Validation of the PSA in San Francisco

By: Alissa Skog and Johanna Lacoe

EXECUTIVE SUMMARY

The Public Safety Assessment (PSA) is an empirically-based risk assessment tool that is used to inform pretrial release decisions across the country. The tool measures the risk of a person failing to appear at a court hearing, being arrested for new criminal activity while on pretrial release, or being arrested for new violent criminal activity while on pretrial release. San Francisco adopted the PSA in May 2016. In addition to the tool, criminal justice stakeholders in the county, including the courts, Sheriff, and District Attorney, developed a local policy document called the Decision-Making Framework (DMF) which translates the PSA score into a recommendation to the judge. The San Francisco DMF includes overrides to the tool for certain charges that increase the supervision level recommend by the PSA or generate a recommendation not to release.

This validation study examines the accuracy and reliability of the PSA in predicting failures to appear, new arrests, and new arrests for violent offenses for persons released pretrial in San Francisco, as required by California Senate Bill 36. The study also investigates whether there is any disparate effect or bias in the tool's scoring based on sex, race, or ethnicity. A tool is considered calibrated if individuals assessed as having the same risk level by the tool have the same outcomes, regardless of their demographic group. A tool is not calibrated if there is evidence of predictive bias, defined as differential prediction based on specified demographic factors. This indicates that the relationship between a given risk score and the outcome measure it aims to predict is weaker for one group of individuals than for another.

Our analysis finds:

- The PSA risk scales are fair to good predictors of the risk of failure to appear, new criminal activity, and new violent criminal activity based on industry standards.
- There is evidence of predictive bias in the PSA risk scales in San Francisco. Specifically, we find:
  - The failure to appear scale is calibrated by ethnicity and sex, but not calibrated by race
  - The new criminal activity scale is calibrated by race and ethnicity, but not calibrated by sex
  - The new violent criminal activity flag is calibrated by ethnicity and sex, but not calibrated by race

Our analytical approach follows those implemented in recent local validations of the PSA in jurisdictions including Kentucky, Los Angeles, and Harris County, Texas. The analysis measures how predictive the tool is in practice in San Francisco for people that are released prior to trial. We acknowledge this results in a biased sample, as we only observe the outcomes for individuals whom judges decided to release. We follow the analytical strategy set forth in previous validations, and as required under SB 36, to facilitate cross-jurisdiction comparison, which we believe is an important contribution to the overall debate about the use of risk assessment tools in pretrial decision-making. The study is not an evaluation of the impact of any particular pretrial release option, nor is it an evaluation of the San Francisco Pretrial Diversion Project.

The California Policy Lab builds better lives through data-driven policy. We are an independent, nonpartisan research institute at the University of California with sites at the Berkeley and Los Angeles campuses. CPL receives general-operating support from Arnold Ventures, (AV) as part of the foundation's investment in state policy labs. We have also received AV funding for research on pretrial justice, such as our recent report on bail reform. AV developed the Public Safety Assessment, which is the subject of this report, and we did not seek their support for this report to avoid any appearance of a conflict of interest. In addition, we conducted this work without input from AV to ensure objective results. After reporting our initial findings to our San Francisco agency partners, the San Francisco Pretrial Diversion Project solicited AV's assistance in implementing changes that respond to the findings from this report. This research publication reflects the views of the authors and not necessarily the views of our funders, our staff, our advisory board, the San Francisco Sheriff's Office, the San Francisco Pretrial Diversion Project, the San Francisco District Attorney's Office, or the Regents of the University of California.

Table of Contents

I. Background

A. Risk Assessments

Jurisdictions across the United States are making efforts to standardize pretrial release recommendations to increase the number of individuals released during the pretrial period while maintaining public safety, and to address disparities in release and detention. Actuarial risk assessments have emerged as one tool to inform release decisions. Most assessments estimate the risk of different pretrial failures separately, such as failure to appear in court, arrest on a new offense, and arrest on a new violent offense. The resulting scores are then translated into a category of risk, ordered from low to high, that classifies a relative risk of pretrial failure.

Local stakeholders set policy to guide the interpretation of different risk scores and categories. These policies are not empirically based, rather they reflect the local tolerance of the risk of pretrial misconduct. One jurisdiction may determine anyone with a likelihood of arrest greater than 45 percent should not be recommended for any type of release, whereas another could recommend releasing individuals with 45-50 percent likelihood of arrest to intensive pretrial supervision.

The risk scores and accompanying recommendations do not replace judicial discretion and decision-making (Desmarais, 2019). They are additional pieces of information provided to the judge to use in his or her release decision. Judges can (and do) decide to release persons assessed as high risk and, conversely, opt to detain those assessed as low-to-medium risk.

Pretrial risk tools use observable factors to predict the aggregate risk for similarly situated individuals; they are not precise predictors of individual outcomes. Therefore, the tool will never predict everyone's actual outcomes – individuals assessed as having a low likelihood of new arrest will be arrested, and those with a high likelihood of a new arrest will complete the pretrial period without one.

Pretrial risk tools are often touted as a strategy to reduce racial disparities in pretrial incarceration as they generate a transparent and empirical prediction of risk that can help correct or reveal bias in release decisions by individual judges. However, risk assessments are not a panacea for the extreme racial disparities in the criminal justice system. Risk predictions rely on historical data which reflect the pervasiveness of racial disparities in the criminal justice system. Black individuals are three times as likely to be arrested for drug-related offenses as White individuals, despite selling and using at similar rates, (Hamilton Project, 2016) and six times more likely to be incarcerated for any offense than White individuals (US Bureau of Justice Statistics, 2018). These disparities are explained by institutional biases in the justice system (Alexander 2020), individual biases of decision makers (Bonilla-Silva, 2006), and structural inequalities that limit access to employment, healthcare, and education (Massey and Denton, 1998). Risk assessment tools do not correct for these racialized differences in criminal history, but instead use criminal history as a predictive factor. Opponents warn that the codification of these historical inequities into risk assessment tools further compounds generations of disparate treatment (Mayson, 2018).

