Extended Analysis of "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions"

Logan Stapleton*† Hao-Fei Cheng‡ Anna Kawakami‡ Venkatesh Sivaraman‡
Yanghuidi Cheng‡ Diana Qing§ Adam Perer‡ Kenneth Holstein‡
Zhiwei Steven Wu‡ Haiyi Zhu‡

April 29, 2022

Abstract

This is an extended analysis of our paper "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions" [1], which looks at racial disparities in the Allegheny Family Screening Tool [5], an algorithm used to help child welfare workers decide which families the Allegheny County child welfare agency (CYF) should investigate. On April 27, 2022, Allegheny County CYF sent us an updated dataset and pre-processing steps. In this extended analysis of our paper, we show the results from re-running all quantitative analyses in our paper with this new data and pre-processing. We find that our main findings in [1] were robust to changes in data and pre-processing. Particularly, the Allegheny Family Screening Tool on its own would have made more racially disparate decisions than workers, and workers used the tool to decrease those algorithmic disparities. Some minor results changed, including a slight increase in the screen-in rate from before to after the implementation of the AFST reported our paper [1].

1 Data & Pre-processing Changes

Our paper "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions" [1] looks at racial disparities in the Allegheny Family Screening Tool [5], an algorithm used to aid child welfare workers deciding which families the county government should investigate. The paper uses a mixed methods approach, reporting both quantitative and qualitative findings. The quantitative findings are based on real, anonymized data on children who were reported to the Allegheny County Office of Children, Youth, and Families (CYF), the AFST scores that these children received, whether the children were screened in for investigation or not, and whether the children were placed in foster care.

We retrieved the data within a secure file location in Carnegie Mellon University used to store sensitive data received from Allegheny County CYF, which is maintained by colleagues of ours. The primary data we used in our paper was stored in files named PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv and PAN_Retro_Run_08012016_07132018.csv. At the time of

*Corresponding author: stapl158@umn.edu
†University of Minnesota
‡Carnegie Mellon University
§University of California, Berkeley

writing our paper, we believed this data to be the same as those used in an CYF-commissioned Impact Evaluation of the AFST done by Stanford researchers [3]. As such, we preprocessed the data in the same way that Goldhaber-Fiebert and Prince [3] describe in their paper: First, we excluded all referrals not marked General Protective Services (GPS) by deleting all data entries where the variable REFER_TYPE_GPS_NULL was not 1. Second, we excluded referrals connected to active and closed cases by deleting entries where the variable CALL_SCRN_OUTCOME was either 'Accept: Actively working with this family' or 'Screen Out: **Assessment Completed on Active Family**' in a secondary dataset named LIM_REFERRAL_CLIENTS_08262021.csv. We exclude referrals which the AFST would not have had an impact on: non-GPS referrals and referrals that are connected to active cases are automatically screened in by workers, so the AFST would have had no impact (because workers had no choice).

On April 19, 2022, we received notice from employees at Allegheny County CYF that they have been using a different dataset and pre-processing steps for their internal analysis. They recommended that we not use the PAN_Retro files (referred to above) to define the population of referrals, the CALL_SCRN_OUTCOME variable to exclude active cases, nor the MCI_ID variable we used to identify each individual child. Particularly, they said the MCI_ID variable contained some errors, where one child could be assigned multiple children. Instead, they recommended that we use a new variable called MCI_UNIQ_ID, which is similar to MCI_ID but removes many of these erroneous redundancies, to identify individual children; and that we use a new variable called ACTIVE_FAMILY_IND to identify active cases. Neither MCI_UNIQ_ID nor ACTIVE_FAMILY_IND were available to us at the time we wrote our paper. On April 21, 2022 we had an online video meeting with a member from the CYF Child Welfare Analytics team and received confirmation on the new pre-processing steps we would take to re-run our analysis. On April 27, 2022, a CYF employee sent us a new dataset LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv which includes information on all referrals from January 1, 2015 through July 13, 2018, including the ACTIVE_FAMILY_IND and MCI_UNIQ_ID variables. They also sent us a file called RETRO_FILES_COMBINED_04212022_v3.csv which contains retroactively-run AFST scores (which corrected for a glitch in AFST V1 scores [2]) for children reported during this time period. Following CYF's recommendations to use this new data they sent us, as well as which pre-processing steps to use, we re-ran all of the quantitative analyses in our paper. For clarity, we will refer to the data, pre-processing, and quantitative analysis in our original paper as **Analysis 1** and this second analysis based on new data and pre-processing as **Analysis 2**.[1]

