policy, Pre-AFST	Change in trend	0.20	0.155	-0.08	0.47
AFST implementation	Change in level	-2.45	0.065	-5.06	0.16
Post-AFST	Change in trend	-0.18	0.170	-0.45	0.08
Total trend in screen-out/no re-referral rates Pre-AFST		0.13	0.016	0.02	0.23
Total trend in screen-out/no re-referral rates Post-AFST		-0.05	0.658	-0.29	0.19

Note: change in trend is expressed in percentage points/month.

TABLE 5B: Accuracy of screen-out, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	86.29%	0.000	83.87%	88.71%
Pre-2014 Policy	Trend	0.00	0.981	-0.23	0.23
2014 Policy	Change in level	-1.87	0.263	-5.18	1.45
Post 2014 policy, Pre-AFST	Change in trend	0.12	0.511	-0.25	0.50
AFST implementation	Change in level	-3.85	0.143	-9.06	1.35
Post-AFST	Change in trend	0.01	0.983	-0.52	0.54
Total trend in screen-out/no re-referral rates Pre-AFST		0.12	0.417	-0.18	0.42
Total trend in screen-out/no re-referral rates Post-AFST		0.13	0.562	-0.31	0.57

Note: change in trend is expressed in percentage points/month.

TABLE 5C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	83.97%	0.000	78.79%	89.15%
Pre-2014 Policy	Trend	0.07	0.763	-0.39	0.53
2014 Policy	Change in level	-1.95	0.398	-6.53	2.64
Post 2014 policy, Pre-AFST	Change in trend	0.16	0.509	-0.32	0.64
AFST implementation	Change in level	-4.74	0.007	-8.11	-1.37
Post-AFST	Change in trend	-0.22	0.148	-0.52	0.08
Total trend in screen-out/no re-referral rates Pre-AFST		0.23	0.005	0.07	0.39
Total trend in screen-out/no re-referral rates Post-AFST		0.01	0.952	-0.25	0.27

Note: change in trend is expressed in percentage points/month.

TABLE 5D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	82.47%	0.000	78.42%	86.52%
Pre-2014 Policy	Trend	0.02	0.922	-0.38	0.42
2014 Policy	Change in level	0.70	0.730	-3.33	4.72
Post 2014 policy, Pre-AFST	Change in trend	0.13	0.518	-0.27	0.54
AFST implementation	Change in level	-1.91	0.154	-4.56	0.74
Post-AFST	Change in trend	-0.32	0.076	-0.67	0.03
Total trend in screen-out/no re-referral rates Pre-AFST		0.15	0.001	0.07	0.23
Total trend in screen-out/no re-referral rates Post-AFST		-0.16	0.336	-0.50	0.18

Note: change in trend is expressed in percentage points/month.

TABLE 5E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	87.39%	0.000	83.98%	90.80%
Pre-2014 Policy	Trend	-0.35	0.123	-0.79	0.10
2014 Policy	Change in level	3.26	0.283	-2.78	9.31
Post 2014 policy, Pre-AFST	Change in trend	0.38	0.151	-0.14	0.90
AFST implementation	Change in level	-0.38	0.855	-4.54	3.79
Post-AFST	Change in trend	-0.14	0.507	-0.55	0.28
Total trend in screen-out/no re-referral rates Pre-AFST		0.03	0.814	-0.24	0.30
Total trend in screen-out/no re-referral rates Post-AFST		-0.11	0.499	-0.42	0.21

Note: change in trend is expressed in percentage points/month.

TABLE 5F: Accuracy of screen-out, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	84.11%	0.000	80.88%	87.33%
Pre-2014 Policy	Trend	0.02	0.909	-0.37	0.42
2014 Policy	Change in level	-3.04	0.261	-8.41	2.34
Post 2014 policy, Pre-AFST	Change in trend	0.28	0.231	-0.18	0.73
AFST implementation	Change in level	-1.85	0.238	-4.97	1.27
Post-AFST	Change in trend	-0.62	0.004	-1.04	-0.21
Total trend in screen-out/no re-referral rates Pre-AFST		0.30	0.012	0.07	0.53
Total trend in screen-out/no re-referral rates Post-AFST		-0.33	0.061	-0.67	0.02

Note: change in trend is expressed in percentage points/month.

TABLE 5G: Accuracy of screen-out, ITSA analysis, Black/African American

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	84.69%	0.000	80.70%	88.67%
Pre-2014 Policy	Trend	-0.21	0.242	-0.57	0.15
2014 Policy	Change in level	3.70	0.121	-1.02	8.41
Post 2014 policy, Pre-AFST	Change in trend	0.20	0.384	-0.25	0.65
AFST implementation	Change in level	-3.20	0.136	-7.46	1.05
Post-AFST	Change in trend	0.18	0.380	-0.23	0.60
Total trend in screen-out/no re-referral rates Pre-AFST		-0.02	0.909	-0.29	0.26
Total trend in screen-out/no re-referral rates Post-AFST		0.17	0.293	-0.15	0.49

Note: change in trend is expressed in percentage points/month.

TABLE 6A: Accuracy of screen-out, adjusted analysis, all children

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	84.69%	0.000	84.01%	85.36%
Post-AFST	83.51%	0.000	82.46%	84.57%
DIFF (Post-Pre)	-1.17%	0.073	-2.46%	0.11%

TABLE 6B: Accuracy of screen-out, adjusted analysis, by age group

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	85.44%	0.000	83.82%	87.06%
4-6 years	84.97%	0.000	83.63%	86.31%
7-12 years	84.80%	0.000	84.02%	85.57%
13-17 years	83.77%	0.000	82.63%	84.92%
Post-AFST				
< 4 years	84.14%	0.000	82.17%	86.11%
4-6 years	82.66%	0.000	80.63%	84.69%
7-12 years	83.52%	0.000	82.03%	85.01%
13-17 years	83.56%	0.000	82.19%	84.93%
Difference Post-Pre				
< 4 years	-1.30%	0.344	-4.00%	1.40%
4-6 years	-2.31%	0.075	-4.86%	0.24%
7-12 years	-1.28%	0.192	-3.21%	0.64%
13-17 years	-0.21%	0.827	-2.11%	1.68%

TABLE 6C: Accuracy of screen-out, adjusted analysis, by race

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	82.65%	0.000	81.50%	83.79%
Black/African American	85.30%	0.000	84.13%	86.47%
Post-AFST				
White	81.63%	0.000	79.84%	83.42%
Black/African American	83.43%	0.000	81.83%	85.03%
Difference Post-Pre				
White	-1.01%	0.322	-3.02%	0.99%
Black/African American	-1.87%	0.067	-3.87%	0.13%
Difference Black-White				
Pre-AFST	2.65%	0.004	0.86%	4.44%
Post-AFST	1.79%	0.173	-0.78%	4.37%

TABLE 7A: Workload, ITSA analysis, all children

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	48.93%	0.000	45.57%	52.28%
Pre-2014 Policy	Trend	0.13	0.415	-0.19	0.46
2014 Policy	Change in level	0.80	0.720	-3.68	5.28
Post 2014 policy, Pre-AFST	Change in trend	-0.58	0.006	-0.99	-0.18
AFST implementation	Change in level	2.16	0.188	-1.09	5.40
Post-AFST	Change in trend	0.44	0.004	0.15	0.73
Total trend in screen-in rates Pre-AFST		-0.45	0.000	-0.69	-0.21
Total trend in screen-in rates Post-AFST		-0.01	0.914	-0.17	0.16

Note: change in trend is expressed in percentage points/month.

TABLE 7B: Workload, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	54.25%	0.000	51.35%	57.16%
Pre-2014 Policy	Trend	-0.03	0.852	-0.34	0.28
2014 Policy	Change in level	4.42	0.053	-0.06	8.91
Post 2014 policy, Pre-AFST	Change in trend	-0.29	0.162	-0.70	0.12
AFST implementation	Change in level	4.07	0.097	-0.76	8.90
Post-AFST	Change in trend	0.08	0.675	-0.30	0.46
Total trend in screen-in rates Pre-AFST		-0.32	0.021	-0.59	-0.05
Total trend in screen-in rates Post-AFST		-0.24	0.087	-0.51	0.04

Note: change in trend is expressed in percentage points/month.

TABLE 7C: Workload, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	46.02%	0.000	41.73%	50.32%
Pre-2014 Policy	Trend	0.16	0.442	-0.26	0.59
2014 Policy	Change in level	-0.59	0.868	-7.71	6.53
Post 2014 policy, Pre-AFST	Change in trend	-0.42	0.182	-1.05	0.21
AFST implementation	Change in level	3.03	0.225	-1.92	7.98
Post-AFST	Change in trend	0.26	0.330	-0.27	0.79
Total trend in screen-in rates Pre-AFST		-0.26	0.266	-0.72	0.20
Total trend in screen-in rates Post-AFST		0.00	0.995	-0.26	0.26

Note: change in trend is expressed in percentage points/month.

TABLE 7D: Workload, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	47.30%	0.000	43.35%	51.25%
Pre-2014 Policy	Trend	0.09	0.676	-0.32	0.50
2014 Policy	Change in level	0.25	0.922	-4.82	5.32
Post 2014 policy, Pre-AFST	Change in trend	-0.39	0.109	-0.87	0.09
AFST implementation	Change in level	-0.38	0.855	-4.57	3.80
Post-AFST	Change in trend	0.42	0.022	0.06	0.78
Total trend in screen-in rates Pre-AFST		-0.31	0.019	-0.56	-0.05
Total trend in screen-in rates Post-AFST		0.12	0.364	-0.14	0.37

Note: change in trend is expressed in percentage points/month.

TABLE 7E: Workload, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	48.74%	0.000	42.85%	54.64%
Pre-2014 Policy	Trend	0.33	0.189	-0.17	0.84
2014 Policy	Change in level	-1.20	0.689	-7.19	4.79
Post 2014 policy, Pre-AFST	Change in trend	-1.24	0.000	-1.81	-0.67
AFST implementation	Change in level	3.60	0.089	-0.57	7.78
Post-AFST	Change in trend	0.97	0.000	0.60	1.35
Total trend in screen-in rates Pre-AFST		-0.91	0.000	-1.17	-0.64
Total trend in screen-in rates Post-AFST		0.06	0.643	-0.21	0.33

Note: change in trend is expressed in percentage points/month.

TABLE 7F: Workload, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	44.22%	0.000	40.66%	47.77%
Pre-2014 Policy	Trend	0.04	0.834	-0.33	0.41
2014 Policy	Change in level	0.70	0.794	-4.63	6.03
Post 2014 policy, Pre-AFST	Change in trend	-0.26	0.262	-0.73	0.20
AFST implementation	Change in level	1.24	0.559	-2.99	5.47
Post-AFST	Change in trend	0.20	0.309	-0.19	0.58
Total trend in screen-in rates Pre-AFST		-0.22	0.112	-0.50	0.05
Total trend in screen-in rates Post-AFST		-0.03	0.844	-0.29	0.24

Note: change in trend is expressed in percentage points/month.

TABLE 7G: Workload, ITSA analysis, Black/African American

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	54.83%	0.000	51.47%	58.18%
Pre-2014 Policy	Trend	0.20	0.169	-0.09	0.49
2014 Policy	Change in level	0.50	0.818	-3.88	4.89
Post 2014 policy, Pre-AFST	Change in trend	-0.87	0.000	-1.27	-0.48
AFST implementation	Change in level	3.40	0.078	-0.40	7.20
Post-AFST	Change in trend	0.68	0.000	0.34	1.02
Total trend in screen-in rates Pre-AFST		-0.67	0.000	-0.94	-0.40
Total trend in screen-in rates Post-AFST		0.01	0.940	-0.20	0.21

Note: change in trend is expressed in percentage points/month.

TABLE 8A: Workload, adjusted analysis, all children

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	48.75%	0.000	46.84%	50.66%
Post-AFST	45.70%	0.000	42.67%	48.73%
DIFF (Post-Pre)	-3.05%	0.017	-6.47%	0.36%

TABLE 8B: Workload, adjusted analysis, by age-group

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	50.16%	0.000	47.75%	52.58%
4-6 years	47.84%	0.000	45.62%	50.05%
7-12 years	49.44%	0.000	47.47%	51.40%
13-17 years	47.25%	0.000	45.12%	49.38%
Post-AFST				
< 4 years	48.46%	0.000	46.09%	50.83%
4-6 years	47.65%	0.000	44.35%	50.94%
7-12 years	46.39%	0.000	42.92%	49.86%
13-17 years	40.99%	0.000	37.68%	44.30%
Difference Post-Pre				
< 4 years	-1.71%	0.233	-4.51%	1.09%
4-6 years	-0.19%	0.913	-3.62%	3.24%
7-12 years	-3.05%	0.139	-7.09%	0.99%
13-17 years	-6.26%	0.003	-10.44%	-2.08%

TABLE 8C: Workload, adjusted analysis, by race

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	48.64%	0.000	46.31%	50.97%
Black/African American	49.98%	0.000	47.59%	52.36%
Post-AFST				
White	46.82%	0.000	43.84%	49.80%
Black/African American	46.03%	0.000	42.84%	49.22%
Difference Post-Pre				
White	-1.82%	0.349	-5.63%	1.99%
Black/African American	-3.95%	0.040	-7.72%	-0.18%
Difference Black-White				
Pre-AFST	1.34%	0.322	-1.31%	3.98%
Post-AFST	-0.79%	0.226	-2.08%	0.49%

TABLE 9: Means of Outcomes (1)–(3) for call screeners included and excluded from Outcome 4/Consistency analyses

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)			POST-AFST (DECEMBER 1, 2016–MAY 31, 2018)			P-VALUE*		
	MEAN	N	95% CI	MEAN	N	95% CI			
Outcome (1) Accuracy of screen-in (Screen-in with further action taken or re-referral within 60 days)									
Screeners excluded	45.12%	4,978	43.74%	46.50%	45.77%	4,352	44.29%	47.25%	0.527
Screeners included	41.73%	10,038	40.77%	42.70%	46.97%	10,247	46.00%	47.94%	0.000
Outcome (2) Accuracy of screen-out (Screen-out with no re-referral within 60 days)									
Screeners excluded	84.43%	4,862	83.41%	85.45%	86.36%	7,641	85.59%	87.13%	0.003
Screeners included	85.23%	9,808	84.52%	85.93%	82.47%	8,787	81.68%	83.27%	0.000
Outcome (3) Workload (Screen-in)									
Screeners excluded	48.25%	10,328	47.28%	49.21%	36.05%	13,117	35.23%	36.87%	0.000
Screeners included	48.22%	20,848	47.54%	48.89%	52.70%	20,407	52.02%	53.39%	0.000

Screeners excluded are those with less than 350 referral calls in either the pre-AFST or the post-AFST. *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.

TABLE 10A: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
			LOWER	UPPER
1	7.58	0.038	0.43	14.74
2	4.94	0.145	-1.71	11.59
3	8.66	0.005	2.65	14.66
4	6.04	0.136	-1.91	14.00
5	5.34	0.149	-1.92	12.60
6	4.18	0.229	-2.62	10.98
7	5.56	0.172	-2.41	13.54
8	0.46	0.944	-12.38	13.29
9	3.28	0.459	-5.41	11.98
10	2.99	0.582	-7.66	13.64
11	-0.55	0.914	-10.55	9.45

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 10B: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by age-group

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	< 4 years	-0.37	0.940	-9.98	9.24
1	4 to 6 years	11.91	0.050	0.01	23.80
1	7 to 12 years	11.21	0.020	1.79	20.63
1	13 to 17 years	8.17	0.128	-2.35	18.69
2	< 4 years	0.59	0.899	-8.61	9.80
2	4 to 6 years	7.44	0.147	-2.62	17.50
2	7 to 12 years	6.64	0.156	-2.54	15.81
2	13 to 17 years	5.40	0.297	-4.76	15.57
3	< 4 years	3.83	0.401	-5.11	12.77
3	4 to 6 years	6.92	0.168	-2.91	16.76
3	7 to 12 years	9.61	0.015	1.88	17.34
3	13 to 17 years	13.25	0.004	4.29	22.22
4	< 4 years	-1.80	0.747	-12.75	9.14
4	4 to 6 years	2.93	0.663	-10.23	16.09
4	7 to 12 years	12.16	0.025	1.56	22.77
4	13 to 17 years	7.80	0.199	-4.11	19.72
5	< 4 years	0.35	0.946	-9.85	10.56
5	4 to 6 years	14.55	0.013	3.05	26.04
5	7 to 12 years	9.05	0.069	-0.70	18.81
5	13 to 17 years	-0.71	0.900	-11.73	10.32
6	< 4 years	-0.96	0.839	-10.25	8.33
6	4 to 6 years	-2.83	0.592	-13.18	7.52
6	7 to 12 years	6.08	0.184	-2.89	15.05
6	13 to 17 years	12.33	0.020	1.95	22.71
7	< 4 years	1.36	0.820	-10.40	13.12
7	4 to 6 years	12.77	0.053	-0.19	25.73
7	7 to 12 years	4.51	0.400	-5.99	15.02
7	13 to 17 years	5.83	0.296	-5.11	16.78
8	< 4 years	3.85	0.630	-11.85	19.55
8	4 to 6 years	3.44	0.758	-18.42	25.29
8	7 to 12 years	-1.11	0.898	-18.14	15.92
8	13 to 17 years	-3.35	0.749	-23.90	17.19
9	< 4 years	5.05	0.361	-5.79	15.90
9	4 to 6 years	6.37	0.323	-6.26	19.01
9	7 to 12 years	1.69	0.768	-9.59	12.97
9	13 to 17 years	0.50	0.945	-13.74	14.74

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
10	< 4 years	-2.23	0.745	-15.67	11.21
10	4 to 6 years	10.04	0.206	-5.52	25.61
10	7 to 12 years	3.66	0.659	-12.60	19.91
10	13 to 17 years	4.73	0.613	-13.60	23.05
11	< 4 years	-16.45	0.032	-31.49	-1.41
11	4 to 6 years	-0.24	0.975	-15.06	14.58
11	7 to 12 years	7.72	0.234	-4.98	20.41
11	13 to 17 years	5.34	0.509	-10.52	21.21

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 10C: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by race

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	white	12.68	0.013	2.63	22.72
1	Black/African American	5.50	0.274	-4.36	15.37
2	white	6.33	0.213	-3.63	16.29
2	Black/African American	3.30	0.479	-5.83	12.43
3	white	3.87	0.401	-5.17	12.91
3	Black/African American	11.88	0.004	3.84	19.92
4	white	-0.79	0.894	-12.40	10.83
4	Black/African American	10.38	0.071	-0.88	21.64
5	white	11.68	0.029	1.17	22.19
5	Black/African American	1.14	0.828	-9.15	11.42
6	white	9.24	0.08	-1.11	19.59
6	Black/African American	1.89	0.697	-7.62	11.40
7	white	10.64	0.073	-1.01	22.29
7	Black/African American	0.23	0.969	-11.11	11.56
8	white	8.08	0.436	-12.25	28.42
8	Black/African American	-3.32	0.71	-20.82	14.19
9	white	10.34	0.117	-2.57	23.25
9	Black/African American	-3.22	0.594	-15.05	8.61
10	white	8.74	0.203	-4.71	22.19
10	Black/African American	-2.03	0.817	-19.28	15.21
11	white	5.71	0.456	-9.30	20.73
11	Black/African American	-3.02	0.661	-16.49	10.46

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 11A: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, for 11 included screeners