B. The Public Safety Assessment

The Public Safety Assessment (PSA) was developed by Arnold Ventures using data from 750,000 cases from nearly 300 jurisdictions across the United States. The PSA predicts the likelihood of three

outcomes during the pretrial period: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA).[1] The development team tested hundreds of factors and found nine to most effectively predict the likelihood of pretrial failure. The risk factors include prior convictions, incarceration, prior failures to appear, violent offenses, pending cases, and age (see Table 1).[2] The factors are counted and weighted to produce a final 6-point risk scale for NCA and FTA and a binary flag for NVCA.

TABLE 1. Factors used to Calculate FTA, NCA, and NVCA scores

RISK FACTOR	IS FACTOR USED FOR THIS PRETRIAL OUTCOME PREDICTION?		
	FTA	NCA	NVCA
Age (22 or Younger) at Arrest		X	
Current Violent Offense			X
Current Violent Offense & 20 years old or younger			X
Pending charge at time of offense	X	X	X
Prior misdemeanor conviction		X	
Prior felony conviction		X	
Prior conviction	X		X
Prior Violent conviction		X	X
Prior failure to appear in past two years	X	X	
Prior failure to appear older than two years	X		
Prior sentence to incarceration		X	

Note. FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity.

San Francisco implemented the PSA in May 2016 as part of a broader effort to expand and accelerate safe non-monetary release to pretrial supervision. While the risk assessment tool is consistent across all jurisdictions implementing the PSA, many jurisdictions create their own policy document – the Decision-Making Framework (DMF) – to set recommended levels of pretrial supervision for each combination of failure to appear and new criminal activity risk scores.[3] The San Francisco DMF recommends release to one of three levels of supervision or does not recommend release (Figure 1). It also includes two overrides to the risk score. The first is for specific offenses that automatically generate a recommendation of "Release Not Recommended", known as Step 2 Exclusions. The initial list included 12 offense groups and has been revised twice to include 7 additional sets of offenses.[4] The

[1] The PSA uses a more expansive list of charges for its "violent" definition than those included in the California Penal code. The PSA definition specifies that an offense is considered violent "when a person causes or attempts to cause physical injury to another person." Examples of offenses that are included but not considered "violent" per the penal code include: misdemeanor battery (PCs 240, 241(b), 241(c), 243.4, etc.), misdemeanor sex assault (PCs 261.5), and misdemeanor elder abuse (PCs 268(b) and 368(c)). Full list of offenses considered violent for the PSA available upon request.

[2] For more information about the PSA risk factors, weights, and scales see: https://advancingpretrial.org/psa/factors/#psa-factors.

[3] Arnold Ventures (AV) no longer recommends the four-step DMF. After San Francisco adopted the PSA, AV replaced the DMF with the Release Conditions Matrix which no longer includes a "Release Not Recommended" option. The DMF is still in place in San Francisco. See the AV guide for further information: https://advancingpretrial.org/guide/guide-to-the-release-condition-matrix/

[4] Possession of a Firearm by a Prohibited Person (PC 29800(a)) and Assaults with Intent to commit Mayhem or Certain Sex Offenses (PC 220) were added in September 2017. Five additional firearm charges (PCs 298065, 29815, 29820, 29825, and 29900) were added in July 2019.

second override is for specific offenses that automatically increase the recommended supervision by one level, known as Step 4 Bump Ups.

Figure 1. San Francisco Decision Making Framework

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	OR - NAS	OR - NAS				
FTA 2	OR - NAS	OR - NAS	OR - NAS	OR - Minimum	SFPDP - ACM	
FTA 3		OR - NAS	OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended
FTA 4		OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended	Release Not Recommended
FTA 5		SFPDP - ACM	SFPDP - ACM	SFPDP - ACM* Release Not Recommended	Release Not Recommended	Release Not Recommended
FTA 6				Release Not Recommended	Release Not Recommended	Release Not Recommended

* Release Not Recommended if any booked offense is a felony or violent misdemeanor per PSA Violent Offenses List; SFPDP – ACM if booked offense(s) are non-violent misdemeanors.

Note: FTA: failure to appear. NCA: new criminal activity. OR-NAS: own release-no active supervision. OR-Minimum: own release, minimum supervision. SFPDP-ACM: San Francisco Pretrial Diversion Project – assertive case management.

The San Francisco Pretrial Diversion Project (SF Pretrial) is a community-based nonprofit organization contracted by the San Francisco Sheriff's Office to administer the PSA for all new misdemeanor and felony bookings as part of the case "workup" that is presented to the judge to inform decisions regarding pretrial release. Analysts at F Pretrial submit a PSA and release recommendation within eighteen hours[5] of ID confirmation of the booked person for persons eligible for pre-arraignment review, within 72 hours of arrest for those presented at arraignment, or at a later pre-disposition court date. Individuals that are cited and released, post bail immediately, or are otherwise released within the first few hours do not have a PSA completed.

After reviewing the completed PSA and supplemental case materials, the judge may deny release or release a person to one of the three levels of supervision provided by SF Pretrial: On Recognizance-No Active Supervision (OR-NAS), On Recognizance-Minimum Supervision (OR-MS), or Assertive Case Management (ACM).[6] OR-NAS does not involve any formally supervision on the part of SF Pretrial; staff only contact clients to remind them of upcoming hearings. Clients released to OR-MS receive court reminders and are also required to check in with SF Pretrial twice weekly by phone. ACM is assigned by judges for higher needs clients and offers more intensive and structured case management.