2 Main Takeaways

Although all of the specific numbers we reported in Analysis 1 changed in Analysis 2, our primary conclusions remain the same. Our primary quantitative results are almost identical under both Analyses 1 and 2, even with changes in the data and pre-processing. This lends support to the robustness of our findings that: 1) **The AFST on its own was more racially disparate than workers, both in terms of screen-in rate and accuracy**; and 2) that **workers were able to reduce this disparity in the algorithm.** This second result is interesting and surprising, given that child welfare workers are known to make racially disparate decisions without algorithms [4]. See Figure 1 for a comparison of screen-in rate disparities across Analyses 1 and 2. See Table 1 for a full comparison of disparities by risk level.

The two Analyses on disparity in accuracy show similar patterns: AFST-only decisions had higher racial disparity in accuracy than worker-AFST decisions (see Figure 2a and 2b). However, one

[1] All of our code is publicly available at `https://github.com/logan-stapleton/AFST_racial_disparity`.

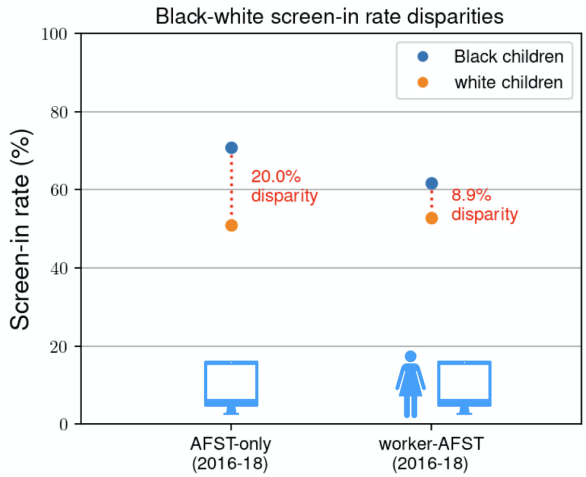
aspect of the accuracy comparison that we explicitly mentioned in the paper changed between Analyses: In Analysis 1, we found that overall AFST-only decisions were more accurate than worker-AFST decisions (by 4.5%). In Analysis 2, we found that overall AFST-only decisions were less accurate than worker-AFST decisions (by 2%). See Figure 2 on page 4. This changes our interpretation of these findings from the original paper. We no longer believe that we can draw clear conclusions based upon the current analyses, about whether the AFST is more accurate than workers or vice versa, even when evaluating performance on the predictive targets used by the AFST itself. It is notable that the results of this comparison changed between Analysis 1 and 2: this highlights the sensitivity of claims about overall accuracy in the AFST to changes in data and pre-processing. However, we **reiterate the arguments presented in our paper that there are additional reasons to be cautious when interpreting and reporting quantitative results about accuracy and comparisons of accuracy** because of (1) **the inherent challenges of accuracy measurement in risk assessments for which the predictions affect the predictive targets** (see Section 4.2 of our paper [1]), and (2) our qualitative finding that workers and the AFST do not agree on prediction outcomes and accuracy measures, meaning that **evaluating human workers' performance on the AFST's benchmarks alone is akin to evaluating a player's performance at a game they are not actually playing** (see Section 6.2 of our paper [1]).

Another quantitative result (**that we did not highlight in the paper due to concerns around confounding factors**) was that the actual screen-in rate increased from before to after the deployment of the AFST. In Analysis 1, we found that, pre-AFST, 52.5% of Black children and 41.2% of white children in discretionary referrals were screened in; and post-AFST, 61.8% of Black children and 52.8% of white children were screened in. In Analysis 2, we found that, pre-AFST, 50.5% of Black children and 42.7% of white children in discretionary referrals were screened in; and post-AFST, 50.2% of Black children and 43.1% of white children were screened in. Thus, following Analysis 1, we calculated that the screen-in rate increased about 10% from pre- to post-AFST, whereas in Analysis 2 there was no noticeable change in the screen-in rate. This indicates that claims about the overall screen-in rate may be more sensitive to changes in data and pre-processing (the design of exclusion criteria).