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	42.48	0.67	2.22	40.99	43.97
Post-AFST	11	46.89	0.92	3.06	44.83	48.95
Difference Post-Pre	11	-4.41	0.84	2.77	-6.27	-2.55
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					p value=0.000	
Levene's test of equality of variance between groups—testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					p value=0.375	

TABLE 11B: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by age-group

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	43.73	1.47	4.89	40.45	47.01
Post-AFST (< 4 years)	11	43.11	1.09	3.61	40.69	45.54
Pre-AFST (4 to 6 years)	11	40.07	1.52	5.03	36.69	43.45
Post-AFST (4 to 6 years)	11	46.74	1.43	4.75	43.54	49.93
Pre-AFST (7 to 12 years)	11	42.41	1.07	3.55	40.03	44.80
Post-AFST (7 to 12 years)	11	48.89	0.92	3.06	46.83	50.95
Pre-AFST (13 to 17 years)	11	42.98	1.27	4.23	40.14	45.82
Post-AFST (13 to 17 years)	11	48.37	0.82	2.73	46.54	50.2019
Difference Pre-Post						
< 4 years	11	0.62	1.74	5.77	-3.26	4.49
4 to 6 years	11	-6.66	1.65	5.49	-10.35	-2.98
7 to 12 years	11	-6.47	1.23	4.06	-9.20	-3.74
13 to 17 years	11	-5.39	1.54	5.12	-8.83	-1.95
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value= 0.731	
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value= 0.002	
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value= 0.000	
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value= 0.006	
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value = 0.375	
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value = 0.642	
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value = 0.492	
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value = 0.067	

TABLE 11C: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by race

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (white)	11	41.05	1.21	4.03	38.35	43.76
Post-AFST (white)	11	48.92	1.60	5.29	45.36	52.47
Pre-AFST(Black)	11	44.59	1.08	3.60	42.17	47.00
Post-AFST(Black)	11	46.65	1.12	3.73	44.15	49.16
Difference Pre-Post						
White	11	-7.87	1.18	3.90	-10.49	-5.25
Black/African American	11	-2.07	1.60	5.32	-5.64	1.50
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.000
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.226
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.309
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.862

TABLE 12A: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
			LOWER	UPPER
1	-2.08	0.488	-7.96	3.80
2	-0.49	0.855	-5.72	4.75
3	-1.41	0.545	-5.98	3.16
4	-3.69	0.286	-10.46	3.09
5	-5.04	0.085	-10.78	0.70
6	-2.15	0.399	-7.13	2.84
7	-0.85	0.798	-7.33	5.64
8	-0.89	0.824	-8.71	6.94
9	-1.76	0.585	-8.07	4.55
10	-9.57	0.013	-17.11	-2.02
11	-1.03	0.771	-7.97	5.91

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 12B: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners, by age-group

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	< 4 years	-0.19	0.969	-9.64	9.26
1	4 to 6 years	-10.65	0.051	-21.34	0.04
1	7 to 12 years	3.46	0.317	-3.31	10.22
1	13 to 17 years	-4.83	0.309	-14.13	4.47
2	< 4 years	-1.97	0.687	-11.58	7.63
2	4 to 6 years	-5.65	0.255	-15.37	4.07
2	7 to 12 years	2.41	0.483	-4.32	9.14
2	13 to 17 years	-0.02	0.994	-7.09	7.04
3	< 4 years	1.55	0.632	-4.79	7.89
3	4 to 6 years	-2.66	0.487	-10.17	4.84
3	7 to 12 years	-1.79	0.582	-8.16	4.58
3	13 to 17 years	-2.36	0.511	-9.40	4.68
4	< 4 years	-4.77	0.350	-14.79	5.24
4	4 to 6 years	-8.90	0.151	-21.05	3.26
4	7 to 12 years	-5.10	0.264	-14.06	3.85
4	13 to 17 years	1.38	0.798	-9.18	11.94
5	< 4 years	-7.87	0.118	-17.75	2.00
5	4 to 6 years	-7.80	0.125	-17.77	2.17
5	7 to 12 years	-0.67	0.853	-7.77	6.42
5	13 to 17 years	-6.81	0.100	-14.92	1.30
6	< 4 years	-7.00	0.124	-15.91	1.91
6	4 to 6 years	-8.65	0.076	-18.22	0.92
6	7 to 12 years	1.61	0.594	-4.31	7.54
6	13 to 17 years	-0.57	0.874	-7.64	6.50
7	< 4 years	8.36	0.123	-2.26	18.98
7	4 to 6 years	3.19	0.577	-8.03	14.41
7	7 to 12 years	-9.08	0.045	-17.95	-0.21
7	13 to 17 years	-0.23	0.965	-10.49	10.03
8	< 4 years	-0.60	0.921	-12.43	11.23
8	4 to 6 years	0.83	0.893	-11.24	12.91
8	7 to 12 years	-7.30	0.213	-18.78	4.18
8	13 to 17 years	6.43	0.427	-9.43	22.29
9	< 4 years	-17.85	0.005	-30.19	-5.52
9	4 to 6 years	-3.27	0.510	-12.99	6.46
9	7 to 12 years	2.53	0.560	-5.96	11.02
9	13 to 17 years	2.89	0.460	-4.77	10.55

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
10	< 4 years	-1.95	0.793	-16.56	12.66
10	4 to 6 years	-18.71	0.001	-29.95	-7.46
10	7 to 12 years	-10.93	0.036	-21.12	-0.73
10	13 to 17 years	-8.51	0.088	-18.30	1.28
11	< 4 years	1.50	0.799	-10.06	13.06
11	4 to 6 years	0.92	0.856	-9.03	10.88
11	7 to 12 years	-5.82	0.181	-14.33	2.70
11	13 to 17 years	2.61	0.670	-9.41	14.63

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 12C: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners, by race

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	white	-0.38	0.932	-9.22	8.45
1	Black/African American	-4.36	0.321	-12.97	4.25
2	white	-3.10	0.479	-11.70	5.49
2	Black/African American	0.86	0.805	-5.95	7.67
3	white	-4.47	0.198	-11.27	2.34
3	Black/African American	1.16	0.730	-5.41	7.73
4	white	-4.70	0.411	-15.91	6.51
4	Black/African American	-0.86	0.848	-9.66	7.93
5	white	-9.22	0.037	-17.88	-0.56
5	Black/African American	-2.54	0.517	-10.24	5.15
6	white	-2.04	0.589	-9.43	5.36
6	Black/African American	-3.89	0.320	-11.55	3.78
7	white	-4.96	0.311	-14.58	4.65
7	Black/African American	4.24	0.411	-5.88	14.35
8	white	7.11	0.278	-5.74	19.97
8	Black/African American	-7.18	0.169	-17.41	3.06
9	white	0.46	0.919	-8.39	9.32
9	Black/African American	-4.06	0.399	-13.48	5.37
10	white	-1.81	0.783	-14.71	11.09
10	Black/African American	-16.43	0.001	-26.45	-6.40
11	white	4.28	0.402	-5.72	14.27
11	Black/African American	-7.03	0.190	-17.56	3.49

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 13A: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, for 11 included screeners

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	85.02	0.43	1.43	84.07	85.98
Post-AFST	11	82.39	0.63	2.09	80.99	83.80
Difference Pre-Post	11	2.63	0.80	2.67	0.84	4.42
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					pvalue=	0.008
Levene's test of equality of variance between groups - testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					pvalue=	0.296

TABLE 13B: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by age-group

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	85.25	1.24	4.10	82.49	88.00
Post-AFST (< 4 years)	11	82.45	1.53	5.07	79.04	85.85
Pre-AFST (4 to 6 years)	11	86.17	0.85	2.80	84.28	88.05
Post-AFST (4 to 6 years)	11	80.59	1.42	4.69	77.44	83.74
Pre-AFST (7 to 12 years)	11	85.54	0.62	2.04	84.16	86.91
Post-AFST (7 to 12 years)	11	82.75	1.22	4.06	80.02	85.48
Pre-AFST (13 to 17 years)	11	83.54	1.03	3.43	81.23	85.84
Post-AFST (13 to 17 years)	11	82.63	0.95	3.15	80.51	84.74
Difference Pre - Post						
< 4 years	11	2.80	2.02	6.70	-1.70	7.30
4 to 6 years	11	5.58	1.90	6.30	1.34	9.81
7 to 12 years	11	2.79	1.54	5.10	-0.64	6.21
13 to 17 years	11	0.91	1.34	4.44	-2.07	3.90
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value=	0.196
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value=	0.015
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value=	0.100
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value=	0.512
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value =	0.346
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value =	0.062
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value =	0.036
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value =	0.779

TABLE 13C: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by race

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (white)	11	82.73	0.52	1.73	81.56	83.89
Post-AFST (white)	11	81.02	1.05	3.49	78.67	83.36
Pre-AFST(Black)	11	85.82	0.96	3.19	83.68	87.97
Post-AFST(Black)	11	82.18	1.10	3.65	79.72	84.63
Difference Pre-Post						
White	11	1.71	1.37	4.54	-1.33	4.76
Black/African American	11	3.65	1.66	5.50	-0.05	7.34
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.239
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.053
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.051
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.693

TABLE 14A: Consistency in workload, adjusted analysis, for 11 included call screeners

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
			LOWER	UPPER
1	6.75	0.008	1.74	11.76
2	1.27	0.613	-3.66	6.21
3	2.82	0.193	-1.43	7.08
4	-3.39	0.269	-9.40	2.62
5	4.75	0.066	-0.31	9.82
6	-0.78	0.742	-5.44	3.88
7	5.28	0.054	-0.10	10.67
8	-5.97	0.184	-14.76	2.83
9	7.47	0.01	1.80	13.14
10	-5.85	0.117	-13.17	1.47
11	7.42	0.048	0.06	14.78

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 14B: Consistency in workload, adjusted analysis, for 11 included call screeners, by age-group

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	< 4 years	1.60	0.662	-5.59	8.80
1	4 to 6 years	0.05	0.990	-8.21	8.32
1	7 to 12 years	12.08	0.000	5.67	18.49
1	13 to 17 years	9.16	0.021	1.37	16.94
2	< 4 years	-0.15	0.968	-7.63	7.33
2	4 to 6 years	4.15	0.283	-3.42	11.72
2	7 to 12 years	2.55	0.440	-3.92	9.01
2	13 to 17 years	-1.52	0.677	-8.67	5.63
3	< 4 years	-0.36	0.909	-6.62	5.89
3	4 to 6 years	8.97	0.009	2.23	15.71
3	7 to 12 years	4.96	0.080	-0.59	10.51
3	13 to 17 years	-2.05	0.534	-8.53	4.42
4	< 4 years	-5.81	0.212	-14.94	3.32
4	4 to 6 years	2.78	0.581	-7.08	12.64
4	7 to 12 years	-0.52	0.894	-8.13	7.10
4	13 to 17 years	-9.72	0.034	-18.70	-0.74
5	< 4 years	-0.18	0.962	-7.69	7.32
5	4 to 6 years	10.59	0.013	2.27	18.90
5	7 to 12 years	5.55	0.099	-1.05	12.15
5	13 to 17 years	3.81	0.307	-3.50	11.11
6	< 4 years	0.43	0.905	-6.66	7.53
6	4 to 6 years	-1.43	0.705	-8.83	5.97
6	7 to 12 years	1.26	0.680	-4.74	7.26
6	13 to 17 years	-4.55	0.194	-11.42	2.32
7	< 4 years	4.07	0.375	-4.92	13.06
7	4 to 6 years	4.33	0.331	-4.40	13.07
7	7 to 12 years	7.29	0.040	0.35	14.23
7	13 to 17 years	3.78	0.339	-3.96	11.52
8	< 4 years	-3.24	0.598	-15.28	8.80
8	4 to 6 years	10.14	0.163	-4.11	24.40
8	7 to 12 years	-8.91	0.118	-20.08	2.26
8	13 to 17 years	-17.23	0.020	-31.74	-2.71
9	< 4 years	12.24	0.004	4.02	20.45
9	4 to 6 years	7.03	0.103	-1.41	15.48
9	7 to 12 years	9.31	0.013	1.97	16.65
9	13 to 17 years	0.10	0.981	-8.46	8.67

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
10	< 4 years	-3.29	0.541	-13.84	7.26
10	4 to 6 years	-1.06	0.859	-12.78	10.67
10	7 to 12 years	-5.22	0.290	-14.91	4.46
10	13 to 17 years	-10.80	0.074	-22.66	1.07
11	< 4 years	9.06	0.078	-1.00	19.12
11	4 to 6 years	8.88	0.089	-1.37	19.13
11	7 to 12 years	2.18	0.658	-7.49	11.85
11	13 to 17 years	12.99	0.021	1.93	24.06

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 14C: Consistency in workload, adjusted analysis, for 11 included call screeners, by race

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.]	
				LOWER	UPPER
1	white	11.69	0.001	4.51	18.87
1	Black/African American	3.45	0.345	-3.72	10.62
2	white	2.81	0.431	-4.18	9.81
2	Black/African American	1.31	0.713	-5.65	8.27
3	white	2.13	0.483	-3.83	8.09
3	Black/African American	3.22	0.305	-2.93	9.38
4	white	-3.47	0.458	-12.62	5.68
4	Black/African American	-4.45	0.292	-12.73	3.83
5	white	10.89	0.002	3.90	17.89
5	Black/African American	0.10	0.979	-7.31	7.51
6	white	-0.25	0.943	-7.08	6.59
6	Black/African American	-2.82	0.408	-9.49	3.86
7	white	12.78	0.001	5.29	20.26
7	Black/African American	-2.79	0.489	-10.69	5.11
8	white	0.29	0.971	-15.02	15.60
8	Black/African American	-7.24	0.223	-18.86	4.39
9	white	7.03	0.078	-0.79	14.84
9	Black/African American	9.62	0.022	1.41	17.82
10	white	-16.18	0.002	-26.35	-6.00
10	Black/African American	4.62	0.387	-5.84	15.08
11	white	6.16	0.236	-4.02	16.34
11	Black/African American	9.68	0.082	-1.21	20.57

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 15A: Means and variance of screener's predicted workload, adjusted analysis, for 11 included screeners

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	49.35	1.20	3.97	46.68	52.02
Post-AFST	11	51.15	1.06	3.52	48.79	53.51
Difference Pre - Post	11	-1.80	1.55	5.13	-5.24	1.65
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					pvalue=	0.272
Levene's test of equality of variance between groups—testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					pvalue=	0.673

TABLE 15B: Means and variance of screener's predicted workload, adjusted analysis, by age-group

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	51.29	1.34	4.43	48.31	54.27
Post-AFST (< 4 years)	11	52.59	0.87	2.88	50.66	54.53
Pre-AFST (4 to 6 years)	11	48.24	1.61	5.34	44.65	51.82
Post-AFST (4 to 6 years)	11	53.19	1.09	3.63	50.75	55.62
Pre-AFST (7 to 12 years)	11	49.95	1.17	3.88	47.34	52.56
Post-AFST (7 to 12 years)	11	52.73	1.45	4.82	49.49	55.97
Pre-AFST (13 to 17 years)	11	47.41	2.07	6.87	42.79	52.03
Post-AFST (13 to 17 years)	11	45.95	1.80	5.96	41.95	49.96
Difference Pre-Post						
< 4 years	11	-1.31	1.62	5.36	-4.91	2.30
4 to 6 years	11	-4.95	1.35	4.49	-7.96	-1.94
7 to 12 years	11	-2.78	1.85	6.13	-6.90	1.34
13 to 17 years	11	1.46	2.69	8.91	-4.53	7.44
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value=	0.438
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value=	0.004
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value=	0.164
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value=	0.600
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value =	0.226
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value =	0.277
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value =	0.578
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value =	0.616

TABLE 15C: Means and variance of screener's predicted workload, adjusted analysis, by race

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (white)	11	49.36	1.77	5.87	45.41	53.30
Post-AFST (white)	11	52.44	1.40	4.65	49.32	55.56
Pre-AFST(Black)	11	49.53	1.76	5.85	45.60	53.45
Post-AFST(Black)	11	50.86	0.99	3.27	48.67	53.06
Difference Pre-Post						
White	11	-3.08	2.50	8.29	-8.65	2.49
Black/African American	11	-1.34	1.65	5.47	-5.01	2.34
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.246
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.437
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.445
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.127

TABLE 16A: Estimated magnitude of monthly impact of AFST on accuracy of screen-in

	ESTIMATED TOTAL # OF CHILDREN WITH ACCURATE SCREEN-IN PER MONTH						ESTIMATED # OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFS-PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	358	346	370	381	369	394	24	8	39
< 4 years	93	89	98	93	89	97	0	-5	4
4-6 years	61	58	65	66	62	69	4	-1	9
7-12 years	123	118	128	136	131	141	13	6	20
13-17 years	83	79	86	89	86	93	7	1	13
white	135	129	142	156	148	164	20	13	27
black	201	191	212	201	193	209	0	-14	14

Estimates are based on predicted probabilities of accuracy of screen-in (Tables 4) and mean number of children screened-in per month over entire analysis period. The total average number of children screened-in per month over the entire analysis period is 819 with 26%, 17%, 34% & 23% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 38%, 53% for white and Black/African American, respectively.

TABLE 16B: Estimated magnitude of impact of AFST on accuracy of screen-out

	ESTIMATED TOTAL # CHILDREN WITH ACCURATE SCREEN-OUT PER MONTH						ESTIMATED # OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFST-PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	774	768	780	763	754	773	-11	-22	1
< 4 years	149	146	151	146	143	150	-2	-7	2
4-6 years	136	134	138	132	129	135	-4	-8	0
7-12 years	287	285	290	283	278	288	-4	-11	2
13-17 years	209	207	212	209	205	212	-1	-5	4
white	340	335	344	336	328	343	-4	-12	4
black	353	348	358	345	339	352	-8	-16	1

Estimates are based on predicted probabilities of accuracy of screen-out (Tables 6) and mean number of children screened-out per month over entire analysis period. The total average number of children screened-out per month over the entire analysis period is 914 with 19%, 17%, 37% & 27% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 45%, 46% for white and Black/African American, respectively.

TABLE 16C: Estimated magnitude of impact of AFST on workload

	ESTIMATED TOTAL # OF CHILDREN SCREENED-IN PER MONTH						ESTIMATED # OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFST-PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	851	817	884	797	745	850	-53	-113	6
< 4 years	198	188	207	191	182	200	-7	-18	4
4-6 years	145	138	152	144	134	154	-1	-11	10
7-12 years	307	295	320	289	267	310	-19	-44	6
13-17 years	208	199	218	181	166	195	-28	-46	-9
white	356	339	373	342	320	364	-13	-41	15
black	430	410	451	396	369	424	-34	-66	-2

Estimates are based on predicted probabilities of workload (Tables 8) and mean number of children in referrals per month over entire analysis period. The total average number of children in referrals per month over the entire analysis period is 1745 with 22%, 17%, 36% & 25% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 42%, 49% for white and Black/African American, respectively.