[5] After the *Buffin* decision taking effect in January 2020, SF Pretrial has 8 hours post-ID confirmation to complete the PSA for pre-arraignment eligible cases. *Buffin v. City & Cty. of San Francisco*, No. 15-cv-04959, 2016 U.S. Dist. LEXIS 142734 (Oct. 14, 2016).

[6] Individuals may also be released to electronic monitoring (EM) - which can be proscribed in tandem with SF Pretrial supervision or as the only form of supervision, to Court OR - a lower level of supervision provided by the Courts, or via local citation or otherwise without supervision.

Additionally, staff at SF Pretrial develop individual treatment plans for their ACM clients and can make referrals to get the client additional support for issues related to substance use, education, employment, and behavioral health.

The PSA and DMF provide the judge with a recommendation; s/he has discretion to make a release determination that agrees with or deviates from that recommendation. During the analysis period, judges presented with a recommendation of release actually released the person at pre- or first arraignment in 70% of cases. Conversely, judges did not release a person by first arraignment in 30% of cases in which a release was recommended. The plurality of those not released had a DMF recommendation of ACM (41%), followed by OR-Minimum (34%), and finally OR-NAS (25%). Judges followed the DMF recommendation at high rates: they detained individuals at pre- and first arraignment in 73% of cases in which the DMF result was Release Not Recommended. In cases where the DMF result was Release Not Recommended, in 27% of cases judges instead released the individual. The majority of those released were placed under ACM (60%).

C. Validation

Pretrial risk assessment tools are often developed using a national sample and are then validated in each jurisdiction to determine how well the tool measures the outcomes it is designed to predict using data from the specific locality. For a tool to be considered valid, it must estimate the likelihood of new arrest and/or failure to appear at rates that are both statistically and politically acceptable (Desmarais, 2019).

Prior to the implementation of the PSA in San Francisco, Christopher Lowenkamp investigated the predictive validity of the PSA using a sample of historic cases in San Francisco between January 1, 2010 and December 25, 2013 (Lowenkamp, 2016). After applying the PSA to these completed cases, the analysis found that all three risk assessment scales adequately predicted their respective outcomes. This analysis did not look into differential predictivity, or whether the predictive accuracy of the tool varies by racial/ethnic groups or gender.

Under SB 36, (passed in October 2019), California requires all counties to validate their pretrial risk assessments by July 1, 2021 and every three years thereafter. We are validating the San Francisco PSA and producing this report in fulfillment of this legislative requirement. In this analysis, we assess the overall predictivity of the PSA and test whether there is disparate effect or bias based on sex or race/ethnicity. We also explore how policy and implementation decisions change the predictivity of tool.

There has been a suite of local validations of the PSA in recent years.[7] These studies, combined with the requirements outlined in SB 36, provide a standard framework to guide local validations. In this study, we closely follow the model set forth in the Kentucky validation (DeMichele et. al, 2018). Critical questions about the implementation of a pretrial risk assessment tool, including how the predictivity of the tool performs compared to human judgement; whether the risk levels were appropriately calibrated for the detained population or what the pretrial outcomes would have been had this population been released; or the impact of different pretrial supervision programs on pretrial outcomes, are often outside the narrow purview of validation studies. We provide some descriptive insight into these questions and plan to address these questions more fully in future analyses.

[7] Completed PSA validation reports available through APPR: https://advancingpretrial.org/psa/research/.

II. Research Design

A. Pretrial Release and Detention

This validation study relies on observed pretrial outcomes. We do not estimate how public safety would change if the population that was detained for the full pretrial period had been released. The detained population is excluded on the assumption that incapacitation will result in near perfect appearance rates (the share of cases that appear at all court hearings) and safety rates (the share of cases with no arrest for a new custodial offense, misdemeanor or felony, during the pretrial period). It is important, however, to understand who is being detained in San Francisco and how they differ from the released population.

Table 2 summarizes the demographics and risk levels of the 13,411 individuals assessed by SF Pretrial and either released or detained for the full pretrial period during our analysis period. A slightly larger share of the detained population is Black and slightly smaller share is Latinx, compared to our analysis sample of released individuals.[8] Notably, the assessed risk of individuals detained for the full pretrial period is much higher: nearly one-third of people detained were flagged as at risk for an arrest on a new violent offense, compared to only 12 percent of the released sample. The detained sample is also more likely to have their release recommendation changed by the charge overrides in San Francisco's DMF (33% compared to 26%).

TABLE 2. Detained Sample

CHARACTERISTICS	DETAINED	RELEASED
Black (share)	0.46	0.39
Latinx (share)	0.19	0.24
White (share)	0.29	0.29
Male (share)	0.89	0.84
Age (mean)	36.28	35.47
Case Duration		
Days in custody (median)	55.35	3.50
Risk scores		
Score - FTA	3.60	2.74
Score - NCA	4.09	3.23
NCA (violent) (share)	0.31	0.12
Charge Overrides*	0.33	0.26
Total observations	3,530	9,881

Note: *Charge override indicates the share of cases whose recommendation was changed because of a Step 2 exclusion or Step 4 bump up. Cases with an override that would be RNR based on their raw risk scores alone are excluded.