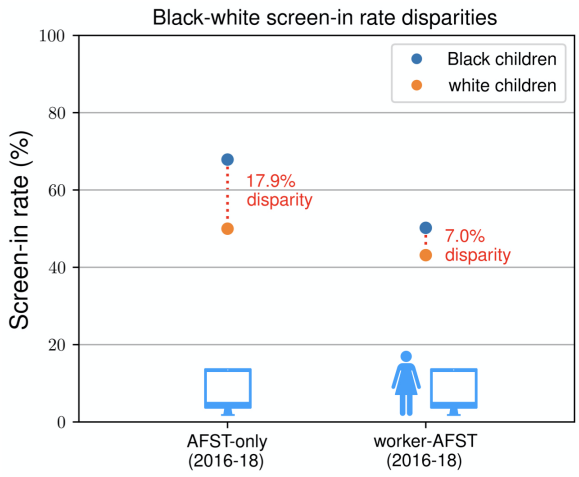
Summary
Our main findings were robust to changes in data and pre-processing. Some results changed. However, as these results were less novel or important, we believe they do not significantly alter the primary contributions of our original paper. See the appendix below for a full comparison of Analyses 1 and 2.

	Analysis 1			**Analysis 2**		
Risk level	**All children**	**Black children**	**white children**	**All children**	**Black children**	**white children**
All	51750	26123 (50.5%)	21623 (41.8%)	35697	17292 (48.4%)	15596 (43.7%)
Mandatory	15182 (29.3%)	9639 (36.9%)	4863 (22.5%)	9431 (26.4%)	5618 (32.5%)	3242 (20.8%)
High	31022 (59.9%)	18536 (71.0%)	11013 (50.9%)	20694 (58.0%)	11737 (67.9%)	7797 (50.0%)
Medium	11778 (22.8%)	5208 (19.9%)	5653 (26.1%)	8391 (23.5%)	3778 (21.8%)	4025 (25.8%)
Low	8950 (17.3%)	2379 (9.1%)	4957 (22.9%)	6612 (18.5%)	1777 (10.3%)	3774 (24.2%)

Table 1: Numbers and proportions of (post-AFST) children by risk level and race between Analyses 1 and 2. Percentages are over total children by race, e.g. 1777 Black children labeled Low risk made up 10.3% of all 17292 Black children referred to CYF from 8/1/16 to 5/13/18.

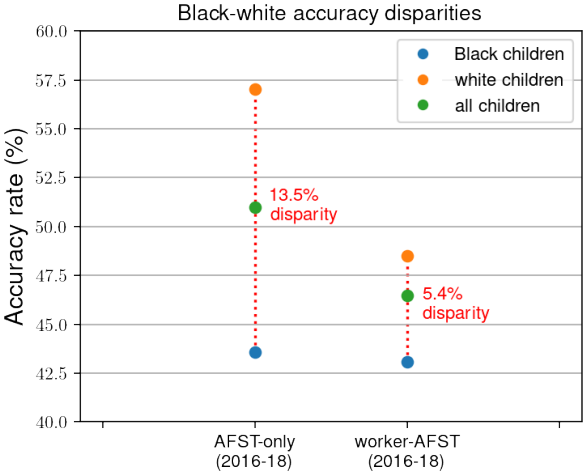


(a) Screen-in rate disparity based on Analysis 1

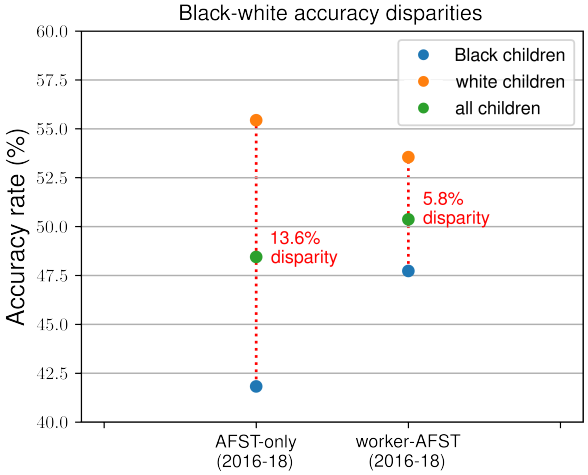


(b) Screen-in rate disparity based on Analysis 2

Figure 1: Comparison of screen-in rate disparities between Analyses 1 and 2



(a) Accuracy rate and disparity based on Analysis 1



(b) Accuracy rate and disparity based on Analysis 2

Figure 2: Comparison of accuracy rates and disparities between Analyses 1 and 2

References

[1] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* `https://dl.acm.org/doi/abs/10.1145/3491102.3501831`

[2] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–12.