FIGURES

FIGURE 1: Example of the AFST Score

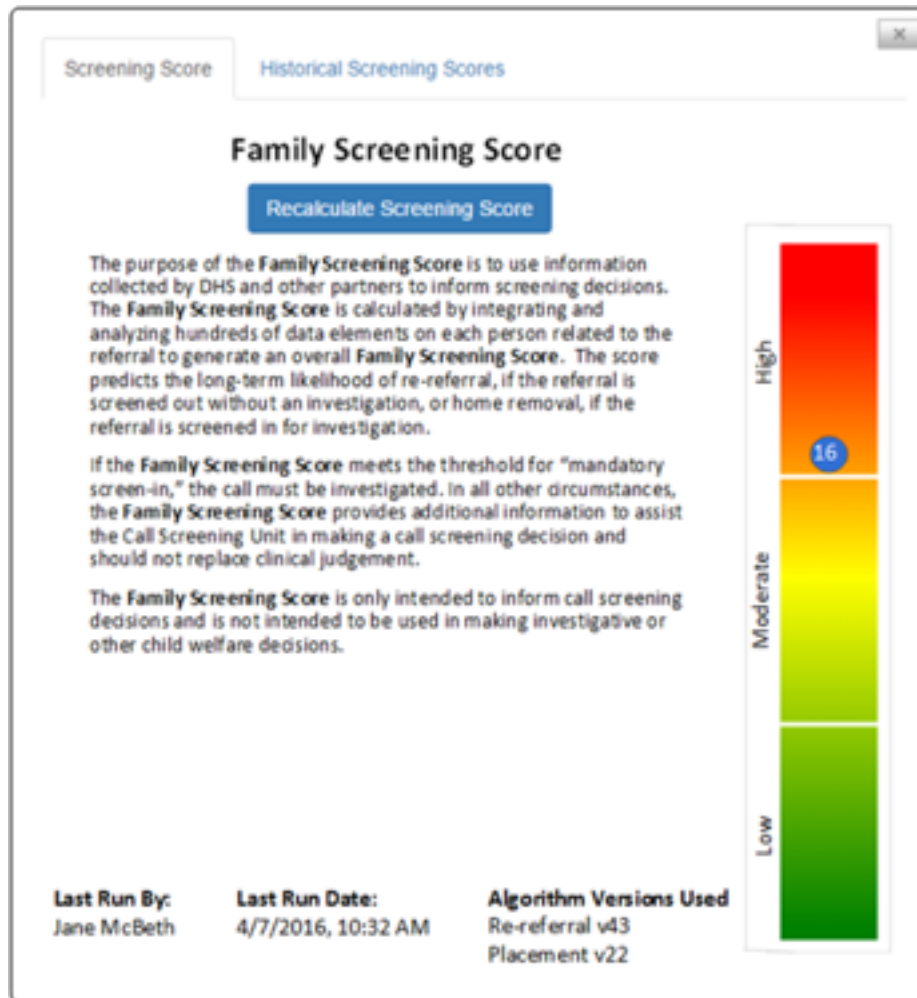


FIGURE 2A: Accuracy of Screen-In, ITSA analysis

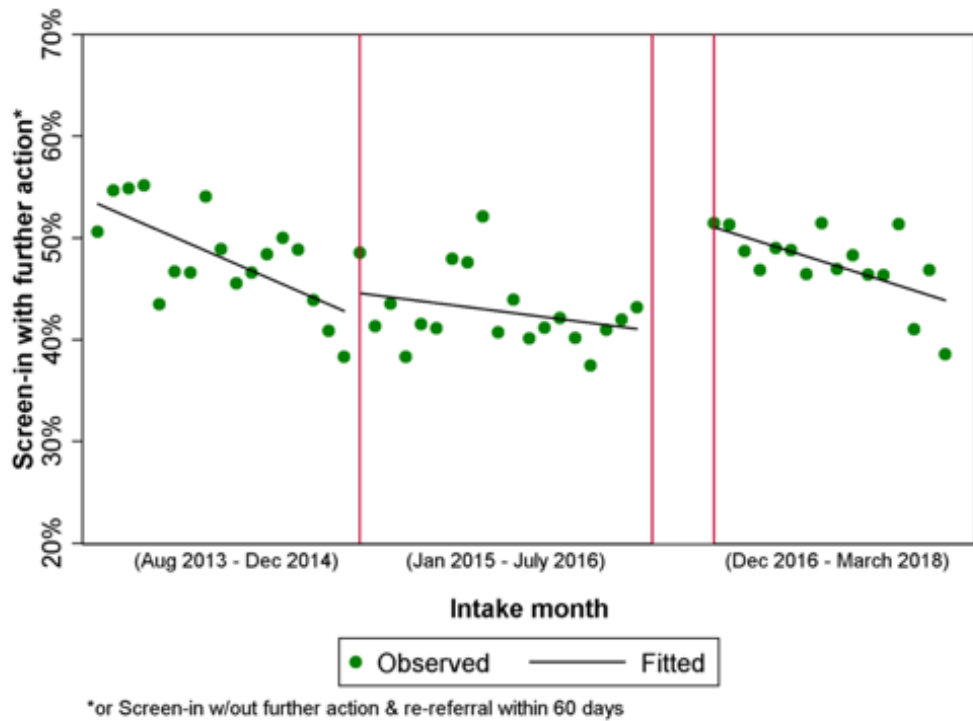


FIGURE 2B: Accuracy of Screen-In, by age-group

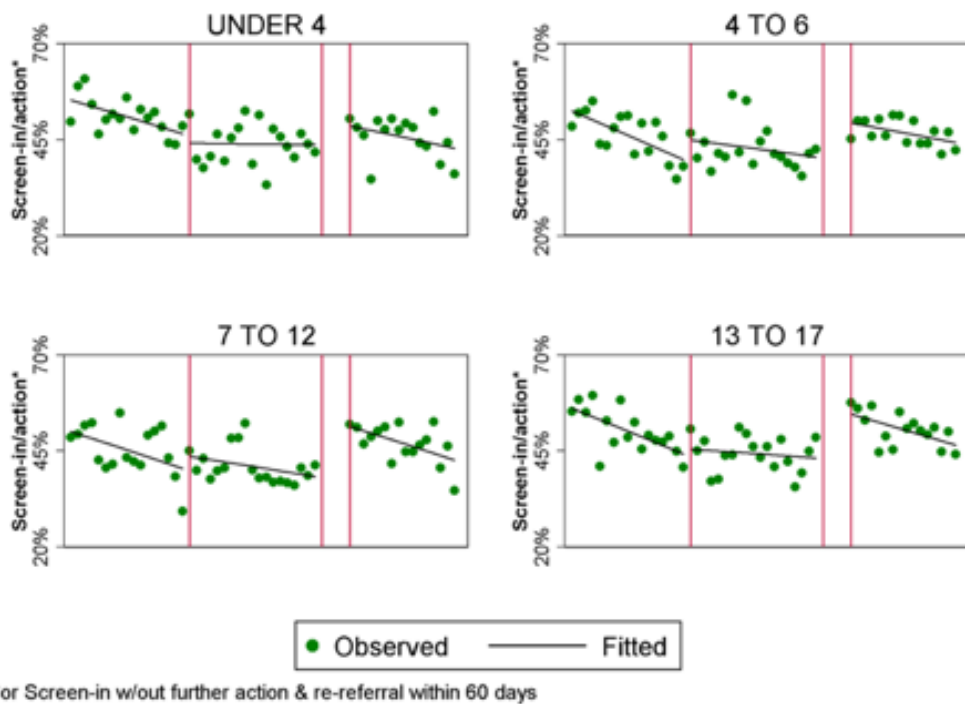


FIGURE 2C: Accuracy of Screen-In, by race

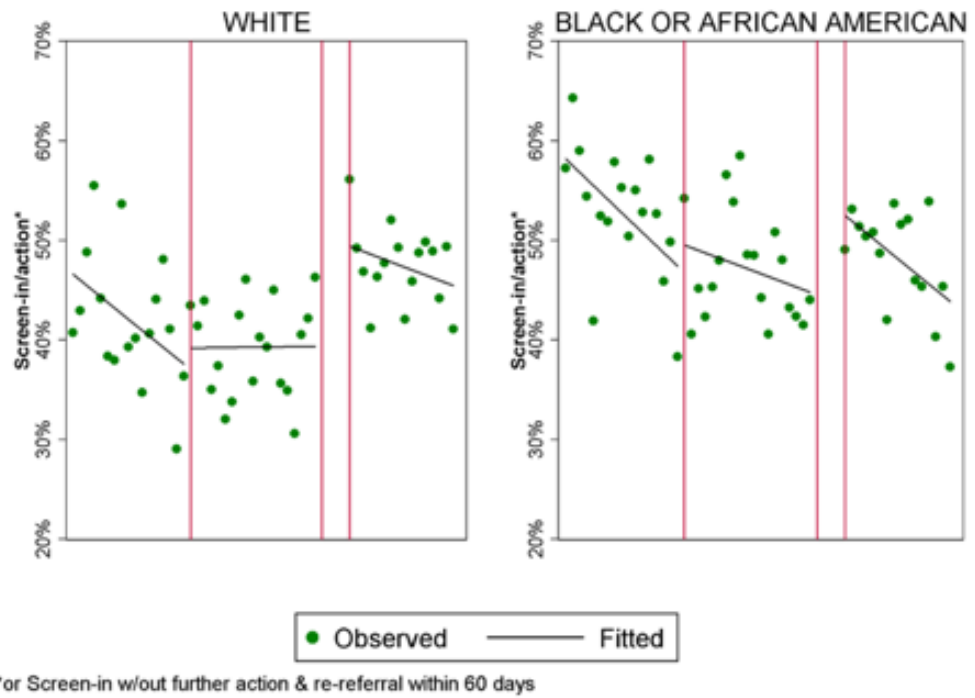


FIGURE 3A: Accuracy of Screen-In, adjusted analysis

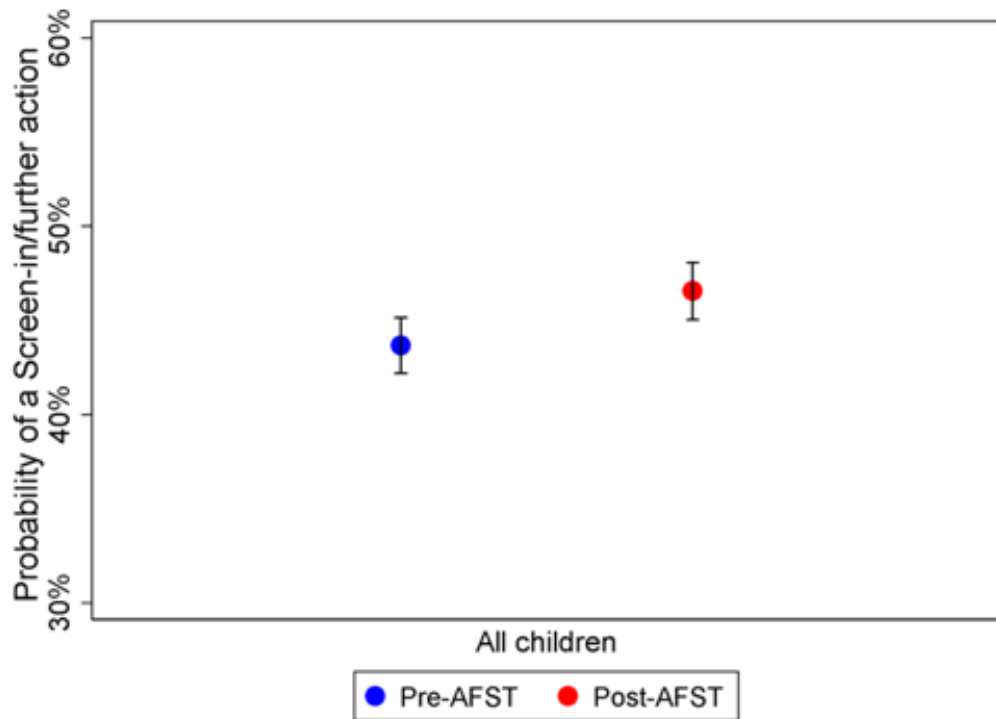


FIGURE 3B: Accuracy of Screen-In, adjusted analysis, by age-groups

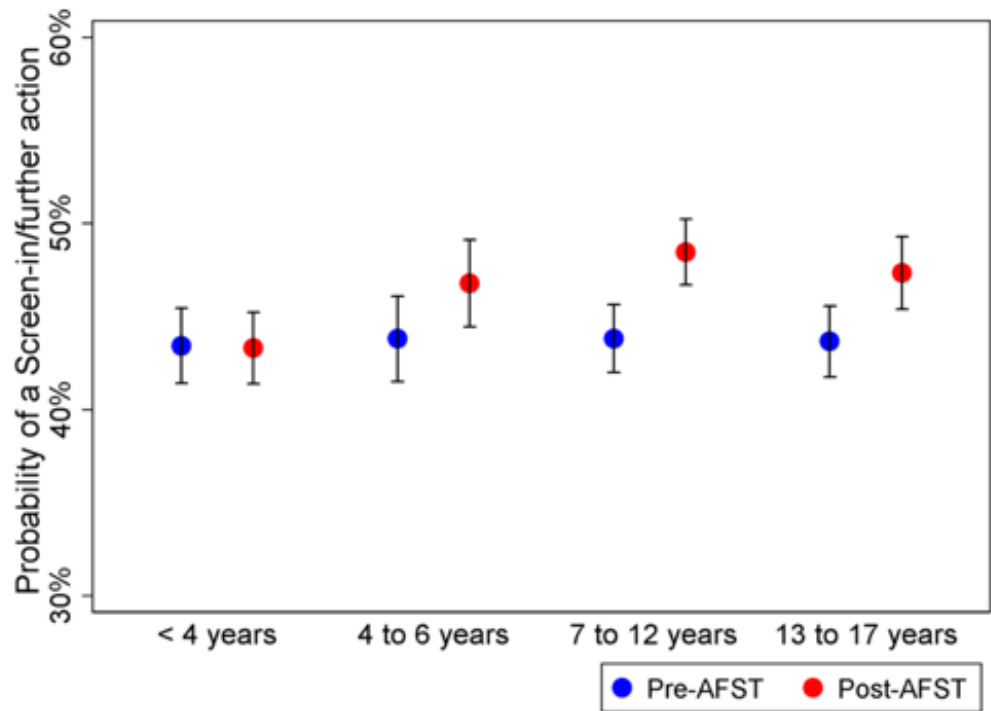


FIGURE 3C: Accuracy of Screen-In, adjusted analysis, by race

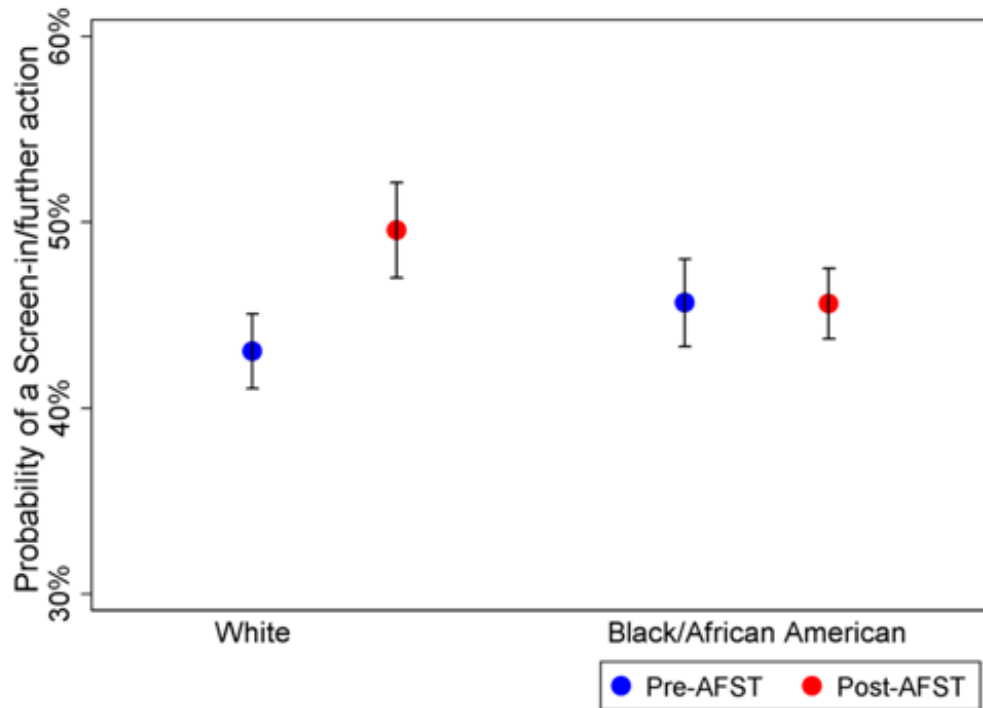


FIGURE 4A: Accuracy of Screen-Out, ITSA analysis

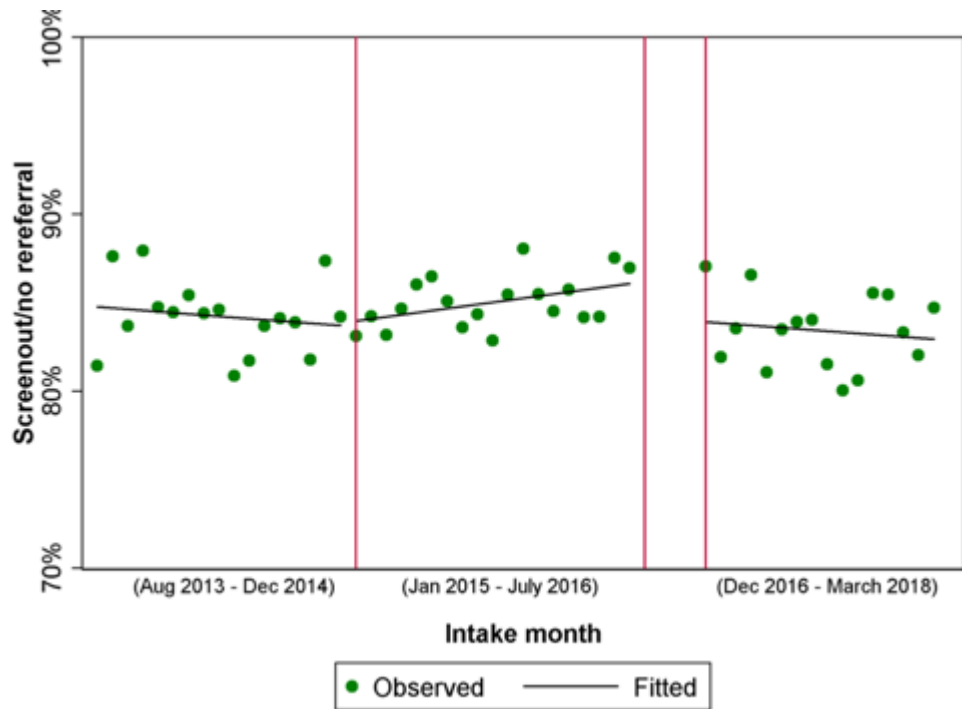


FIGURE 4B: Accuracy of Screen-Out, ITSA analysis, by age-group

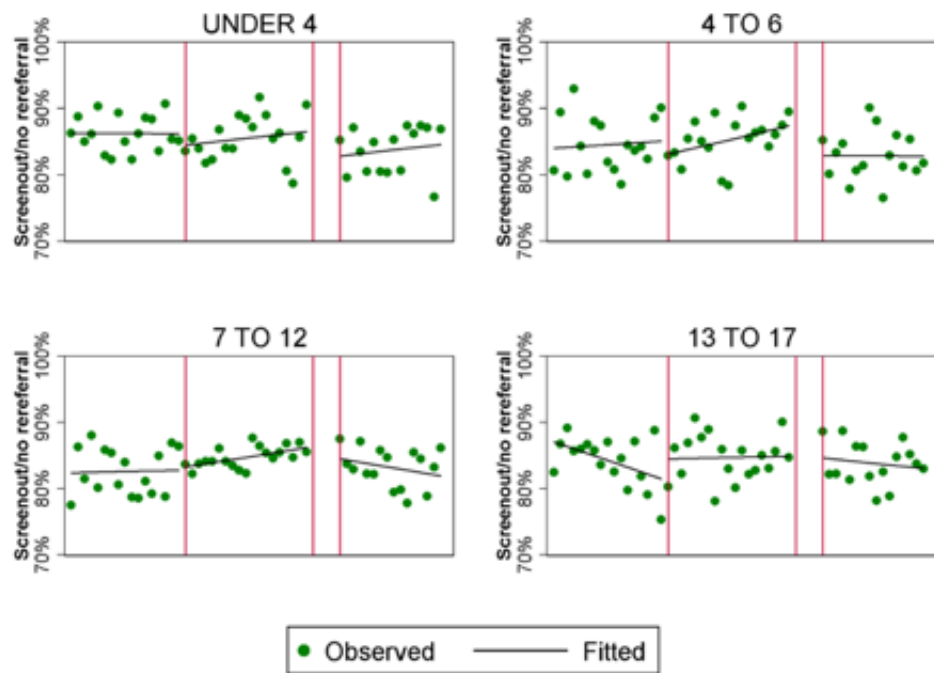


FIGURE 4C: Accuracy of Screen-Out, ITSA analysis, by race