A main limitation of many validation studies is that the outcomes used in the analysis are observed only for individuals who are released during the pretrial period. As evidenced in San Francisco, the detained

[8] The data are coded as White, Black, Latinx, Asian, or Other. The samples of individuals identified as Asian and Other are very small, and therefore are not broken out in this analysis.

population looks substantially different than the released population. A key concern for the interpretation of the validation results is whether there is are different detention rates based on the characteristics that we observe, such as race/ethnicity, that affects the sample of individuals included in the analysis. In San Francisco, we find that race/ethnicity is predictive of detention on its own: compared to White individuals, Black individuals are more likely to be detained and Latinx individuals are less likely to be detained. But when we combine race/ethnicity with either the PSA risk score or the DMF recommendation, we find race/ethnicity is no longer predictive of detention (See Table A-1). This suggests that there is not differential selection into detention based on the factors that we observe in the data. In other words, Black individuals assessed at a 4 on the new criminal activity risk scale are not more likely to be detained pretrial than a White individual with the same assessed risk level. This does not mean that there is not differential treatment; there is continued evidence of Black and White individuals experiencing different rates of contact with the justice system – from policing and arrest, to filing, conviction, and sentencing – leading to differences in assessed risk scores.

B. Analysis Sample

This analysis utilizes a linked longitudinal dataset of people's interactions with the criminal justice system from booking to case disposition from the San Francisco District Attorney's Office, San Francisco Sheriff's Office, San Francisco Pretrial Diversion Project (SF Pretrial), and California Department of Justice. Additional details for each data set can be found in Appendix A-2: Data Sources. We analyze all new bookings between May 1, 2016 and December 31, 2019 that resulted in a charge being filed by the District Attorney's Office. We restrict our sample to cases in which a person was released before their case disposition and thus exclude individuals that are incarcerated for the full pretrial period. [9] We then applied the following restriction criteria to identify cases for inclusion in our research sample:

1) Case is reviewed by SF Pretrial and a release recommendation was generated;
2) Case is released at pre-arraignment, arraignment, post-arraignment, or via bail before the date of disposition or case closure;
3) Age at booking is at least 18 years because the PSA is not applied juveniles;
4) Release from custody by December 31, 2019
5) Case is closed by April 30, 2020.[10]

The final sample is 9,881 unique cases. Table 3 provides descriptive statistics of the analysis sample. Overall, the groups of individuals released to SF Pretrial, on bail, and on another type of release (such as court own release, diversion, or electronic monitoring) have similar demographic characteristics. The groups differ in case duration, with individuals released on "other releases" spending more days in custody than those released to SF Pretrial or on bail. The PSA scores for both failure to appear and new criminal activity are higher, on average, for the "other release" group relative to the SF Pretrial and bail groups.

[9] Approximately 15,000 new bookings resulted in a filing during our analysis period. Slightly more than 75 percent were released prior to their case close date.
[10] Results do not change when we rerun all models without this restriction, allowing open cases to remain in the sample.

TABLE 3. Analysis Sample

CHARACTERISTICS	FULL SAMPLE	RELEASED TO SF PRETRIAL	RELEASED ON BAIL	OTHER RELEASE
Black (share)	0.39	0.38	0.45	0.41
Latinx (share)	0.24	0.26	0.24	0.19
White (share)	0.29	0.29	0.22	0.33
Male (share)	0.84	0.84	0.87	0.85
Age (mean)	35.47	35.53	33.34	36.47
Case Duration				
Days in custody (median)	3.50	2.78	4.02	17.02
Days in community (median)*	144	143	191	127
Risk scores				
Score - FTA	2.74	2.57	2.58	3.40
Score - NCA	3.23	3.05	3.16	3.88
NCA (violent) (share)	0.12	0.08	0.17	0.20
DMF Recommendation				
OR-NAS	0.28	0.35	0.19	0.13
OR-Min	0.19	0.21	0.18	0.12
ACM	0.16	0.17	0.10	0.16
Release Not Recommended	0.36	0.27	0.53	0.59
Total observations	9,881	6,736	1,131	2,014

Note: Other releases include releases to Court OR, Collaborative Court or pretrial diversion programs, electronic monitoring, another jurisdiction, or via citation. FTA: failure to appear. NCA: new criminal activity. OR-NAS: non-active supervision. OR-Min: minimum supervision. ACM: Assertive case management.

C. Empirical Strategy

Using the sample of pretrial releasees described above, this validation addresses the following research objectives:

(1) **Validation of the PSA**: How accurately do the PSA risk scales and risk factors predict likelihood of failure to appear, new criminal activity, and new violent criminal activity between release from custody and case disposition? Does the predictive validity of the PSA differ by race/ethnicity or sex?

(2) **Assessment of predictive bias and fairness:** Does the predictivity of the PSA perform differently by race/ethnicity or sex?

The results use the PSA risk scales for FTA and NCA, and the binary NVCA flag (unless otherwise stated) because this is how recommendations are given to the judge. Results are reported as success rates (appearance rate instead of failure to appear rate) in line with recommendations from Advancing Pretrial Policy and Research (APPR) that aim to emphasize success over failure.[11] The one exception is the use of failure rates when examining the outcome error rates, presented in Validation of the PSA:

[11] Recommendation to use success rates: https://advancingpretrial.org/guide/guide-to-outcomes-and-oversight/

Differential Predictivity. In this analysis, we focus on the three outcomes the PSA is designed to predict (Table 4). See Appendix B for full details of the empirical strategy.

TABLE 4. Outcome Measures

OUTCOME	DEFINITION	DATA SOURCES
Appearance Rate	The share of cases where individuals appeared at all hearings in San Francisco and did not have a bench warrant issued by the court on the case for which they were released. Cases in which a bench warrant was recalled are included because the recall reason is unknown.[12]	SF District Attorney; SF Pretrial
Safety Rate	The share of cases with no arrest for a new custodial offense (misdemeanor or felony) during the pretrial period. For this measure, we capture any eligible new arrest in the state of California.	SF District Attorney; CA DOJ
Safety Rate (Violent Offense)	The share of cases with no arrest for a new offense that is considered violent per the Public Safety Assessment Violent Offense List for California during the pretrial period. An offense is considered violent when "a person causes or attempts to cause violence to another person." For this measure, we capture any eligible new arrest in the state of California.	SF District Attorney; CA DOJ

Note: FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity. DMF: decision making framework.