[3] Jeremy D Goldhaber-Fiebert and Lea Prince. 2019. Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office. *Pittsburgh: Allegheny County* (2019).

[4] Hyunil Kim, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. 2017. Lifetime

Prevalence of Investigating Child Maltreatment Among US Children. *American Journal of Public Health* 107, 2 (2017), 274–280. https://doi.org/10.2105/AJPH.2016.303545

[5] Rhema Vaithianathan, Nan Jiang, Tim Maloney, Parma Nand, and Emily Putnam-Hornstein. 2017. Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf

A Detailed Comparisons of Analyses 1 and 2

Analysis 1	Analysis 2
• Fully automated AFST-only decisions would have screened in 71.0% of Black children and 51.0% of white children in discretionary referrals from 2016-18, a racial disparity of 20%. Over the same time period, workers using the AFST screened in 61.8% of Black children and 52.8% of white children, a disparity of 9%. **Thus, worker-AFST decisions reduced disparities in the AFST by 11%.**	• AFST-only decisions would have screened in 67.9% of Black children and 50.0% of white children in discretionary referrals from 2016-18, a racial disparity of 17.9%. Over the same time period, workers using the AFST screened in 50.2% of Black children and 43.1% of white children, a disparity of 7.1%. **Thus, worker-AFST decisions reduced disparities in the AFST by 10.8%.**
• AFST-only decisions would have had a 13.5% Black-white disparity in terms of accuracy, compared to 5.4% for worker-AFST decisions. Thus, **worker-AFST were less disparate than the AFST on its own in terms of accuracy**.	• AFST-only decisions would have had a 13.6% Black-white disparity in terms of accuracy, compared to 5.8% for worker-AFST decisions. Thus, **worker-AFST were less disparate than the AFST on its own in terms of accuracy**.
• AFST-only decisions would have been accurate for 51.0% of all children, versus 46.5% for worker-AFST decisions. Thus, **the AFST on its own would have been more accurate than workers.**	• AFST-only decisions would have been accurate for 48.4% of all children, versus 50.4% for worker-AFST decisions. Thus, **the AFST on its own would have been slightly less accurate than workers.**

Table 2: Comparison of quantitative findings between Analyses 1 and 2

	Analysis 1	Analysis 2
Population data	PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv, PAN_Retro_Run_08012016_07132018.csv	LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv
Exclusions	LIM_REFERRAL_CLIENTS_08262021.csv	LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv
Placements	PLACEMENTS_DEID_FULL_07232020.csv	LIM_PLACEMENTS_08262021.csv
AFST scores	PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv, PAN_Retro_Run_08012016_07132018.csv, RR_scorecutoffs.csv, PL_scorecutoffs.csv	RETRO_FILES_COMBINED_04212022_v3.csv

Table 3: Comparison of data used between Analyses 1 and 2

Analysis 1	Analysis 2
• Non-GPS referrals (where REFER_TYPE_GPS_NULL!=1 in PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv) • Referrals connected to completed cases (CALL_SCRN_OUTCOME = 'Screen Out: **Assessment Completed on Active Family**' in LIM_REFERRAL_CLIENTS_08262021.csv) • Referrals connected to active cases (CALL_SCRN_OUTCOME = 'Accept: Actively working with this family' in LIM_REFERRAL_CLIENTS_08262021.csv)	In LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv • Non-GPS referrals (REF_TYPE ≠ 'GPS') • Active referrals (ACTIVE_FAMILY_IND = 1) • Referrals from truancy court (TRUANCY_ONLY_COURTS_REF = 1) • Non-children (ALL_CHILD ≠ 1) • Intake didn't screen (CALL_SCRN_CODE = -9)