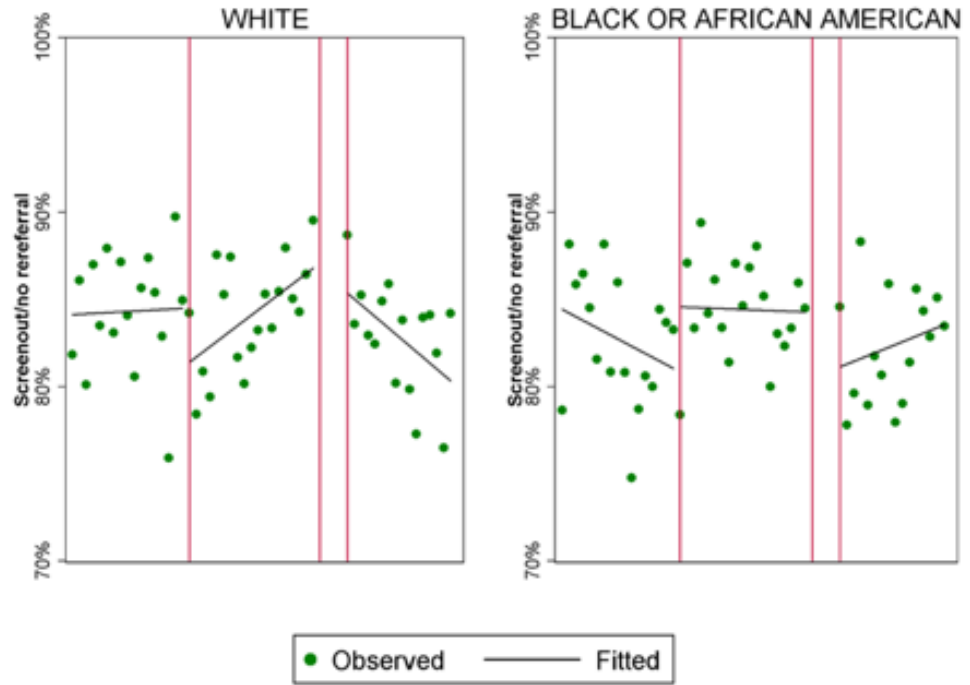


FIGURE 5A: Accuracy of Screen-Out, adjusted analysis

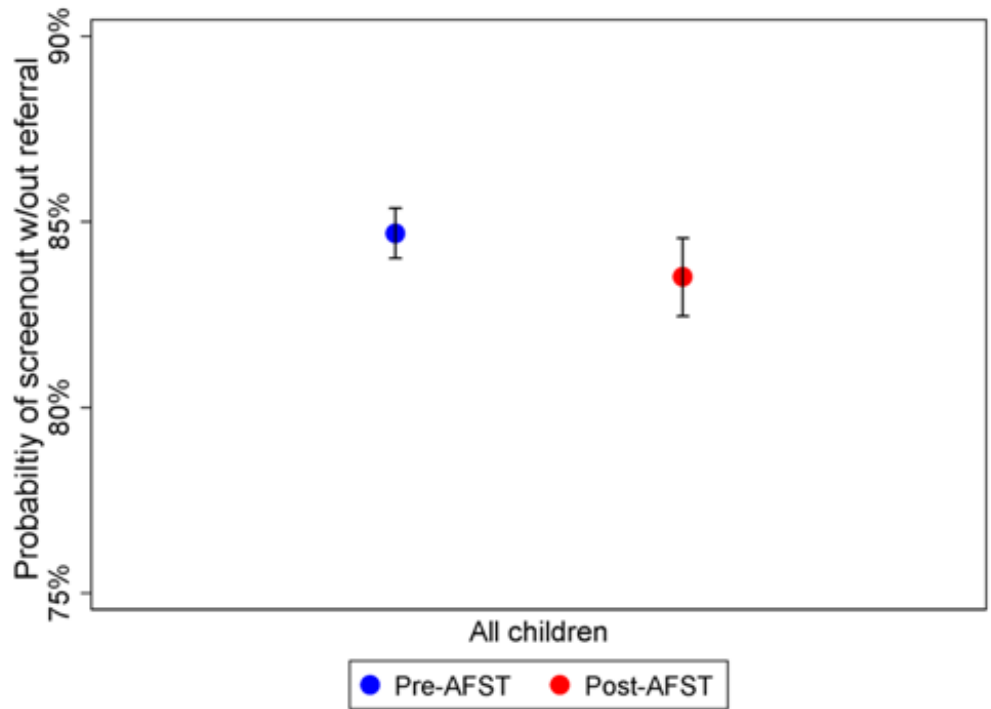


FIGURE 5B: Accuracy of Screen-Out, adjusted analysis, by age-group

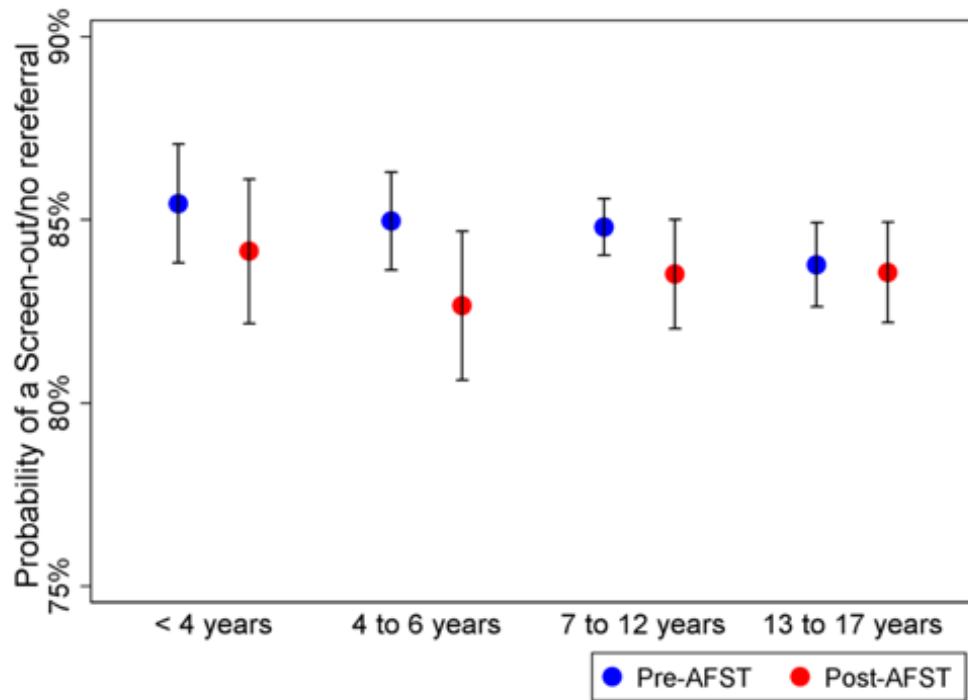


FIGURE 5C: Accuracy of Screen-Out, adjusted analysis, by race

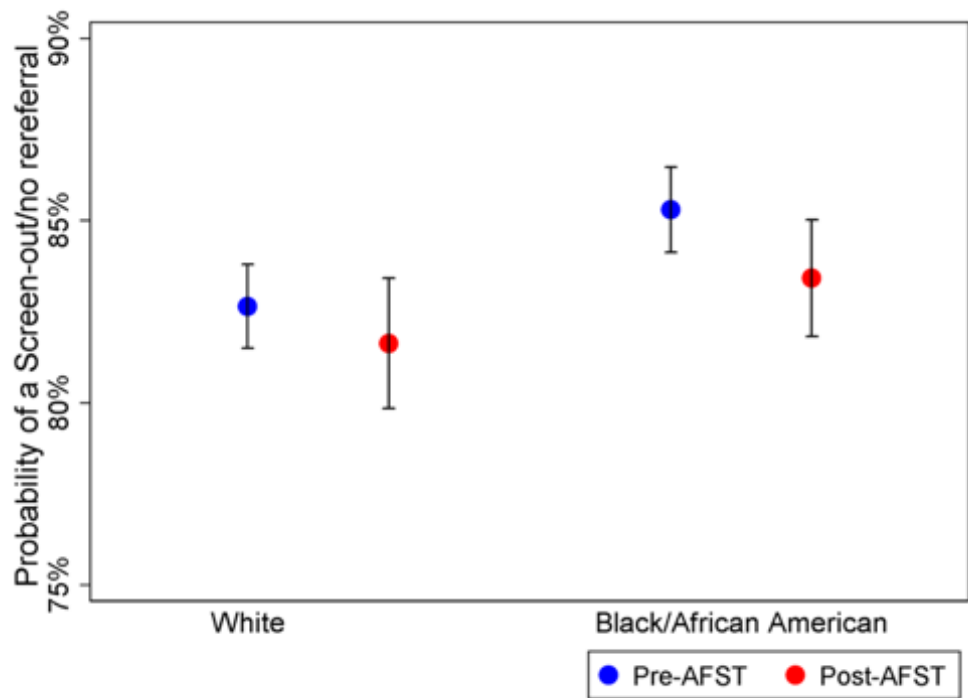


FIGURE 6A: Workload, ITSA Analysis

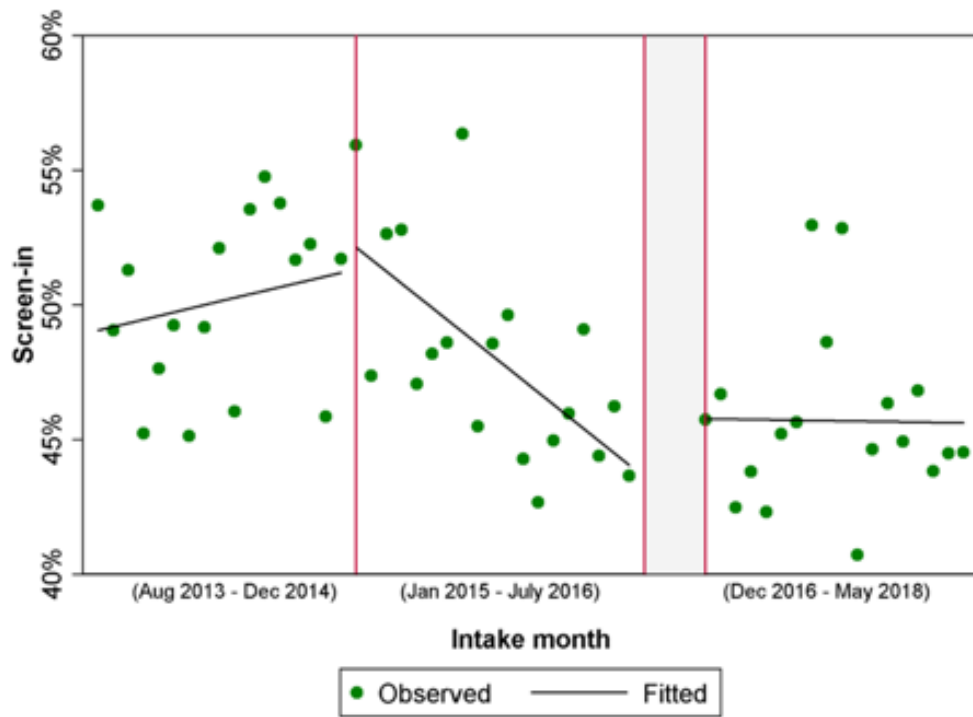


FIGURE 6B: Workload, ITSA Analysis, by age-group

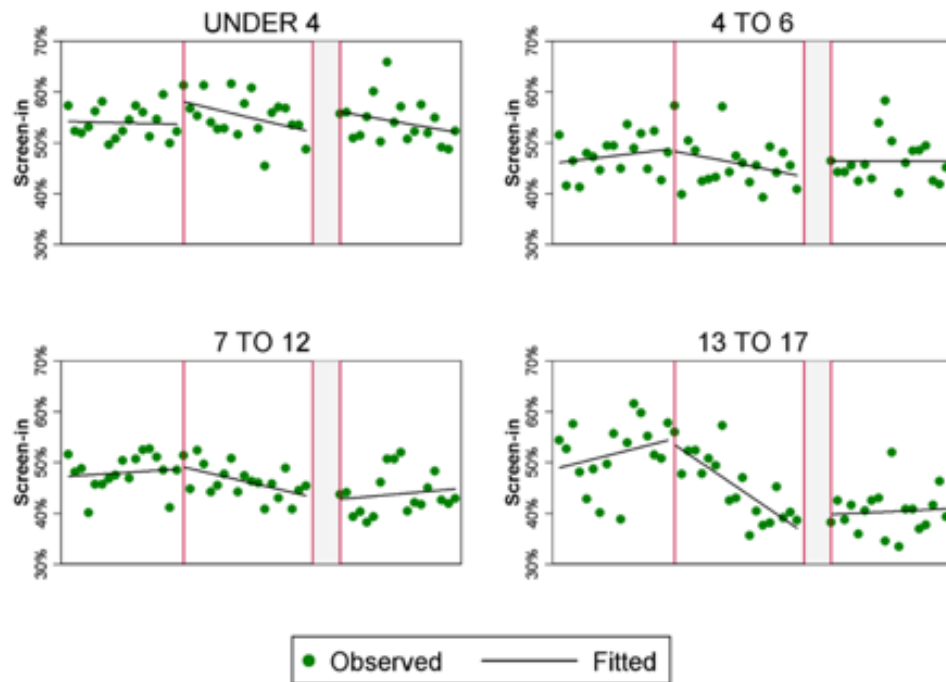


FIGURE 6C: Workload, ITSA Analysis, by race

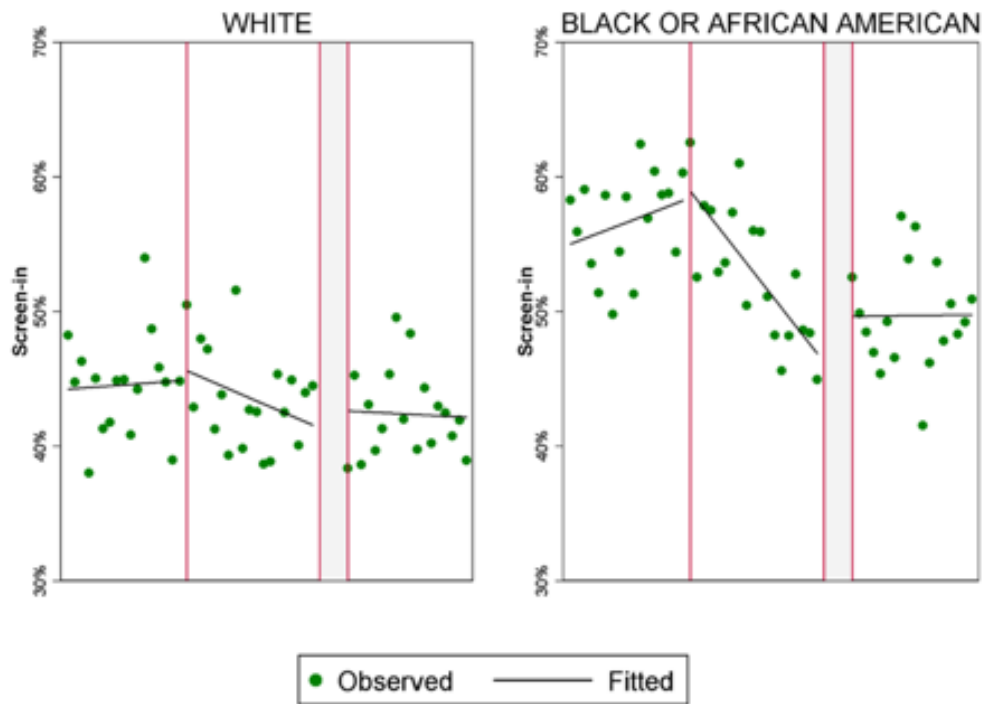


FIGURE 7A: Workload, adjusted analysis

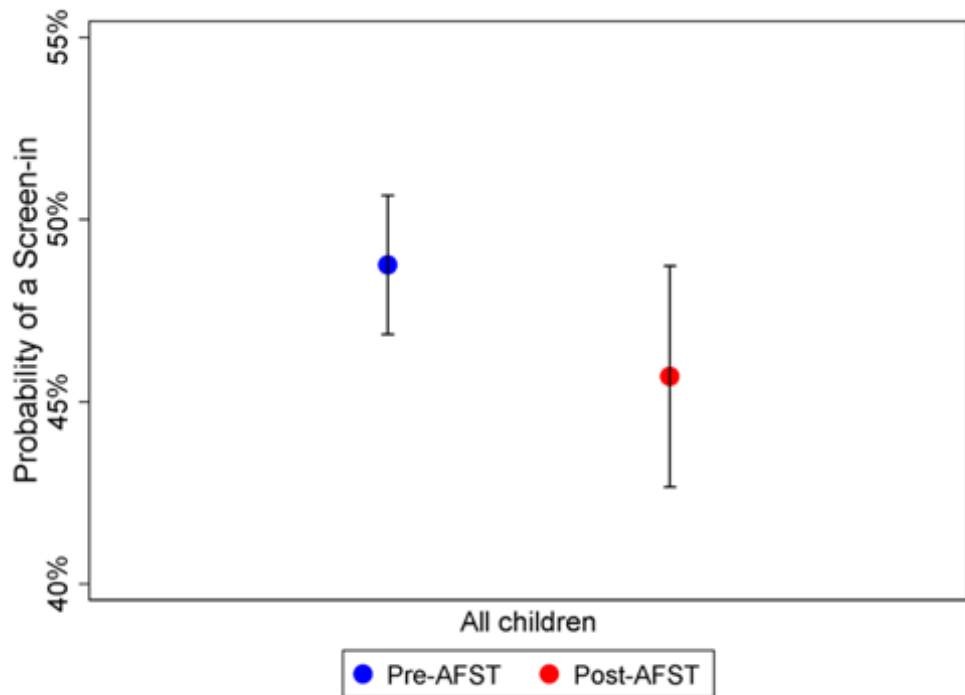


FIGURE 7B: Workload, adjusted analysis, by age-group

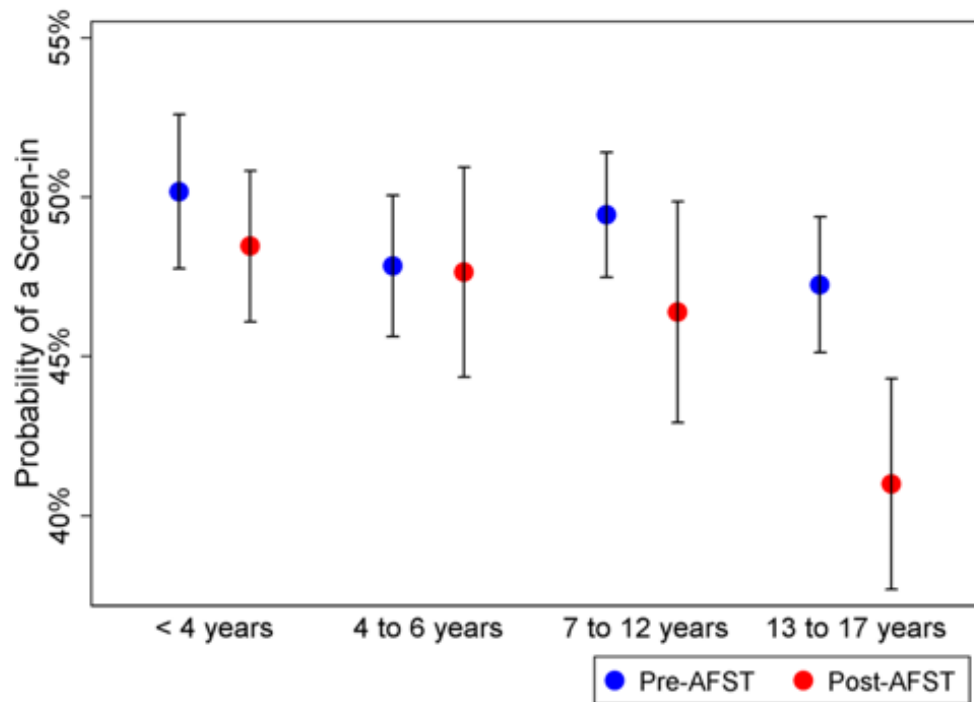


FIGURE 7C: Workload, adjusted analysis, by race

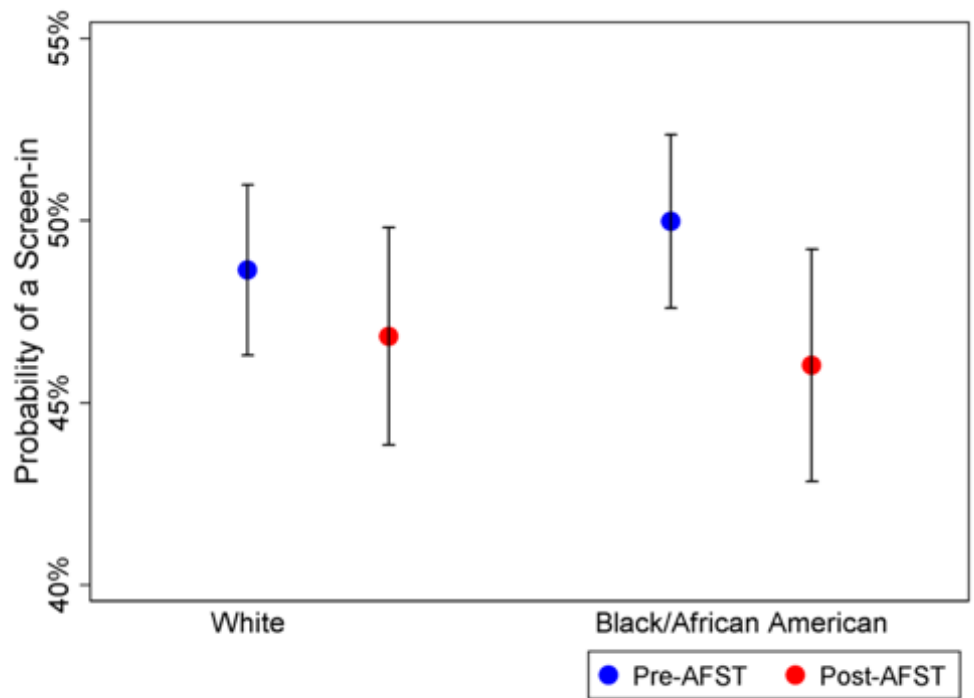