Following the literature, we first estimate the performance of the PSA using Area Under the Curve (AUC) Receiver Operator Characteristics (ROC) estimates and explore variation in the estimates by race/ethnicity and sex (DeMichele et. al., 2018; Grenier et. al., 2020; and Grenier et. al., 2021). Next, we extend the same approach to evaluate how the predictivity of the PSA changes when the DMF is applied. To investigate differential predictivity by race/ethnicity and sex, we apply moderator regression techniques and present outcome-focused error rates (Skeem and Lowenkamp, 2020). These methods are described briefly below, with greater details in Appendix A-3.

III. Results

The results are summarized in three sections. First, we present the overall performance of the PSA risk scales and evaluate the predictive power of each scales' components. Next, we explore differences in predictivity of the PSA risk scores by race, ethnicity, and sex. We conclude by using two approaches to assess differential predictivity by race, ethnicity, and sex: the moderator regression and outcome-focused error rates.

A. Validation of the PSA: Overall Performance

The AUC-ROC provides a single measure of accuracy and is comparable to other validation studies. However, it may be the case that some of the specific factors used to calculate a risk score are accurate predictors of an outcome, while others are less predictive. To unpack the overall findings, we explore the relationship between the observed and predicted outcome rates for each of the PSA risk scales, and consider the predictive power of each scales' components.

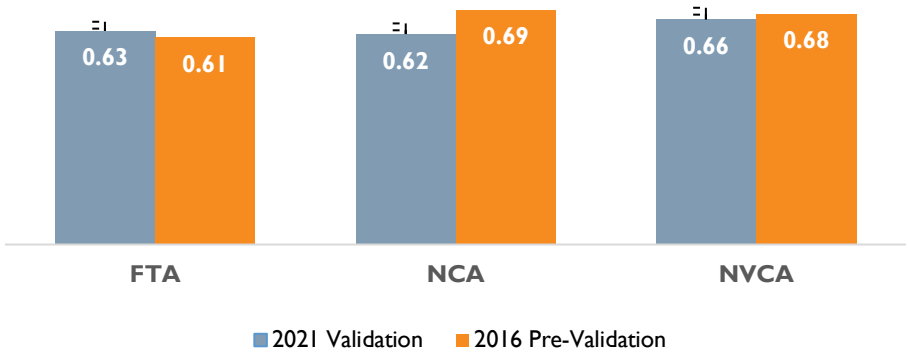
[12] Bench warrants are not counted as an FTA if they are recalled because a person is in custody in another jurisdiction. We cannot tell the reason for a recall in the data and thus cannot distinguish which would be counted as an FTA. Our main models count all bench warrants for FTA, even if they are recalled. As a robustness check, we excluded all bench warrants that are recalled for any reason and find this does not impact our results.

AUC-ROC

An AUC-ROC measure provides the probability that a randomly selected person that was successful for the full pretrial period (i.e. did not miss any hearings) has a lower FTA score than a randomly selected person who missed at least one hearing. AUC values range from 0 to 1 and an AUC score of 0.50 indicates that the tool performs no better than a coin toss. Although there are no proscribed guidelines, standard benchmarks to evaluate the predictivity of risk assessment tools in the criminal justice space do exist (Demarais and Singh, 2013). These benchmarks suggest a measure less than 0.54 indicates no predictive validity; 0.55-0.63 indicates fair, but not strong evidence of, predictivity; 0.64-0.70 are considered moderately predictive; and anything above 0.71 indicates strong predictivity.

Using this framework, the PSA risk scales in San Francisco have fair to moderate predictivity. Figure 2 presents the AUC-ROC values for each outcome predicted by the PSA and compares these rates to the 2016 pre-implementation validation (Lowenkamp, 2016). The results indicate the FTA and NCA risk scales are fair predictors (0.63 and 0.62 respectively) and the NVCA risk scale is a good predictor (0.66).[13] The results for NCA and NVCA indicate better predictivity in the pre-implementation validation than in the current validation. This could be explained by the data used to calculate PSA scores, as the pre-implementation validation used local data to estimate the risk scores and we are using the actual PSA scores, the application of the DMF, and/or differences in data used to measure safety rate, as the original validation was restricted to new arrests in San Francisco only and the current analysis uses statewide arrest data.

FIGURE 2. AUC-ROC Values by PSA Risk Scale



Note: The NVCA AUC-ROC model for the 2020 validation uses the weighted risk scale in the model. Standard errors bars are available for the 2021 validation only. FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity. Benchmarks: AUC < 0.54 = no predictive validity; 0.55-0.63 fair, but not strong evidence of, predictivity; 0.64-0.70 moderately predictive; and AUC > 0.71 strong predictivity (Demarais and Singh, 2013).

The overall predictivity of the PSA risk scales in San Francisco align with the findings from other jurisdictions with a recent validation of the PSA. Table 5 summarizes the AUC-ROC values for five other jurisdictions and we see the results fall within the fair to good range, with the exception of the FTA and NCA scales in Los Angeles, which are considered strongly predictive, and the NVCA scale in Harris County, which has no predictive value. The predictivity of the FTA and NVCA scales in San Francisco fall within the range of predictive values in the other jurisdictions. The NCA scale, however, has the lowest AUC in comparison to the other jurisdictions.

[13] The AUC results presented use the NVCA risk scale. If we change the measure to the NVCA risk flag, a binary measure indicating whether the person is at risk for a new violent arrest, the AUC value is reduced to 0.57.