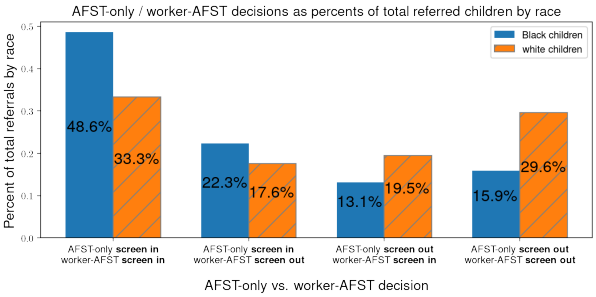
Table 4: Comparison of data exclusions steps between Analyses 1 and 2

Analysis 1	Analysis 2
• white: 'RACE_WHITE_NULL' = 1 and 'RACE_BLACK_NULL', 'RACE_HISPANIC_NULL', 'RACE_NATIVE_NULL', 'RACE_ASIAN_NULL', 'RACE_OTHER_NULL', 'RACE_UNKNOWN_NULL' = 0 in PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv • Black: 'RACE_BLACK_NULL' = 1 in PAN_Retro_Run_Referrals_for_2014-2016_provided_2018-08-29.csv	• white: RACE = 'White' in LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv • Black: RACE contains 'Black or African American' in LIM_REFERRAL_CLIENTS_UNIQapp_04212022_v3.csv

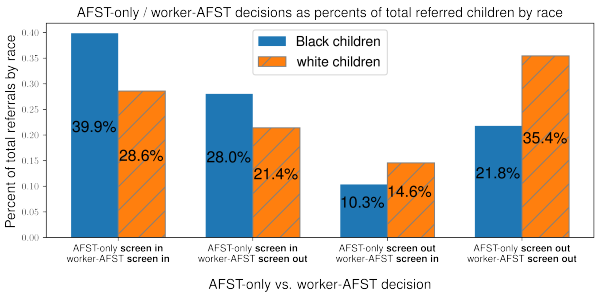
Table 5: Comparison of race (Black and white) coding between Analyses 1 and 2

B Comparisons of all figures in the paper [1] based on Analyses 1 and 2

Figure 3: Comparison of worker-AFST compliance rates (Figure 4 in the paper [1]) based on Analyses 1 and 2

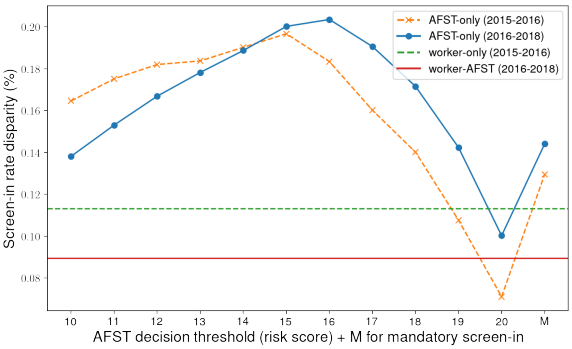


(a) Figure 4 in the paper [1] based on Analysis 1

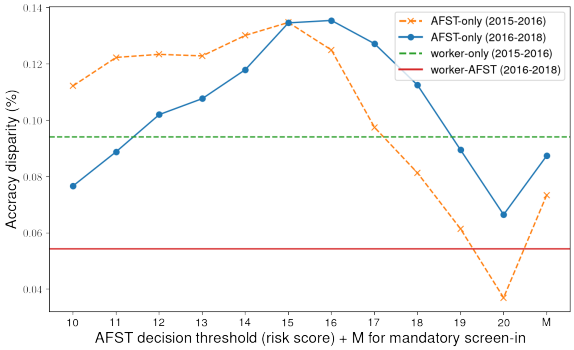


(b) Figure 4 in the paper [1] based on Analysis 2

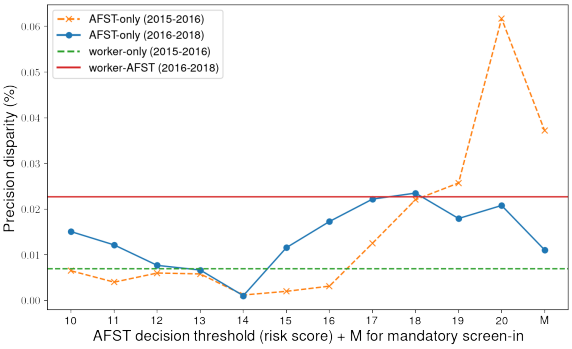
Figure 4: Comparisons of common group fairness metrics (Figure 6 in the paper [1]) based on Analysis 1