FIGURE 8A: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis

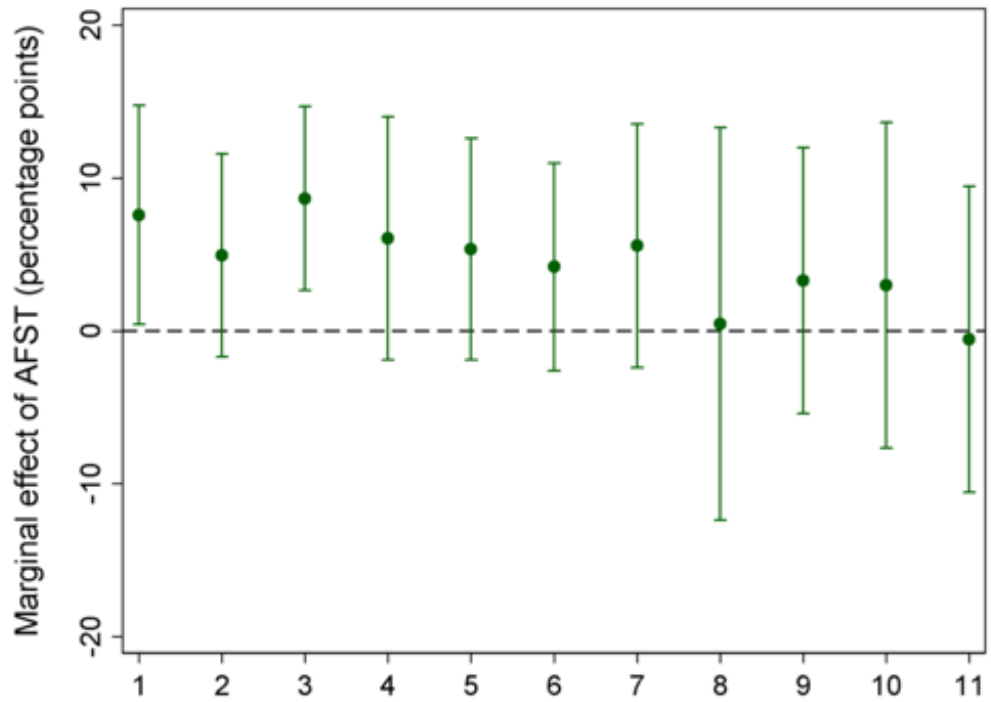


FIGURE 8B: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by age-group

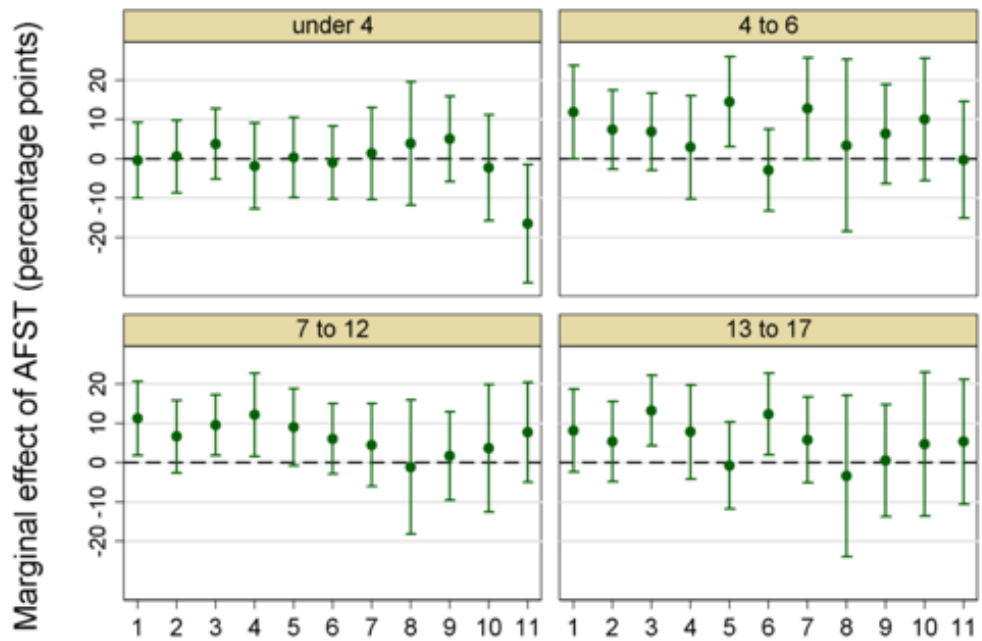


FIGURE 8C: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by race

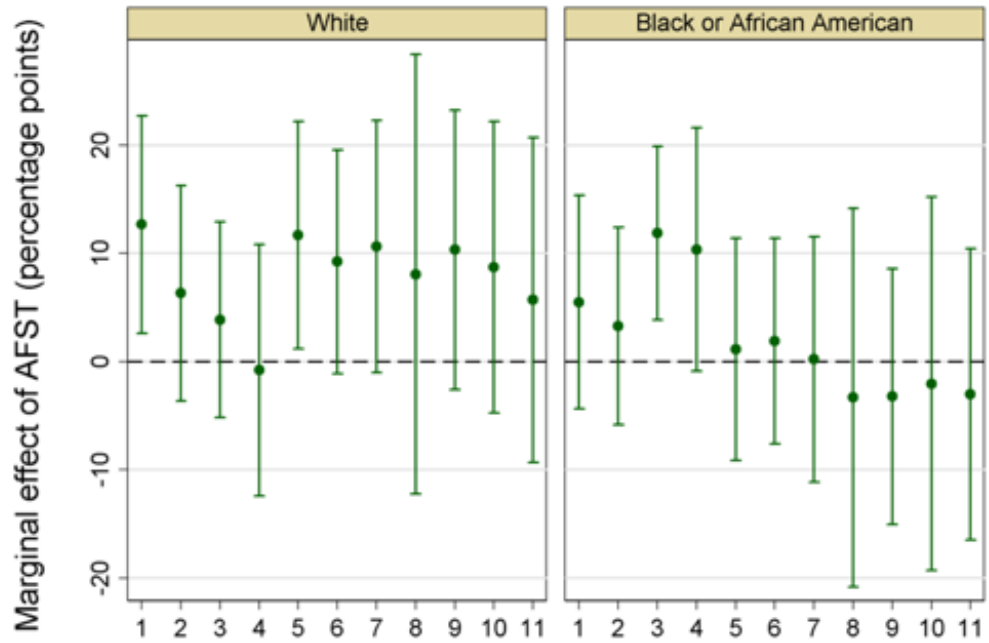


FIGURE 9A: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis

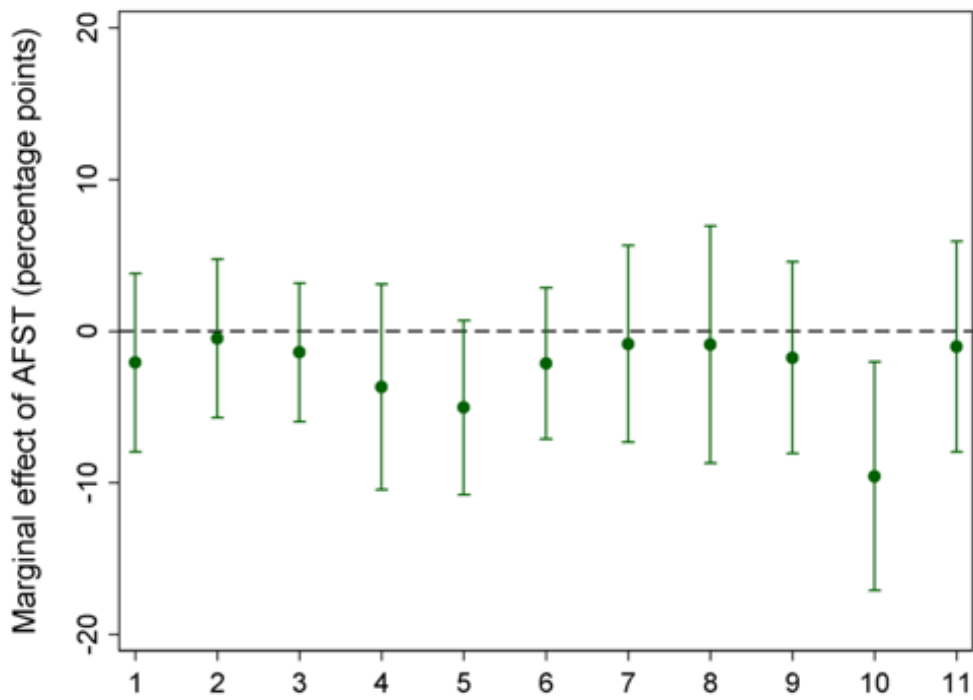


FIGURE 9B: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by age-group

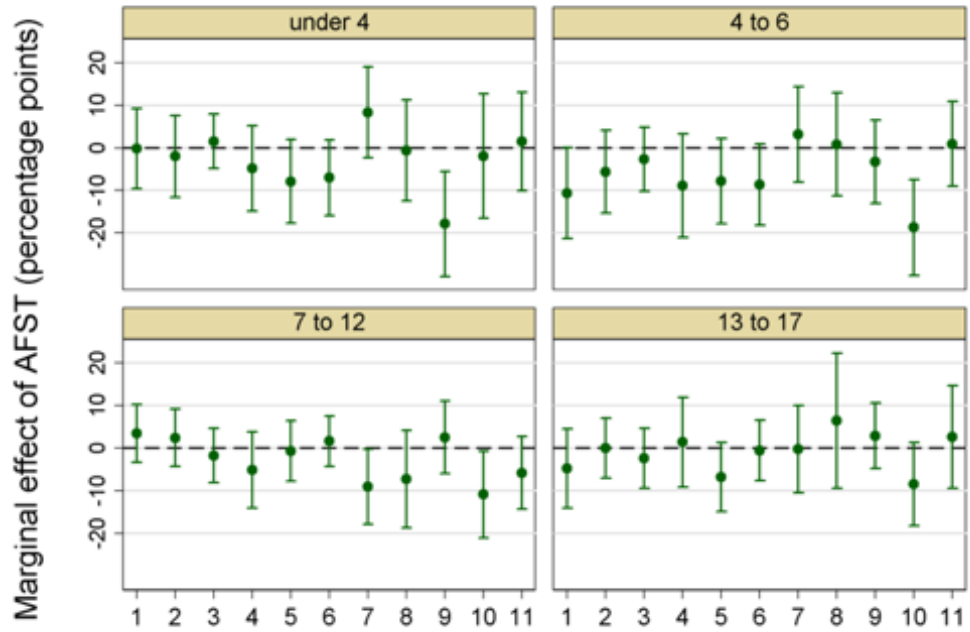


FIGURE 9C: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by race

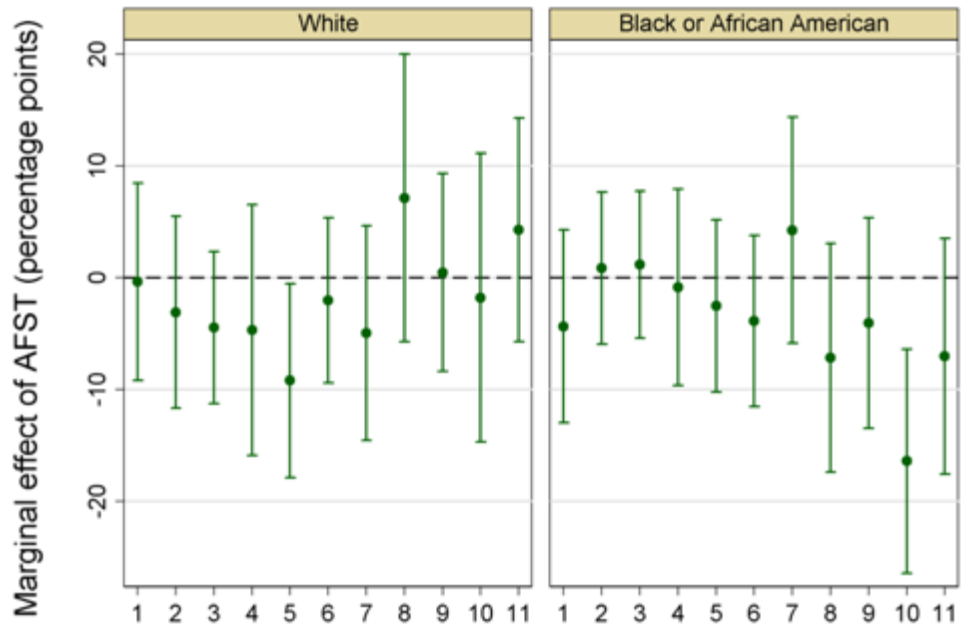


FIGURE 10A: Consistency of workload for 11 call screeners, adjusted analysis

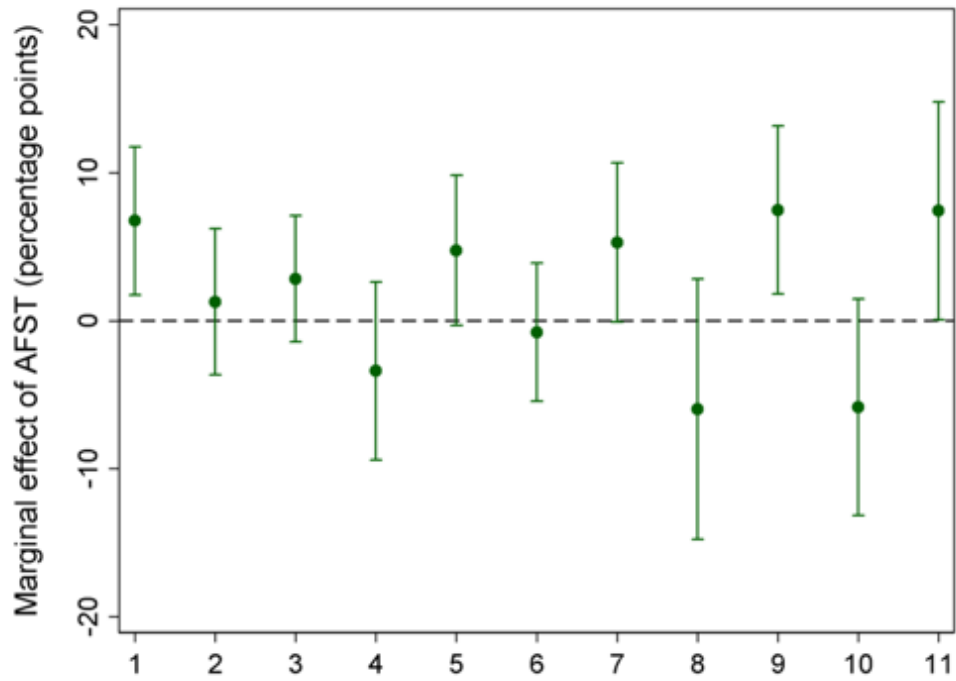


FIGURE 10B: Consistency of workload for 11 call screeners, adjusted analysis, by age-group