TABLE 5. AUC-ROC Values from other PSA Validations

JURISDICTION	DATA	FTA	NCA	NVCA
Harris County, TX	County	0.60	0.66	0.55
McLean County, IL	County	0.70	0.67	0.61
Lucas County, OH	Regional	0.62	0.63	0.68
Kentucky	State	0.65	0.65	0.66
Los Angeles	State	0.73	0.72	0.67
San Francisco	State	0.63	0.62	0.66

Note: Differences in data sources may impact the AUC-ROC values, as jurisdictions that are restricted to local data (such as Harris, McClean, and Lucas counties) will likely be overestimating the actual safety rate at each risk level. FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity. Citations: Grenier et. al., 2020; Grenier et. al., 2020; Lowenkamp, et. al., 2020; DeMichele, et. al. 2018; and Hess and Turner, 2021).

It may be that some of the factors identified as strong predictors of pretrial outcomes in the national study are not predictive in the local San Francisco context, making the overall scores less predictive of outcomes. Therefore, in the next section, we examine the predictivity of the PSA score and each score's component factors, for each pretrial outcome separately.

Appearance

Overall, 51% of the people in the sample appeared at all of their hearings during the pretrial period.[14] Table 6 shows the appearance rate by FTA score and the number of cases released at each score. As we expect, the appearance rate declines as the risk score increases: the group assessed to have the lowest risk (FTA Score of 1) has the highest appearance rate (66% appeared at all hearings), and the groups assessed to have the highest risk have the lowest appearance rates (31% and 34% with scores of 5 and 6, respectively, appeared at all hearings). We see a slightly higher appearance rate for the cases with an FTA score of six compared to those with a five, though this should be interpreted with caution as it is likely driven by the small number of cases in the final category.

TABLE 6. Appearance Rate by FTA Score

FTA SCORE	CASES	APPEARANCE RATE (SF)	VALIDATED APPEARANCE RATE (PRE-VALIDATION)	VALIDATED APPEARANCE RATE (AV)
1	2,102	0.66	0.84	0.90
2	2,879	0.56	0.73	0.85
3	2,215	0.49	0.66	0.80
4	1,205	0.40	0.61	0.69
5	1,151	0.31	0.56	0.65
6	330	0.34	0.60	0.60

Note: The Validated Appearance Rate (Pre-Validation) includes the appearance rates from the 2016 pre-implementation validation using 15,876 cases in San Francisco and estimating the FTA score (Lowenkamp, 2016). The validated AV appearance rate is based on a national sample of 500,000 cases (retrospective) from three localities and two states (VanNorstrand, 2015). FTA: failure to appear.

For comparison, we present validated rates from the 2016 pre-validation assessment (Lowenkamp, 2016) and the validation completed by Arnold Ventures prior to the launch of the PSA (VanNorstrand, 2015). The observed appearance rates are substantially lower than what would be expected based on both the local and national validation, particularly as the FTA score increases. Some of this difference may be explained by the FTA scores themselves: the current analysis uses the actual FTA scores, produced by SF Pretrial using California Record of Arrest and Prosecution (RAP) sheets. The pre-validation used local data, which does not include arrests or prosecutions that occurred outside of San

[14]See Table A-2 for a comparison of San Francisco's appearance and safety rates to other jurisdictions using the PSA.

Francisco, to estimate the FTA risk scores, which likely results in an under estimate of the risk scores.[15]

We next test the relationship between each FTA score and the observed appearance rate within a regression framework (Table 7). With the lowest score as the reference category (FTA score =1), we see that the relationship between each score and the outcome measure (appearance at all court hearings) is statistically significant, and the odds of appearing at all hearings generally declines as the risk score increases. The exception is the odds of appearing for individuals with an FTA score of six is greater than that for individuals with an FTA score of five. Again, this is likely driven by the small number of cases that were released with an FTA score of six.

Table 7. Logistic Regression of FTA Score on Appearance Rate

FTA SCORE	ODDS RATIO
FTA = 2	0.641***
	(0.0381)
FTA =3	0.495***
	(0.0310)
FTA =4	0.347***
	(0.0259)
FTA =5	0.233***
	(0.0183)
FTA = 6	0.264***
	(0.0330)
Constant	1.973***
	(0.0911)
Pseudo R-squared	0.0370
N	9,881

The PSA uses four factors to predict the likelihood that a person will appear at all hearings. Table 8 shows that all four factors are negatively associated with appearance rates (the appearance rate is lower when the factor is present). For example, the group of individuals with a pending charge has a lower overall appearance rate (45% appear at all hearings) than the group of individuals who do not having a pending charge (53% appear at all hearings). There is a 27-percentage point difference in the appearance rate of pretrial releasees with two or more FTAs in the prior two years (30% appear at all hearings), and those without any FTA in the prior two years (57% appear at all hearings). We see a smaller difference (eight percentage points) between individuals with a single FTA in the prior two years and those with two or more.

[15] In our analysis sample, 50% of observations have an FTA score of one or two, compared to 76% of the pre-validation sample. We assume this difference is largely explained by prior failures to appear in other counties that are not observable in the San Francisco data alone.

TABLE 8. Factors in the FTA scale

RISK FACTOR		#	APPEARANCE RATE
	No	7,351	0.54
Prior Conviction	Yes	6,990	0.47
	No	2,891	0.62
	2 or more	1,221	0.30
	1	1,163	0.38
	None	7,497	0.57
Prior FTA older than 2 years	Yes	3,369	0.44
	No	6,512	0.55

The results of a bivariate logistic regression assessing the relationship between each of the four risk factors individually and the likelihood of appearance (Table 9) confirm the descriptive results. The presence of each factor individually decreases the odds that a person will appear at all hearings ($p<0.001$) (column I). The strongest association is between prior FTAs and a decrease in the odds of appearance.