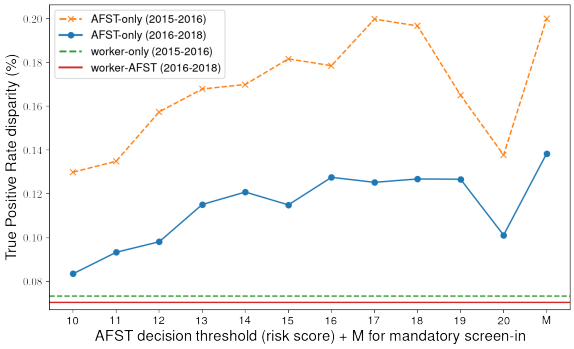
(a) Screen-in rate disparity



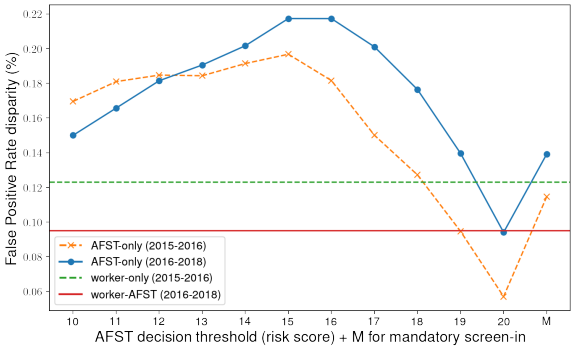
(b) Accuracy disparity



(c) Precision rate disparity

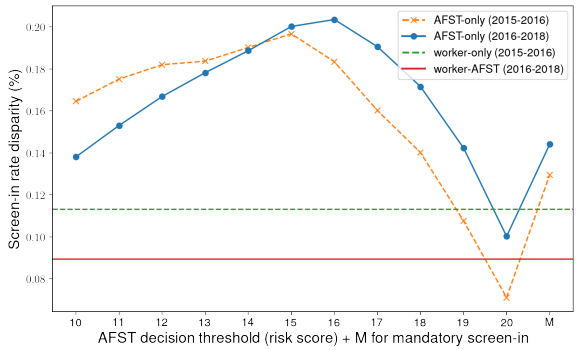


(d) True positive rate disparity

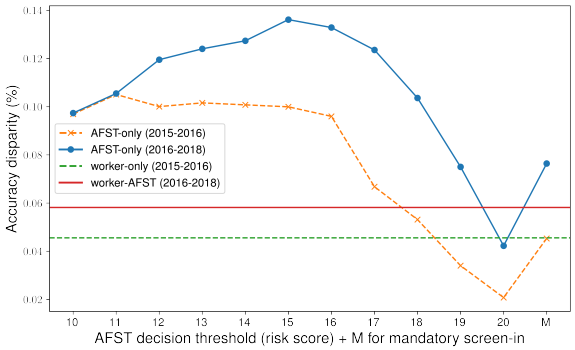


(e) False positive rate disparity

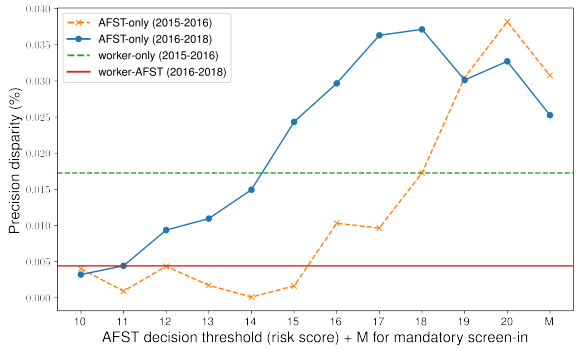
Figure 5: Comparisons of common group fairness metrics (Figure 6 in the paper [1]) based on Analysis 2