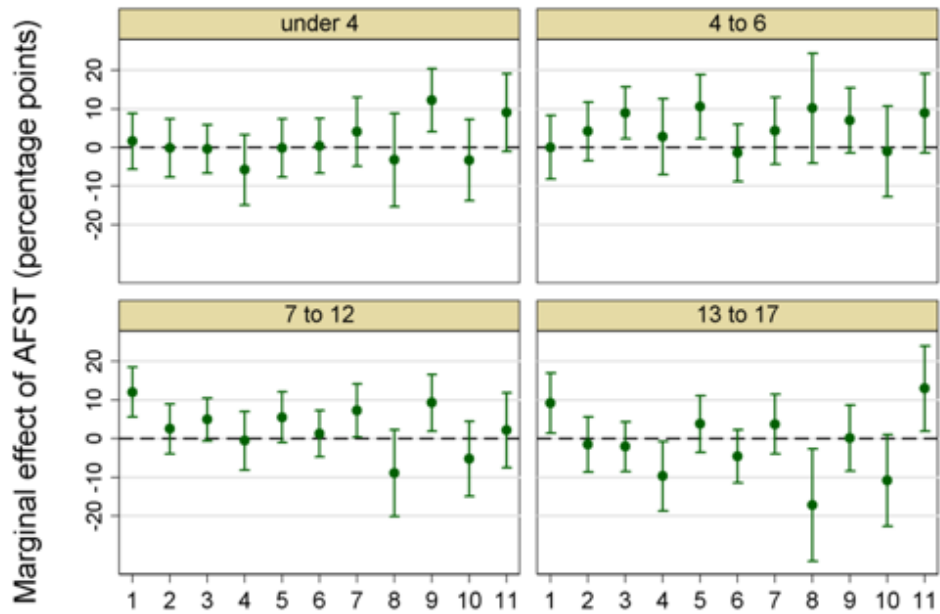
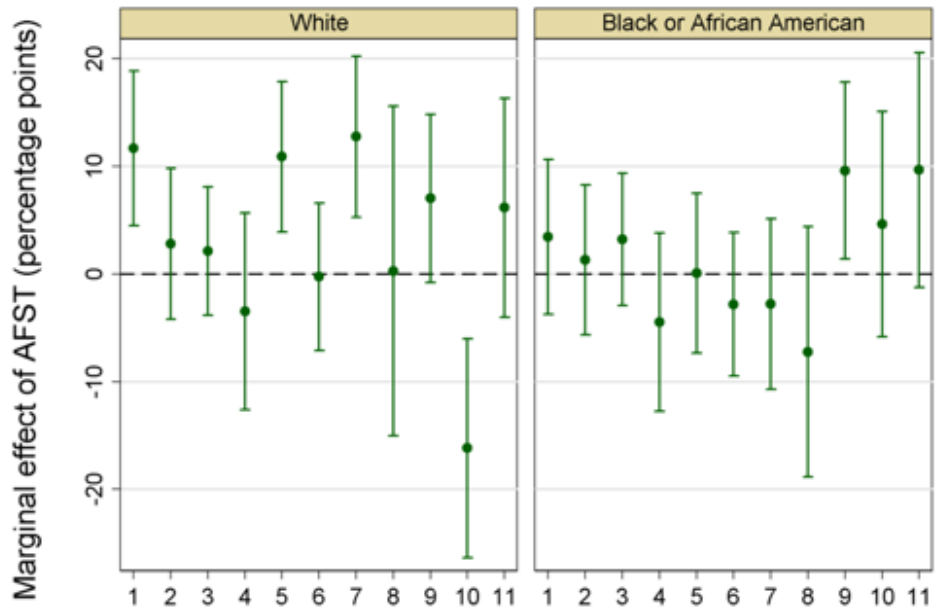


FIGURE 10C: Consistency of workload for 11 call screeners, adjusted analysis, by race



APPENDIX A1: ANALYTIC DATASET AND VARIABLE CONSTRUCTION

Construction of the outcome variables

The screen-in variable is constructed from the Referrals data, using the variable “call_scrn_outcome”. A child was coded as “screen-out” if the call-screen-outcome contained the words “screen” and “out” (after accounting for case-sensitivity). A “screen-in” was any case which was not a “screen-out” and which did not have missing information for the call-screen-outcome. Referrals for the entire post-AFST period (December 2016 – May 2018) were included in this outcome.

A **screen-in with further action upon investigation or a screen-in with no further action and a re-referral within a 2-month window** was constructed for children identified as “screen-in” and for whom a service decision was available (i.e. not missing). “Further action” status was given to children with a service decision other than “Do Not Accept for Service”, regardless of whether the case was connected to an open or closed case. Referrals for a truncated post-AFST period (December 2016–March 2018) and rereferrals for the entire post-AFST period (December 2016–May 2018) were included in this outcome.

A **screen-out with no re-referral within a 2-month window** was constructed for children identified as “screen-out”. For each child, a referral was considered the “index event” if it was not within 60 days of a previous index referral, or if it was the first time a child entered the dataset. A re-referral was any subsequent call within 60 days of the initial referral date, regardless of outcome or service decision dates. Although we account for re-referrals occurring in the months of April and May 2018, we do not include index events occurring as of April 2018. Notably, while index referrals were for GPS calls, re-referral could be for either CPS or GPS calls. Referrals for a truncated post-AFST period (December 2016 – March 2018) and rereferrals for the entire post-AFST period (December 2016 – May 2018) were included in this outcome.

Exclusions in the analytic dataset

All children in all referrals were included in the primary analytic dataset with the following exceptions. Children > 17 years of age at the time of the referral were excluded (although we account for 18-year-old children in re-referral calls). Children in any CPS referral were excluded from the analytic dataset, as there is no variation in screening-decision for these children. Any referral which had a call screen outcome (variable call_scrn_outcome) coded as “Accept: Actively working with this family” was excluded, as were those with call screen outcome coded as “Assessment Completed on Active Family”. The latter two exclusions were at the recommendation of Allegheny County's analysts to perform analyses on data consistent with Allegheny's in-house analyses.

Construction of control variables

Child characteristics include age in years (grouped into categories as under 4, 4–6 years, 7–12 years, 13–17 years), legal sex category (male, female, or undetermined), race category (Black / African American, white, other, unable to determine). A child was considered Black / African

**Appendix
(continued)**

American if their race was coded as such, or any combination of another race and Black / African American and white if the child was coded as “white”, and not a mixed race. All other children fall into either the other category (race was specified, but was not black/African American or white, or unable to determine).

Household characteristics include the composition of household members or the number of other people in the referral who fall between specified age ranges (e.g. <1, 1-5, 6-12, 13-17, adult parents, other adults), the mean age in years of adults in a referral (18 to <30; 30 to <50, 50 to <66, 66+). Household characteristics also included a proxy measure for socioeconomic status. This measure was designed to be consistent with the measure used as an input to the AFST. Specifically, the Zip Code in which the household was located was coded in terms of the fraction of residents falling below the federal poverty line based on the American Community Survey (2008-2012). The constructed socioeconomic status variable has five categories: living in areas where 1) 0 to <10%; 2) 10 to <15%; or 3) 15 to <20%; 4) 20% to <25%; or 5) 25%+ of households fall below the federal poverty line. As an indicator of the risk that any referred child faces, we use a maximum risk score category (low, medium, high, mandatory risk) for the household. Maximum risk score is based on the maximum of the binned risk scores for the placement and the re-referral score, based on cutoffs as determined by Allegheny. The risk score used as a control in regression analyses was not the AFST risk score shown to the call screeners in the Post-AFST Period, but rather the risk score, exactly comparable to that constructed for the Pre-AFST Period.

APPENDIX A2: NOTES ON INTERRUPTED TIME-SERIES (ITSA)

ITSA is estimated as an autoregressive model, to account for the form of correlation between observations. For example, observations which occur within a closer timeframe may be more correlated than observations further apart in time. This type of pattern could reflect secular trends or seasonal patterns. Traditionally, there are two general approaches to account for autocorrelation in ITSA, the autoregressive integrated moving-average (ARIMA) models and ordinary least-squares (OLS), with adjustments for autocorrelation. We utilize the `itsa` command in Stata (v14) which relies on OLS, due to its more flexible and more broadly applicable nature (1–3). We assume that the error term follows an autoregressive process:

$$e_t = \rho e_{t-1} + u_t$$

Where ρ is the correlation between error terms that are adjacent in time and the remaining disturbances, u_t , are independent.

We can specify the maximum number of lags in Stata, as part of the ITSA command and test for the correctness of this specification using `actest` which performs the Cumby-Huizinga general specification test of serial correlation.

Causal inference based on ITSA requires several assumptions:

Assumption 1: Outcomes (levels/trends) remain unchanged in the absences of the program.

Assumption 2: Relative to rapid rate of change in outcomes attributed to the abrupt implementation of the policy of interest, all unobserved time-varying variables change slowly, such that their impact on outcomes would be distinguishable.

Assumption 3: There are no other policies/changes that occur at or around the same time “as the AFST implementation that would impact outcomes substantially.

Assumption 4: Full implementation of the AFST occurs at a discrete point in time.

Assumption 5: The AFST did not materially alter the collection of data on outcomes or covariates or the quality of the data collected.

REFERENCES

1. Box GEP, and Jenkins GM. Time series analysis : forecasting and control. Rev. ed. San Francisco; London: Holden-Day; 1976. xxi, 575 p. p.
2. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal*. 2015;15(2):280-500.
3. Velicer WF, and Harrop J. The reliability and accuracy of time series model identification. *Evaluation Review*. 1983;7:551 - 60.

Appendix
(continued)

APPENDIX TABLES

TABLE A1A: Accuracy of Screen-in, ITSA analysis, all children, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	63.92%	0.000	61.34%	66.49%
Pre-2014 Policy	Trend	-0.58	0.002	-0.93	-0.23
2014 Policy	Change in level	1.38	0.670	-5.13	7.90
Post 2014 policy, pre-AFST	Change in trend	0.53	0.021	0.09	0.98
AFST implementation	Change in level	6.27	0.001	2.75	9.79
Post-AFST	Change in trend	-0.05	0.835	-0.50	0.40
Total trend in screen-in rates pre-AFST		-0.05	0.771	-0.38	0.29
Total trend in screen-in rate post-AFST		-0.10	0.476	-0.36	0.17

TABLE A1B: Accuracy of Screen-in, ITSA analysis, < 4 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	63.86%	0.000	58.98%	68.75%
Pre-2014 Policy	Trend	-0.50	0.050	-1.00	0.00
2014 Policy	Change in level	-2.77	0.452	-10.11	4.58
Post 2014 policy, pre-AFST	Change in trend	0.73	0.031	0.07	1.39
AFST implementation	Change in level	0.18	0.956	-6.33	6.69
Post-AFST	Change in trend	-0.05	0.890	-0.76	0.66
Total trend in screen-in rates pre-AFST		0.23	0.287	-0.20	0.66
Total trend in screen-in rate post-AFST		0.18	0.526	-0.39	0.75

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A1C: Accuracy of Screen-in, ITSA analysis, 4 to 6 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	67.80%	0.000	64.10%	71.50%
Pre-2014 Policy	Trend	-0.97	0.000	-1.40	-0.54
2014 Policy	Change in level	7.33	0.025	0.96	13.70
Post 2014 policy, pre-AFST	Change in trend	0.72	0.006	0.22	1.23
AFST implementation	Change in level	6.55	0.005	2.10	11.01
Post-AFST	Change in trend	0.14	0.587	-0.38	0.67
Total trend in screen-in rates pre-AFST		-0.25	0.077	-0.52	0.03
Total trend in screen-in rate post-AFST		-0.10	0.647	-0.55	0.35

Note: change in trend is expressed in percentage points/month.

TABLE A1D: Accuracy of Screen-in, ITSA analysis, 7 to 12 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	60.47%	0.000	55.70%	65.24%
Pre-2014 Policy	Trend	-0.37	0.210	-0.96	0.22
2014 Policy	Change in level	1.29	0.757	-7.04	9.62
Post 2014 policy, pre-AFST	Change in trend	0.19	0.580	-0.50	0.88
AFST implementation	Change in level	8.77	0.001	3.64	13.89
Post-AFST	Change in trend	0.07	0.855	-0.67	0.80
Total trend in screen-in rates pre-AFST		-0.18	0.319	-0.54	0.18
Total trend in screen-in rate post-AFST		-0.11	0.720	-0.75	0.52

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A1E: Accuracy of Screen-in, ITSA analysis, 13 to 17 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	65.64%	0.000	61.16%	70.13%
Pre-2014 Policy	Trend	-0.64	0.000	-0.98	-0.31
2014 Policy	Change in level	1.25	0.662	-4.48	6.98
Post 2014 policy, pre-AFST	Change in trend	0.65	0.024	0.09	1.21
AFST implementation	Change in level	9.56	0.001	3.92	15.20
Post-AFST	Change in trend	-0.41	0.205	-1.06	0.23
Total trend in screen-in rates pre-AFST		0.01	0.982	-0.44	0.45
Total trend in screen-in rate post-AFST		-0.41	0.087	-0.88	0.06

Note: change in trend is expressed in percentage points/month.

TABLE A1F: Accuracy of Screen-in, ITSA analysis, White, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	47.28%	0.000	41.03%	53.52%
Pre-2014 Policy	Trend	-0.57	0.061	-1.17	0.03
2014 Policy	Change in level	2.33	0.535	-5.17	9.82
Post 2014 policy, pre-AFST	Change in trend	0.57	0.129	-0.17	1.32
AFST implementation	Change in level	10.12	0.006	3.13	17.12
Post-AFST	Change in trend	-0.27	0.414	-0.92	0.39
Total trend in screen-in rates pre-AFST		0.00	0.991	-0.44	0.45
Total trend in screen-in rate post-AFST		-0.27	0.272	-0.75	0.22

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A1G: Accuracy of Screen-in, ITSA analysis, Black/African American, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	59.46%	0.000	53.37%	65.56%
Pre-2014 Policy	Trend	-0.72	0.018	-1.31	-0.13
2014 Policy	Change in level	2.91	0.468	-5.09	10.90
Post 2014 policy, pre-AFST	Change in trend	0.49	0.180	-0.23	1.21
AFST implementation	Change in level	7.71	0.002	3.01	12.40
Post-AFST	Change in trend	-0.36	0.257	-0.99	0.27
Total trend in screen-in rates pre-AFST		-0.23	0.273	-0.65	0.19
Total trend in screen-in rate post-AFST		-0.59	0.016	-1.06	-0.11

Note: change in trend is expressed in percentage points/month.

TABLE A2A: Accuracy of screen-in, adjusted analysis, all children, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	54.96%	0.000	53.85%	56.07%
Post-AFST	58.78%	0.000	57.36%	60.20%
DIFF (Post - Pre)	3.82%	0.000	2.15%	5.49%

TABLE A2B: Accuracy of screen-in, adjusted analysis, by age group, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	54.55%	0.000	52.98%	56.12%
4-6 years	55.70%	0.000	53.85%	57.55%
7-12 years	54.99%	0.000	53.40%	56.57%
13-17 years	54.86%	0.000	53.15%	56.58%
Post-AFST				
< 4 years	56.22%	0.000	54.03%	58.42%
4-6 years	58.79%	0.000	56.59%	60.98%
7-12 years	60.25%	0.000	58.71%	61.79%
13-17 years	59.57%	0.000	57.04%	62.10%
Difference Post-Pre				
< 4 years	1.67%	0.166	-0.69%	4.04%
4-6 years	3.09%	0.019	0.51%	5.68%
7-12 years	5.27%	0.000	3.00%	7.54%
13-17 years	4.71%	0.007	1.29%	8.13%

Appendix
(continued)

TABLE A2C: Accuracy of screen-in, adjusted analysis, by race, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	55.72%	0.000	54.20%	57.25%
Black/African American	56.58%	0.000	54.96%	58.20%
Post-AFST				
White	61.34%	0.000	59.10%	63.58%
Black/African American	58.73%	0.000	56.62%	60.85%
Difference Post - Pre				
White	5.62%	0.000	3.53%	7.70%
Black/African American	2.15%	0.168	-0.91%	5.21%
Difference Black - White				
Pre-AFST	0.86%	0.449	-1.37%	3.08%
Post-AFST	-2.61%	0.105	-5.75%	0.54%

TABLE A3A: Accuracy of screen-out, ITSA analysis, all children, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	72.58%	0.000	68.84%	76.32%
Pre-2014 Policy	Trend	-0.01	0.953	-0.36	0.34
2014 Policy	Change in level	1.27	0.528	-2.75	5.29
Post 2014 policy, pre-AFST	Change in trend	0.10	0.653	-0.33	0.52
AFST implementation	Change in level	-2.19	0.386	-7.25	2.86
Post-AFST	Change in trend	-0.38	0.197	-0.96	0.20
Total trend in screen-in rates pre-AFST		0.09	0.454	-0.15	0.32
Total trend in screen-in rate post-AFST		-0.29	0.266	-0.82	0.23

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A3B: Accuracy of screen-out, ITSA analysis, < 4 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	75.68%	0.000	70.16%	81.21%
Pre-2014 Policy	Trend	-0.16	0.530	-0.65	0.34
2014 Policy	Change in level	0.54	0.867	-5.93	7.01
Post 2014 policy, pre-AFST	Change in trend	0.25	0.505	-0.50	1.01
AFST implementation	Change in level	-1.75	0.746	-12.56	9.06
Post-AFST	Change in trend	-0.27	0.665	-1.54	0.99
Total trend in screen-in rates pre-AFST		0.10	0.736	-0.47	0.66
Total trend in screen-in rate post-AFST		-0.18	0.752	-1.31	0.95

Note: change in trend is expressed in percentage points/month.

TABLE A3C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	70.53%	0.000	66.23%	74.83%
Pre-2014 Policy	Trend	0.18	0.302	-0.17	0.54
2014 Policy	Change in level	0.28	0.906	-4.46	5.02
Post 2014 policy, pre-AFST	Change in trend	-0.07	0.803	-0.61	0.47
AFST implementation	Change in level	-6.28	0.051	-12.59	0.03
Post-AFST	Change in trend	0.19	0.643	-0.63	1.01
Total trend in screen-in rates pre-AFST		0.12	0.559	-0.29	0.52
Total trend in screen-in rate post-AFST		0.31	0.391	-0.41	1.02

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A3D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	69.46%	0.000	65.04%	73.87%
Pre-2014 Policy	Trend	0.19	0.352	-0.21	0.59
2014 Policy	Change in level	-0.30	0.894	-4.89	4.28
Post 2014 policy, pre-AFST	Change in trend	0.00	0.998	-0.45	0.45
AFST implementation	Change in level	-1.68	0.570	-7.58	4.23
Post-AFST	Change in trend	-0.83	0.030	-1.57	-0.09
Total trend in screen-in rates pre-AFST		0.19	0.074	-0.02	0.40
Total trend in screen-in rate post-AFST		-0.64	0.077	-1.35	0.07

Note: change in trend is expressed in percentage points/month.

TABLE A3E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	75.84%	0.000	71.38%	80.31%
Pre-2014 Policy	Trend	-0.36	0.114	-0.80	0.09
2014 Policy	Change in level	5.95	0.054	-0.10	12.00
Post 2014 policy, pre-AFST	Change in trend	0.25	0.366	-0.30	0.79
AFST implementation	Change in level	-0.34	0.856	-4.05	3.38
Post-AFST	Change in trend	-0.16	0.500	-0.64	0.32
Total trend in screen-in rates pre-AFST		-0.11	0.489	-0.42	0.21
Total trend in screen-in rate post-AFST		-0.27	0.139	-0.63	0.09

Note: change in trend is expressed in percentage points/month.