Table 9. Logistic Regression Results: FTA Risk Factors and Appearance Rate

MEASURE	(I) BIVARIATE	(II) MULTIVARIATE
Pending Charge	0.699***	0.998
	(0.0324)	(0.0526)
Prior Conviction	0.524***	0.649***
	(0.0237)	(0.0328)
Two FTAs (<2years)	0.320***	0.364***
	(0.0214)	(0.0265)
One FTA (<2years)	0.468***	0.523***
	(0.0303)	(0.0361)
FTAs (>2years)	0.653***	0.878**
	(0.0279)	(0.0421)
Constant		1.817***
		(0.0735)
Pseudo R-squared		0.0380

Note: The bivariate results test each risk factor individually and the third factor (FTAs in prior two years) is tested as a categorical value. The multivariate results test all risk factors concurrently. Standard errors in parentheses. *p<0.05 ** p<0.01 *** p<0.001. FTA: failure to appear.

Testing the joint relationship between all risk factors and appearance we find pending charge, though individually is associated with the appearance rate, is no longer significant when combined with all other risk factors (column II). This loss of significance suggests that pending charge is strongly related to at least one other risk factor or there is omitted variable bias in the simple regression. All other risk factors are statistically significant, indicating that when used together they are still predictive of appearance.

Safety

Almost half (45%) of the people in the sample were not arrested on a new misdemeanor or felony offense in California during their pretrial release period. Table 10 shows the observed safety rate by NCA score. As we expect, the safety rate declines as the risk score increases. Sixty-seven percent of individuals with an NCA score of 1, did not have new criminal activity during the pretrial period, while among those with an NCA score of 6, only 26% did not have new criminal activity. Similar to the findings for appearance rate, San Francisco's observed safety rate is substantially lower than the local and national validated rates.[16]

TABLE 10. Safety Rates by NCA Score

NCA SCORE	CASES	SAFETY RATE (SF)	VALIDATED APPEARANCE RATE (SF PRE-VALIDATION)	VALIDATED SAFETY RATE (AV)
1	1,294	0.67	0.93	0.90
2	1,872	0.51	0.86	0.85
3	2,362	0.45	0.79	0.77
4	2,612	0.38	0.73	0.70
5	1,067	0.32	0.69	0.52
6	674	0.26	0.63	0.45

Note: The Validated Appearance Rate (Pre-Validation) includes the appearance rates from the 2016 pre-implementation validation using 15,876 cases in San Francisco and estimating the FTA score (Lowenkamp, 2016). The validated AV appearance rate is based on a national sample of 500,000 cases (retrospective) from three localities and two states. NCA: new criminal activity.

We next test the relationship between each NCA score and the observed safety rate (Table 11). With the lowest score as the omitted variable (NCA=1), we see each score is statistically significant and the safety rate declines as the risk score increases, consistent with the descriptive results above.

TABLE 11. Logistic Regression of NCA Score on Safety Rate

NCA WEIGHTED SCORE	ODDS RATIO
NCA = 2	0.516***
	(0.0388)
NCA =3	0.398***
	(0.0288)
NCA =4	0.297***
	(0.0201)
NCA=5	0.227***
	(0.0201)
NCA = 6	0.170***
	(0.0180)
Constant	2.059***
	(0.122)
Pseudo R-squared	0.039

[16] There could be an underestimate of risk in the 2016 validation, as the validation used restricted local data to estimate PSA scores. In our sample, only 32% have an NCA score of one or two, compared to 66% in the pre-validation study. The difference in the scores may be due to the 2016 validation omitting prior criminal activity that occurred outside of San Francisco county in the calculation of the risk scores.

The PSA uses seven factors to estimate the likelihood that a person will not be arrested on a new misdemeanor or felony while on pretrial release. We find, descriptively, that the safety rate does not vary substantially for individuals 22 or younger compared to those who are 23 and above at the time of arrest (Table 12). The group of individuals with any prior violent conviction has a lower safety rate, but we do not find a difference in safety rates when we disaggregate those with one versus those with two or more violent convictions (37 percent). Individuals with FTAs in the prior two years have a lower safety rate than those without (20 percentage point difference for those with one FTA and 27 percentage point difference for those without any prior FTAs).

TABLE 12. Factors in the NCA scale

FACTOR		#	SAFETY RATE
22 or Younger at Arrest	Yes	1,624	0.46
	No	8,257	0.44
Pending Charge	Yes	2,530	0.33
	No	7,351	0.49
	No	3,472	0.53
Prior Felony Conviction	Yes	4,654	0.39
	No	5,227	0.50
	1	2,386	0.37
	None	6,732	0.48
FTA in Prior Two Years	2 or more	1,221	0.23
	1	1,163	0.30
	None	7,497	0.50
Prior Incarceration	Yes	5,628	0.39
	No	4,253	0.52

We test the relationship between each factor and the safety rate separately, using a bivariate logistic regression, and together, with a multivariate regression (Table 13). Differences emerge when looking at risk factors individually compared to their combined relationship with the safety rate. Individuals who are 22 or younger at the time of arrest do not have a statistically significantly different safety rate in the bivariate model (column I). However, when evaluated with the other six factors, being 22 years old or young at arrest is associated with a lower safety rate (column II). Conversely, a prior felony conviction or two or more prior violent convictions are independently associated with the safety rate, but they lose significance in the joint model, suggesting they may be strongly related to other factors.