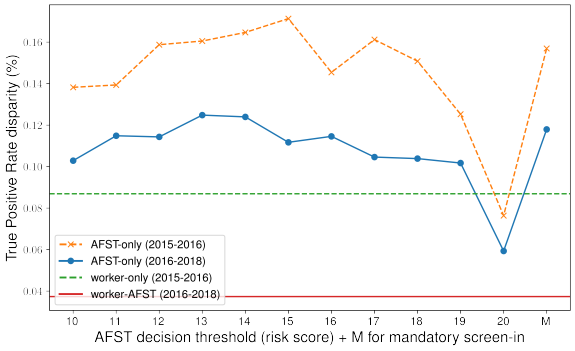
(a) Screen-in rate disparity



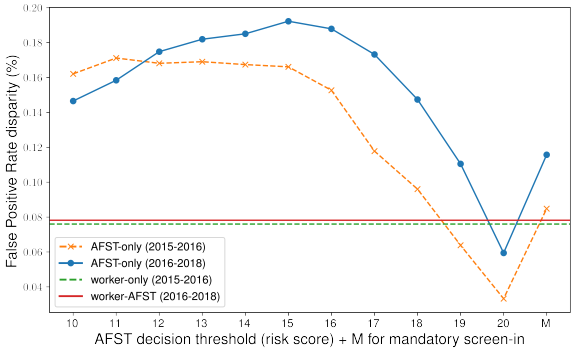
(b) Accuracy disparity



(c) Precision rate disparity

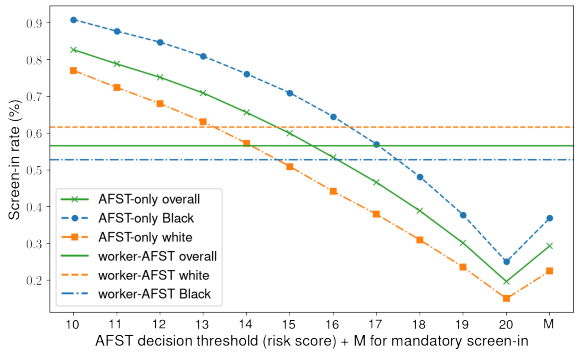


(d) True positive rate disparity

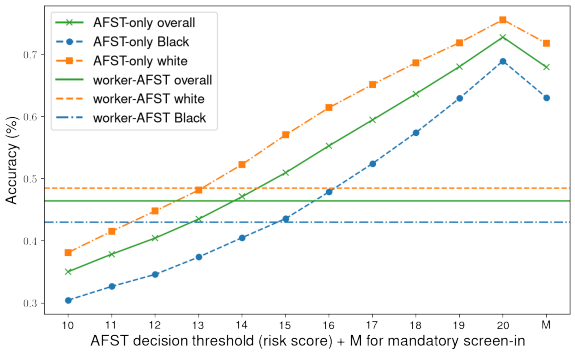


(e) False positive rate disparity

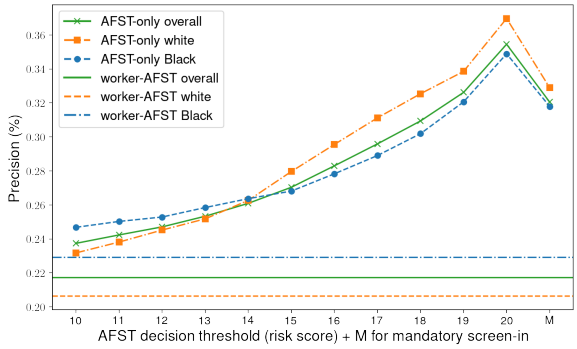
Figure 6: Comparisons of decision outcomes between worker-AFST and AFST-only decisions (Figure 7 in the paper [1]) based on Analysis 1