Appendix
(continued)

TABLE A3F: Accuracy of screen-out, ITSA analysis, White, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	71.67%	0.000	66.68%	76.65%
Pre-2014 Policy	Trend	0.09	0.704	-0.40	0.59
2014 Policy	Change in level	-3.93	0.169	-9.61	1.74
Post 2014 policy, pre-AFST	Change in trend	0.32	0.254	-0.24	0.88
AFST implementation	Change in level	-0.80	0.747	-5.79	4.19
Post-AFST	Change in trend	-1.18	0.000	-1.72	-0.65
Total trend in screen-in rates pre-AFST		0.41	0.003	0.15	0.68
Total trend in screen-in rate post-AFST		-0.77	0.002	-1.23	-0.31

Note: change in trend is expressed in percentage points/month.

TABLE A3G: Accuracy of screen-out, ITSA analysis, Black/African American, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > T	[95% CI]	
Start point (August 2013)	Level	71.98%	0.000	65.87%	78.10%
Pre-2014 Policy	Trend	-0.17	0.470	-0.65	0.30
2014 Policy	Change in level	5.57	0.067	-0.41	11.56
Post 2014 policy, pre-AFST	Change in trend	-0.06	0.864	-0.72	0.61
AFST implementation	Change in level	-3.61	0.348	-11.30	4.07
Post-AFST	Change in trend	0.40	0.365	-0.48	1.27
Total trend in screen-in rates pre-AFST		-0.23	0.322	-0.69	0.23
Total trend in screen-in rate post-AFST		0.17	0.650	-0.58	0.91

Note: change in trend is expressed in percentage points/month.

TABLE A4A: Accuracy of screen-out, adjusted analysis, all children, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	73.90%	0.000	73.40%	74.40%
Post-AFST	72.31%	0.000	71.07%	73.54%
DIFF (Post - Pre)	-1.59%	0.014	-2.86%	-0.32%

Appendix
(continued)

TABLE A4B: Accuracy of screen-out, adjusted analysis, by age group, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	73.00%	0.000	71.08%	74.92%
4-6 years	74.33%	0.000	73.12%	75.54%
7-12 years	74.10%	0.000	73.10%	75.09%
13-17 years	73.97%	0.000	72.29%	75.65%
Post-AFST				
< 4 years	72.45%	0.000	69.35%	75.55%
4-6 years	71.76%	0.000	69.46%	74.05%
7-12 years	72.22%	0.000	70.18%	74.27%
13-17 years	72.66%	0.000	70.66%	74.67%
Difference Post-Pre				
< 4 years	-0.55%	0.746	-3.86%	2.77%
4-6 years	-2.57%	0.071	-5.36%	0.22%
7-12 years	-1.88%	0.160	-4.49%	0.74%
13-17 years	-1.31%	0.363	-4.13%	1.51%

TABLE A4C: Accuracy of screen-out, adjusted analysis, by race, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	70.55%	0.000	69.30%	71.81%
Black/African American	74.08%	0.000	72.92%	75.24%
Post-AFST				
White	70.45%	0.000	68.71%	72.18%
Black/African American	70.52%	0.000	68.62%	72.42%
Difference Post-Pre				
White	-0.11%	0.929	-2.41%	2.20%
Black/African American	-3.56%	0.001	-5.67%	-1.44%
Difference Black-White				
Pre-AFST	3.53%	0.001	1.39%	5.66%
Post-AFST	0.08%	0.952	-2.41%	2.57%

Appendix
(continued)

TABLE APPENDIX A5: Regression results for further action or no further action and re-referral within 60 days, conditional on screen-in (Outcome 1: accuracy of screen-in)

VARIABLES	(1)	(2)	(3)
	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE-GROUP
Post-AFST	0.12** [0.03-0.21]	0.29*** [0.16-0.42]	-0.02 [-0.15-0.10]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.30*** [-0.47--0.12]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			0.15* [-0.01-0.30]
Post-AFST x age 7 to 12 years			0.21*** [0.06-0.35]
Post-AFST x age 13 to 17 years			0.21** [0.04-0.38]
Race (comparator is White)			
Black/African American	-0.03 [-0.13-0.07]	0.11* [-0.02-0.24]	-0.03 [-0.13-0.07]
Age-group (age < 4 is comparator)			
age 4 to 6 years	0.10** [0.02-0.17]	0.09** [0.02-0.17]	0.02 [-0.08-0.13]
age 7 to 12 years	0.13*** [0.06-0.20]	0.13*** [0.06-0.19]	0.03 [-0.07-0.13]
age 13 to 17 years	0.11*** [0.03-0.19]	0.11*** [0.03-0.19]	0.01 [-0.11-0.12]
Legal sex (comparator is female)			
Male	-0.04 [-0.09-0.02]	-0.04 [-0.09-0.02]	-0.04 [-0.09-0.02]
HH composition counts			
< 1	-0.23*** [-0.35--0.12]	-0.22*** [-0.34--0.11]	-0.23*** [-0.35--0.12]
1 to 5 years	0.07** [0.01-0.13]	0.07** [0.01-0.13]	0.07** [0.01-0.13]
6 to 12 years	0.04* [-0.01-0.09]	0.05* [-0.01-0.10]	0.04* [-0.01-0.09]
13 to 17 years	-0.08** [-0.15--0.02]	-0.08** [-0.15--0.02]	-0.08** [-0.15--0.02]
Parents	-0.00 [-0.06-0.05]	-0.00 [-0.05-0.05]	-0.00 [-0.06-0.05]
Other adults	-0.03 [-0.09-0.03]	-0.03 [-0.09-0.03]	-0.03 [-0.09-0.03]

Appendix
(continued)

VARIABLES	(1)	(2)	(3)
	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE-GROUP
Mean age of all adults in referral (comparator is no adult age reported)			
18–29 years	-0.30 [-0.73–0.13]	-0.30 [-0.74–0.13]	-0.30 [-0.73–0.13]
30–49 years	-0.34 [-0.76–0.08]	-0.34 [-0.77–0.08]	-0.34 [-0.76–0.09]
50–65 years	-0.44* [-0.89–0.01]	-0.44* [-0.90–0.01]	-0.44* [-0.89–0.01]
66/max years	-0.72* [-1.46–0.03]	-0.74* [-1.48–0.00]	-0.71* [-1.46–0.04]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	-1.25*** [-1.56–-0.94]	-1.26*** [-1.57–-0.94]	-1.25*** [-1.56–-0.93]
Poor	-1.22*** [-1.53–-0.92]	-1.22*** [-1.53–-0.92]	-1.22*** [-1.52–-0.91]
Mid	-1.09*** [-1.42–-0.76]	-1.09*** [-1.42–-0.76]	-1.09*** [-1.41–-0.76]
Wealthier	-1.15*** [-1.46–-0.84]	-1.15*** [-1.46–-0.84]	-1.14*** [-1.46–-0.83]
Wealthiest	-1.12*** [-1.44–-0.81]	-1.13*** [-1.44–-0.81]	-1.12*** [-1.43–-0.81]
Risk score (historical plus projected, comparator is no risk score)			
Low	-2.46*** [-3.41–-1.51]	-2.45*** [-3.39–-1.51]	-2.46*** [-3.41–-1.50]
Middle	-1.68*** [-2.62–-0.74]	-1.67*** [-2.60–-0.74]	-1.68*** [-2.62–-0.74]
High	-0.96** [-1.89–-0.03]	-0.95** [-1.88–-0.03]	-0.96** [-1.90–-0.02]
Mandatory	-0.04 [-0.98–0.89]	-0.03 [-0.96–0.89]	-0.04 [-0.98–0.90]
Observations	26,010	26,010	26,010

Appendix
(continued)TABLE APPENDIX A6: Regression results for no re-referral within 60 days, conditional on screen-out
(Outcome 2: accuracy of screen-out)

VARIABLES	(1)	(2)	(3)
	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Post-AFST	-0.09* [-0.19-0.01]	-0.07 [-0.21-0.07]	-0.10 [-0.32-0.11]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.07 [-0.29-0.15]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			-0.07 [-0.30-0.16]
Post-AFST x age 7 to 12 years			0.00 [-0.29-0.30]
Post-AFST x age 13 to 17 years			0.09 [-0.15-0.32]
Race (comparator is White)			
Black/African American	0.16*** [0.04-0.28]	0.20*** [0.06-0.34]	0.16*** [0.04-0.28]
Other race	0.14* [-0.02-0.30]		0.14* [-0.02-0.30]
Unable to determine race	0.97*** [0.62-1.33]		0.98*** [0.62-.33]
Age-group (age < 4 is comparator)			
age 4 to 6 years	-0.07 [-0.17-0.03]	-0.05 [-0.15-0.05]	-0.04 [-0.17-0.09]
age 7 to 12 years	-0.05 [-0.15-0.05]	-0.03 [-0.13-0.06]	-0.05 [-0.22-0.12]
age 13 to 17 years	-0.08 [-0.20-0.04]	-0.06 [-0.18-0.06]	-0.13* [-0.27-0.01]
Legal sex (comparator is female)			
Male	0.03 [-0.06-0.12]	0.03 [-0.06-0.11]	0.03 [-0.06-0.12]
HH composition counts			
< 1	0.01 [-0.17-0.18]	-0.02 [-0.21-0.17]	0.01 [-0.17-0.19]
1 to 5 years	-0.01 [-0.09-0.07]	-0.03 [-0.11-0.05]	-0.01 [-0.09-0.07]
6 to 12 years	-0.07*** [-0.12--0.02]	-0.07** [-0.13--0.01]	-0.07*** [-0.12--0.02]
13 to 17 years	0.02 [-0.05-0.08]	0.01 [-0.05-0.07]	0.02 [-0.05-0.08]

Appendix
(continued)

VARIABLES	(1)	(2)	(3)
	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Parents	-0.03 [-0.08-0.03]	-0.02 [-0.08-0.03]	-0.03 [-0.08-0.03]
Other adults	-0.02 [-0.08-0.04]	-0.02 [-0.08-0.04]	-0.02 [-0.08-0.04]
Mean age of all adults in referral (comparator is no adult age reported)			
18 - 29 years	-0.83*** [-1.23--0.44]	-0.82*** [-1.24--0.40]	-0.84*** [-1.23--0.44]
30 - 49 years	-0.76*** [-1.12--0.39]	-0.73*** [-1.12--0.34]	-0.76*** [-1.12--0.39]
50 - 65 years	-0.61*** [-1.00--0.22]	-0.61*** [-1.02--0.19]	-0.61*** [-1.00--0.23]
66/max years	0.02 [-0.68-0.72]	-0.09 [-0.80-0.61]	0.02 [-0.68-0.72]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	-0.22 [-0.53-0.08]	-0.17 [-0.48-0.14]	-0.22 [-0.53-0.08]
Poor	-0.30** [-0.60--0.00]	-0.25 [-0.56-0.06]	-0.30** [-0.60--0.00]
Mid	-0.34** [-0.60--0.08]	-0.30** [-0.57--0.02]	-0.34** [-0.60--0.08]
Wealthier	-0.39*** [-0.64--0.13]	-0.32** [-0.58--0.06]	-0.39*** [-0.64--0.13]
Wealthiest	-0.36*** [-0.61--0.11]	-0.29** [-0.56--0.03]	-0.36*** [-0.61--0.11]
Risk score (historical plus projected, comparator is no risk score)			
Low	0.41 [-0.33-1.15]	0.37 [-0.35-1.09]	0.42 [-0.32-1.15]
Middle	0.10 [-0.65-0.85]	0.08 [-0.66-0.82]	0.11 [-0.64-0.86]
High	-0.39 [-1.13-0.36]	-0.38 [-1.11-0.35]	-0.38 [-1.12-0.36]
Mandatory	-0.38 [-1.07-0.31]	-0.38 [-1.06-0.30]	-0.38 [-1.07-0.31]
Observations	28,957	25,740	28,957

All regression results are based on a GLM model, with standard errors clustered at the screener-level

Appendix
(continued)

TABLE APPENDIX A7: Regression results for screen-in (Outcome 3: workload)

	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Post-AFST	-0.13 [-0.30-0.03]	-0.06 [-0.24-0.13]	-0.06 [-0.19-0.08]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.13* [-0.27-0.01]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			0.08 [-0.04-0.19]
Post-AFST x age 7 to 12 years			-0.07 [-0.20-0.06]
Post-AFST x age 13 to 17 years			-0.25*** [-0.40--0.10]
Race (comparator is White)			
Black/African American	0.01 [-0.06-0.09]	0.07 [-0.05-0.20]	0.01 [-0.06-0.08]
Other race	0.06 [-0.04-0.15]		0.06 [-0.03-0.16]
Unable to determine race	0.12*** [0.03-0.22]		0.12** [0.02-0.21]
Age-group (age < 4 is comparator)			
age 4 to 6 years	-0.07** [-0.14--0.01]	-0.07** [-0.13--0.01]	-0.11*** [-0.18--0.04]
age 7 to 12 years	-0.06** [-0.11--0.00]	-0.05* [-0.11-0.01]	-0.02 [-0.10-0.05]
age 13 to 17 years	-0.25*** [-0.32--0.18]	-0.26*** [-0.32--0.19]	-0.13** [-0.23--0.02]
Legal sex (comparator is female)			
Male	-0.02 [-0.06-0.01]	-0.03* [-0.06-0.00]	-0.02 [-0.06-0.01]
HH composition counts			
< 1	0.65*** [0.55-0.76]	0.66*** [0.54-0.78]	0.65*** [0.55-0.76]
1 to 5 years	0.10*** [0.05-0.15]	0.10*** [0.04-0.15]	0.10*** [0.05-0.15]
6 to 12 years	0.12*** [0.08-0.17]	0.12*** [0.08-0.17]	0.12*** [0.08-0.17]
13 to 17 years	-0.03 [-0.08-0.02]	-0.03 [-0.08-0.02]	-0.03 [-0.08-0.02]
Parents	-0.04 [-0.08-0.01]	-0.03 [-0.08-0.02]	-0.04 [-0.08-0.01]

Appendix
(continued)

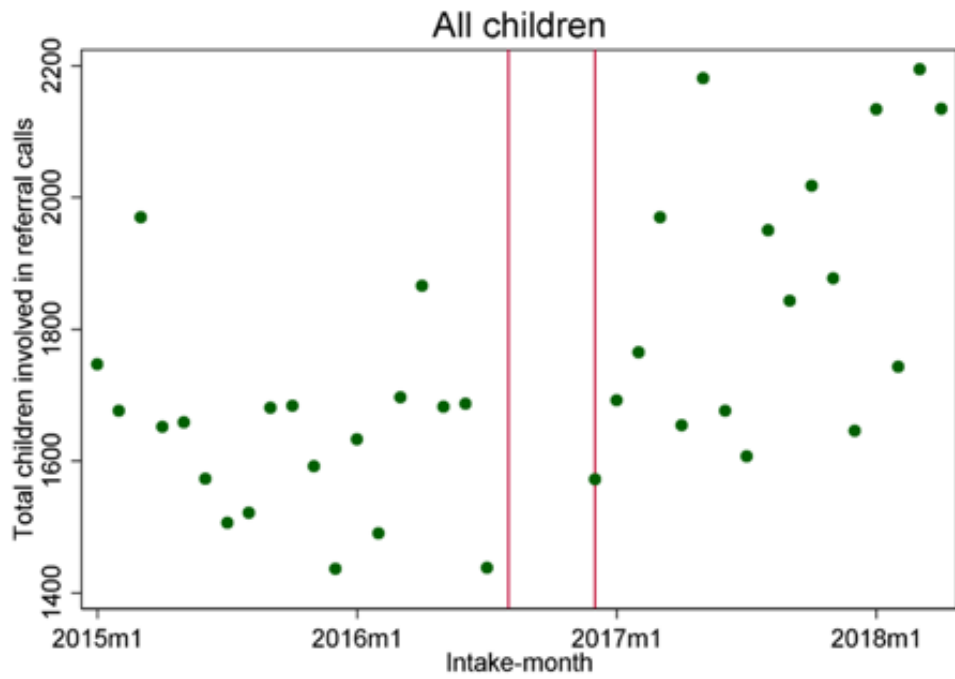
VARIABLES	(1)	(2)	(3)
	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Other adults	0.09*** [0.06-0.13]	0.10*** [0.06-0.14]	0.09*** [0.06-0.13]
Mean age of all adults in referral (comparator is no adult age reported)			
18-29 years	1.08*** [0.90-1.26]	1.02*** [0.84-1.21]	1.08*** [0.90-1.26]
30-49 years	1.08*** [0.90-1.27]	1.01*** [0.82-1.20]	1.09*** [0.90-1.27]
50-65 years	1.15*** [0.90-1.40]	1.09*** [0.82-1.36]	1.15*** [0.90-1.40]
66/max years	1.05*** [0.56-1.55]	1.12*** [0.64-1.59]	1.05*** [0.56-1.55]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	0.57** [0.04-1.10]	0.55** [0.00-1.09]	0.57** [0.04-1.10]
Poor	0.66** [0.14-1.18]	0.62** [0.09-1.16]	0.66** [0.14-1.18]
Mid	0.65*** [0.17-1.14]	0.59** [0.09-1.09]	0.65*** [0.17-1.14]
Wealthier	0.73*** [0.22-1.24]	0.69** [0.16-1.21]	0.72*** [0.22-1.23]
Wealthiest	0.66*** [0.17-1.16]	0.61** [0.09-1.12]	0.66*** [0.16-1.15]
Risk score (historical plus projected, comparator is no risk score)			
Low	-0.05 [-0.57-0.47]	-0.05 [-0.58-0.48]	-0.06 [-0.58-0.46]
Middle	0.53* [-0.01-1.06]	0.51* [-0.03-1.05]	0.52* [-0.02-1.06]
High	0.93*** [0.38-1.47]	0.92*** [0.38-1.47]	0.92*** [0.37-1.46]
Mandatory	2.19*** [1.62-2.75]	2.18*** [1.61-2.75]	2.18*** [1.61-2.74]
Observations	60,287	54,388	60,287

All regression results are based on a GLM model, with standard errors clustered at the screener-level

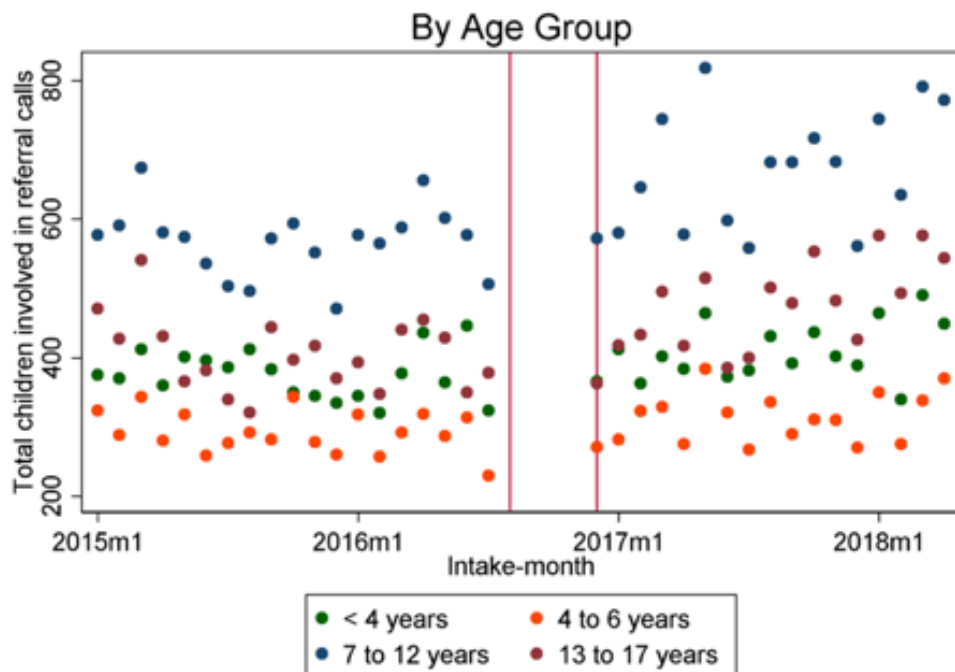
Appendix
(continued)

APPENDIX FIGURES

APPENDIX FIGURE 1A: Total children in referral calls, by month

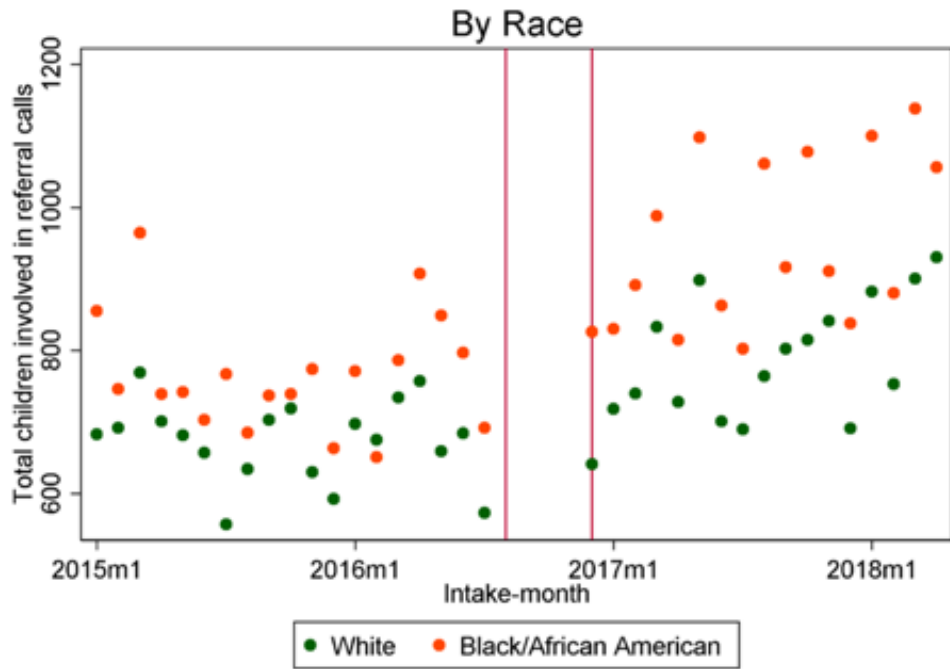


APPENDIX FIGURE 1B: Total children in referral calls, by month and age-group

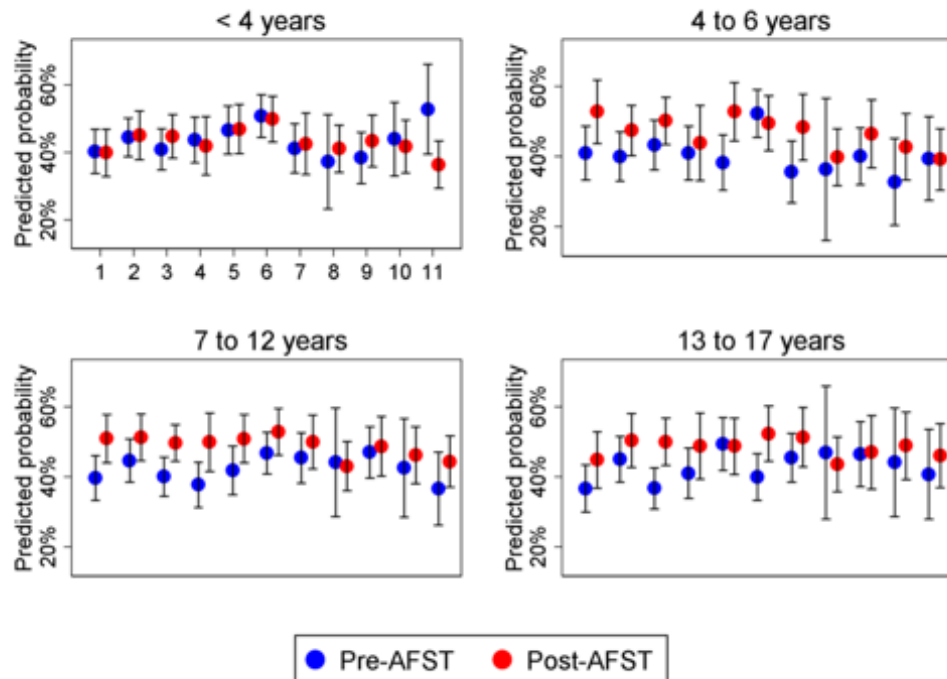


Appendix
(continued)

APPENDIX FIGURE 1C: Total children in referral calls, by month and race

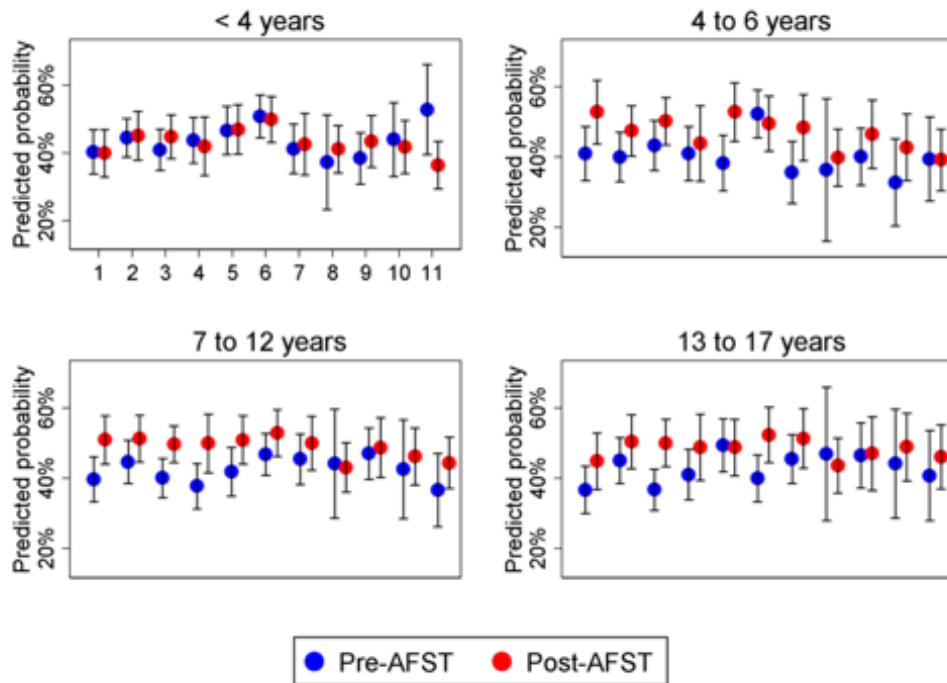


APPENDIX FIGURE 2A: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis

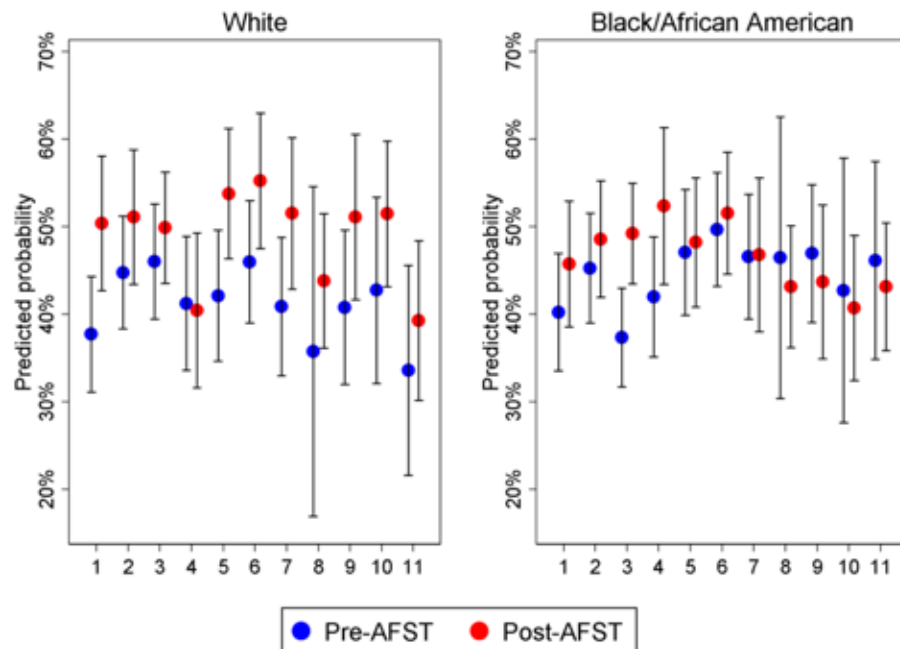


Appendix
(continued)

APPENDIX FIGURE 2B: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by age-group

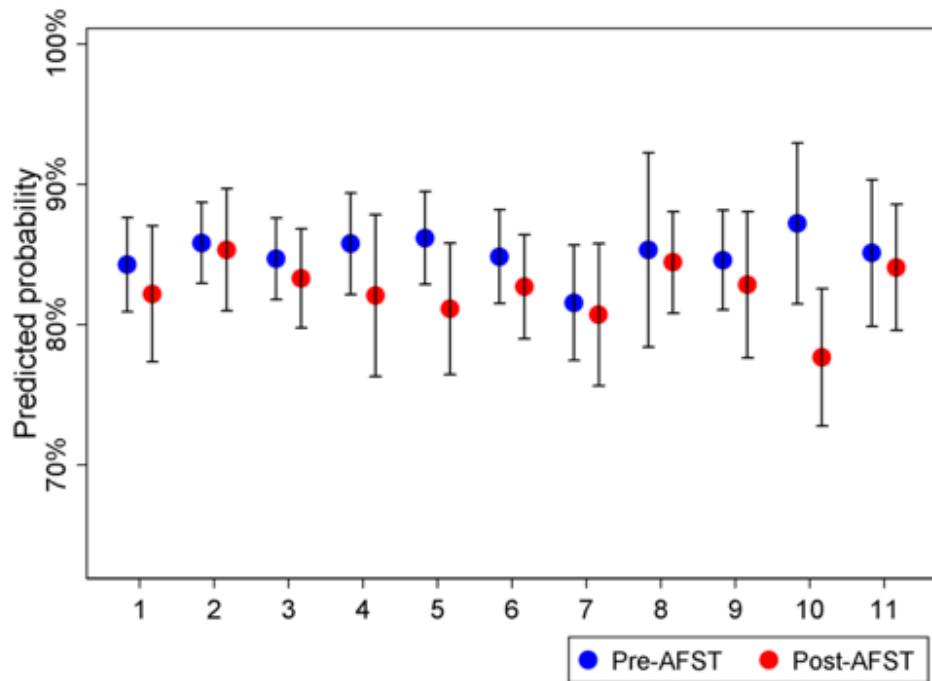


APPENDIX FIGURE 2C: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by race

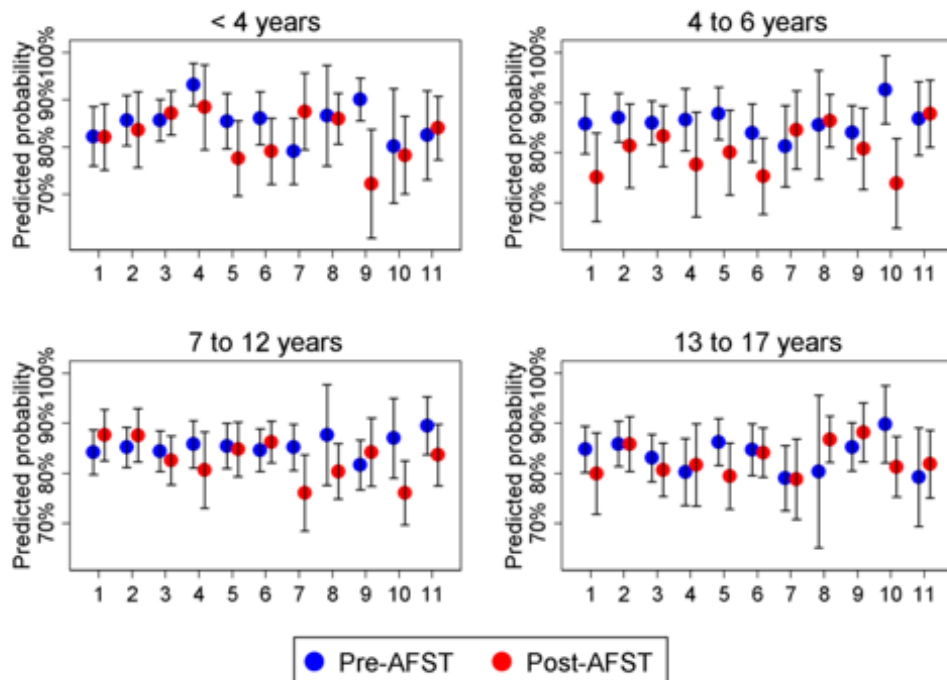


Appendix
(continued)

APPENDIX FIGURE 3A: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis

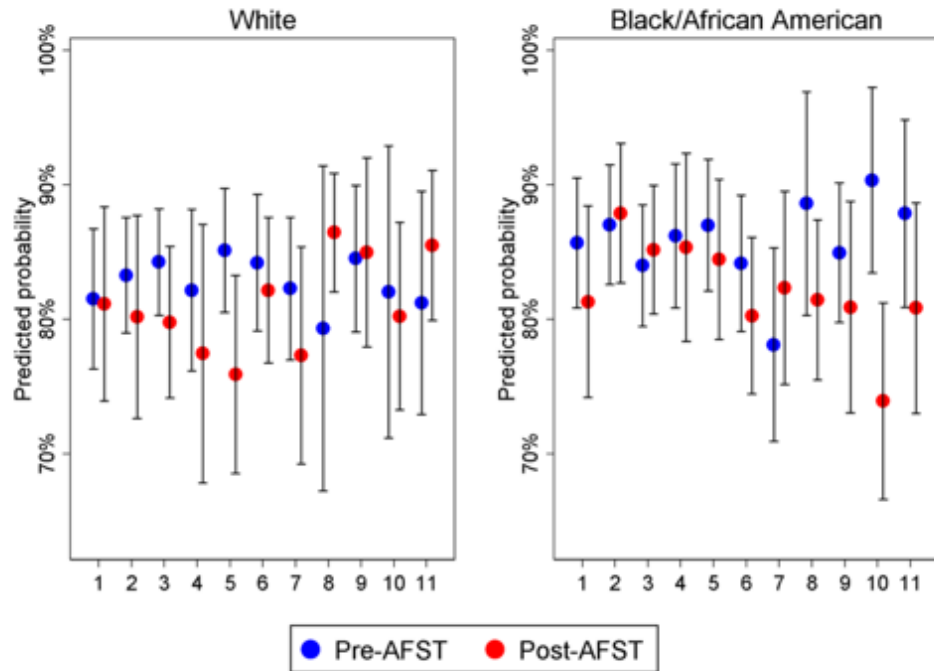


APPENDIX FIGURE 3B: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by age-group

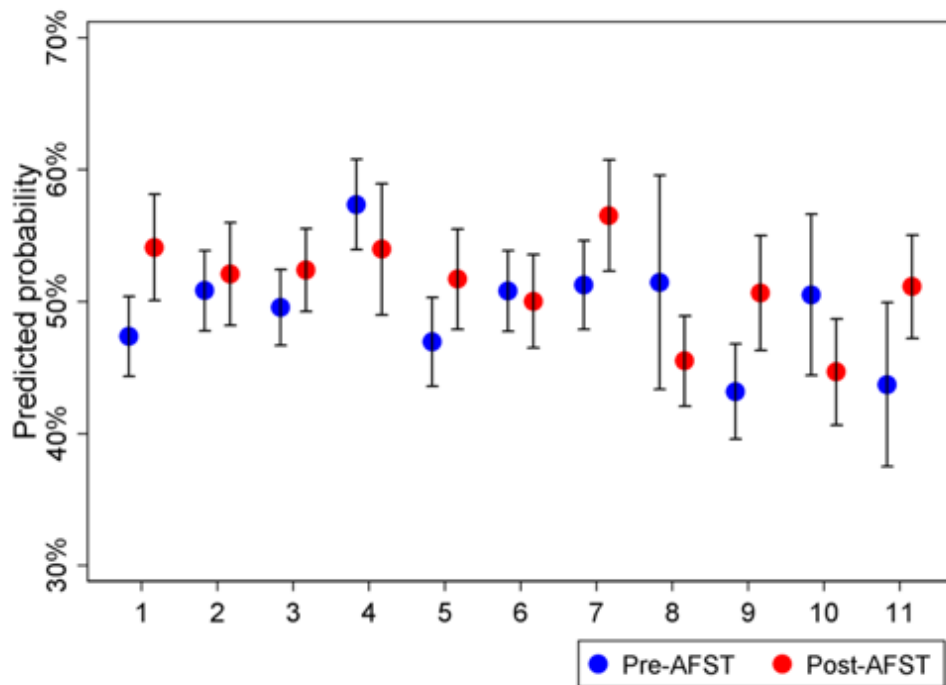


Appendix
(continued)

APPENDIX FIGURE 3C: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by race

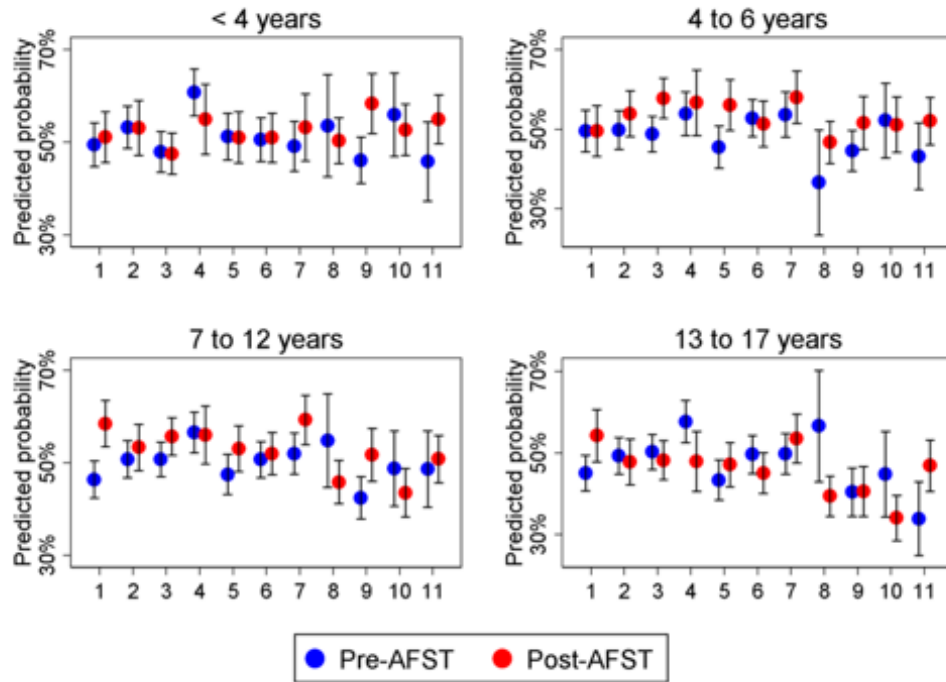


APPENDIX FIGURE 4A: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis



Appendix
(continued)

APPENDIX FIGURE 4B: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by age-group



APPENDIX FIGURE 4C: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by race