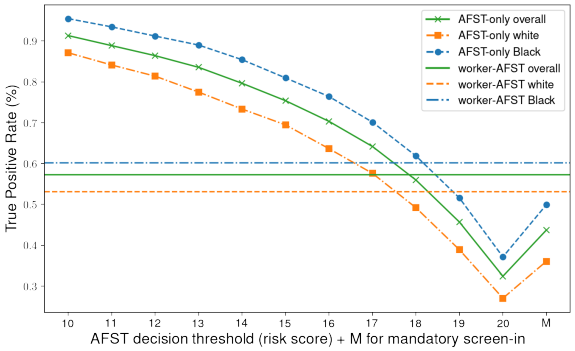
(a) Screen-in rate



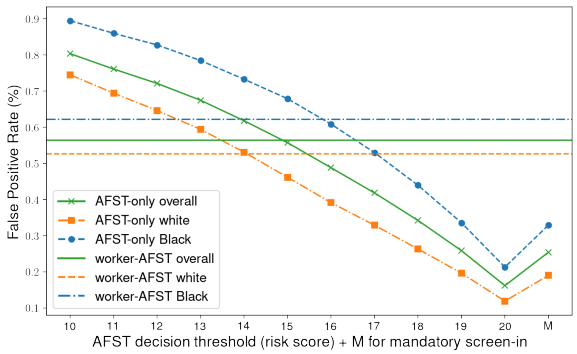
(b) Accuracy



(c) Precision

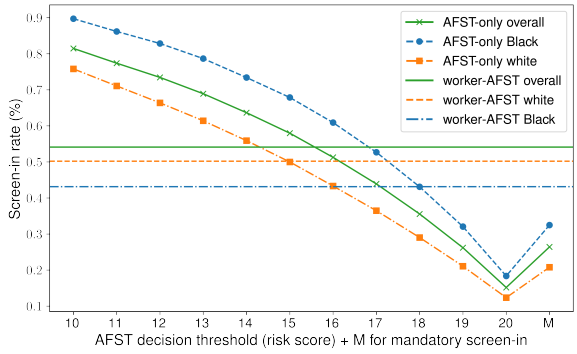


(d) True positive rate

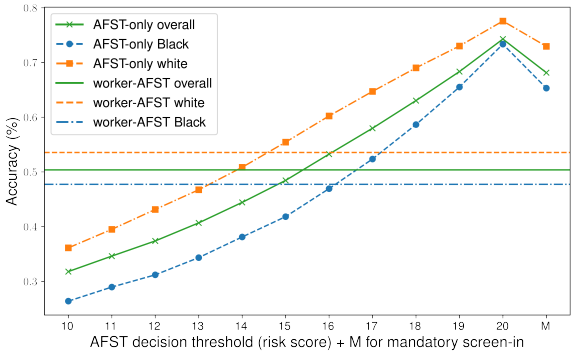


(e) False positive rate

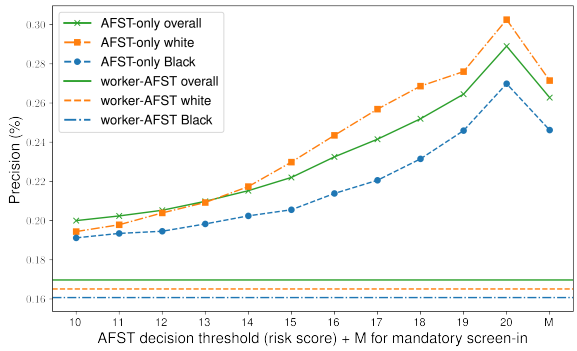
Figure 7: Comparisons of decision outcomes between worker-AFST and AFST-only decisions (Figure 7 in the paper [1]) based on Analysis 2



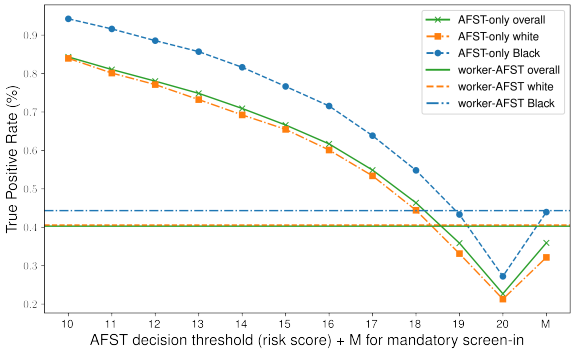
(a) Screen-in rate



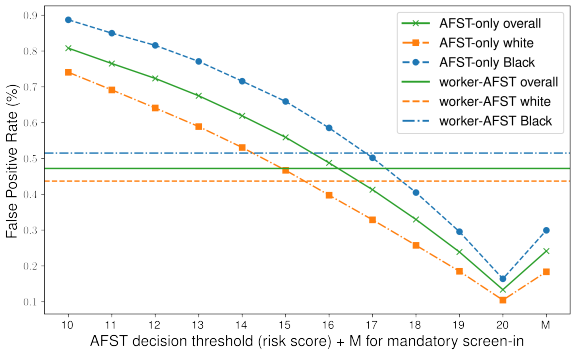
(b) Accuracy



(c) Precision



(d) True positive rate



(e) False positive rate