TABLE 13. Logistic Regression Results: NCA Risk Factors and Safety Rate

MEASURE		(I) BIVARIATE	(II) MULTIVARIATE
22 or Younger at Arrest	Yes	1.058	0.682***
		(0.0578)	(0.0423)
Pending Charge	Yes	0.515***	0.706***
		(0.0249)	(0.0381)
Prior Misdemeanor Conviction	Yes	0.580***	0.775***
		(0.0246)	(0.0463)
Prior Felony Conviction	Yes	0.648***	0.875*
		(0.0265)	(0.0527)
Prior Violent Conviction	Two+	0.642***	0.871
		(0.506)	(0.0758)
	One	0.629***	0.817***
		(0.0308)	(0.0466)
FTAs (<2years)	Two+	0.286***	0.373***
		(0.0207)	(0.0290)
	One	0.429***	0.541**
		(0.0291)	(0.0389)
Prior Incarceration	Yes	0.574***	0.827**
		(0.0236)	(0.0581)
Constant			1.648***
			(0.0740)
R-Squared			0.0483

Note: The bivariate results test each risk factor individually and the multivariate results test all risk factors concurrently. Prior FTAs and Prior Violent Conviction are modeled as categorical variables in the bivariate models. Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. N= 9,881. FTA: failure to appear. NCA: new criminal activity.

Safety (no new violent criminal activity)

The majority (82%) of the people in the sample were not arrested on a new violent offense during their pretrial release. Unlike the weighted FTA and NCA scores, the NVCA score is used as a binary measure indicating whether or not the person is at risk for committing a new violent offense while on release. Table 14 shows the safety rate for individuals flagged for NVCA compared to those who are not flagged. A larger share of individuals without the NVCA flag avoided a new arrest for a violent offense during the pretrial period than those with the flag. The validated safety rates generated from the national sample used to create the PSA are substantially higher than what we observe in San Francisco: approximately 30 percentage points higher for those with the NVCA flag and 14 percentage points higher for those without the flag.[17]

[17] More than double the sample in our validation is flagged for NVCA compared to the pre-validation sample (12 percent compared to 4 percent).

TABLE 14. Safety Rate (New Violent Offense) by NVCA Flag

NVCA	CASES	SAFETY RATE (SF)	VALIDATED SAFETY RATE (SF PRE-VALIDATION)	VALIDATED SAFETY RATE (AV)
Yes	1,139	0.64	0.94	0.918
No	8,742	0.84	0.98	0.976

Note: NVCA: new violent criminal activity. Validated safety rate (SF pre-validation) is from the 2016 pre-implementation validation (Lowenkamp, 2016). Validated safety rate (AV) is based on a national sample of 500,000 cases (retrospective) from three localities and two states.

When evaluated in a regression framework, the relationship between the NVCA flag and the safety rate is statistically significant, indicating that the presence of the flag is associated with a lower odds of completing the pretrial period without a new arrest for a violent offense (Table 15).

TABLE 15. Bivariate Logistic Regression Results of NVCA flag on safety rate (new violent offense)

NVCA FLAG	ODDS RATIO
Constant	5.391***
	(0.159)

The PSA uses five factors to predict the likelihood that someone will be arrested for a new violent offense while on pretrial release. Table 16 shows that all factors are negatively associated with the safety rate, indicating their presence is associated with a higher rate of new violent activity. The difference in magnitude is often much smaller for the NVCA risk factors compared to those in the FTA and NCA risk scales: there is a small difference in the safety rate between the population with a pending charge or prior conviction and those without. Measures of current or prior violent offenses generate the largest difference between the groups. Individuals whose current booked offense is violent have an 18-percentage point lower safety rate than those who are not booked on a violent offense (69 percent compared to 87 percent). Individuals who are 20 years or younger and booked on a violent offense have a 16-percentage point lower safety rate than their peers who are not booked on a violent offense, and those who are above 20 years of age.

TABLE 16. Factors in the NVCA scale

RISK FACTOR		#	SAFETY RATE
Current Offense is Violent	Yes	2,729	0.69
	No	7,152	0.87
Current Offense is Violent & 20 and Under	Yes	266	0.66
	No	9,616	0.82
	No	7,351	0.83
Prior Conviction	Yes	6,990	0.81
	No	2,891	0.85
Prior Violent Conviction	Two or more	763	0.72
	One	2,386	0.77
	None	6,732	0.85

The bivariate logistic regression results, testing the relationship between each risk factor and the safety rate, confirms the descriptive results that independently each factor is significantly related to the safety rate (Table 17, column I). Prior conviction, however, is no longer significant when evaluating in the multivariate model (column II). This suggests there may be omitted variable bias in the bivariate model, or that the factor is strongly correlated with one or more of the other factors used to generate the score. Descriptively there is a small difference in the safety rate for those with a pending charge, but this factor remains significant in both models, indicating it has a significant association with the safety rate. The three factors based on prior violent offenses are significant in terms of magnitude in both models, though we lose significance on the coefficient for violent and 20 or younger in the multivariate model, likely due to its strong correlation with the violent offense risk factor.

Table 17. Logistic Regression Results: NVCA Risk Factors and Safety Rate

MEASURE	(I) Bivariate	(II) Multivariate
Violent Offense	0.333***	0.320***
	(0.0181)	(0.0185)
Violent Offense & >20	0.416***	0.670**
	(0.0551)	(0.0949)
Pending Charge	0.834***	0.673***
	(0.0487)	(0.0417)
Prior Conviction	0.775***	0.899
	(0.0463)	(0.0649)
Two+ Prior Violent Conviction	0.457***	0.470***
	(0.0401)	(0.0450)
One Prior Violent Conviction	0.579***	0.578***
	(0.0342)	(0.0394)
Constant		10.19***
		(0.669)
R-Squared		0.062

Note: The bivariate results test each risk factor individually. The multivariate results test all risk factors concurrently. Standard errors in parentheses.
* p<0.05 ** p<0.01 *** p<0.001. N= 9,881.

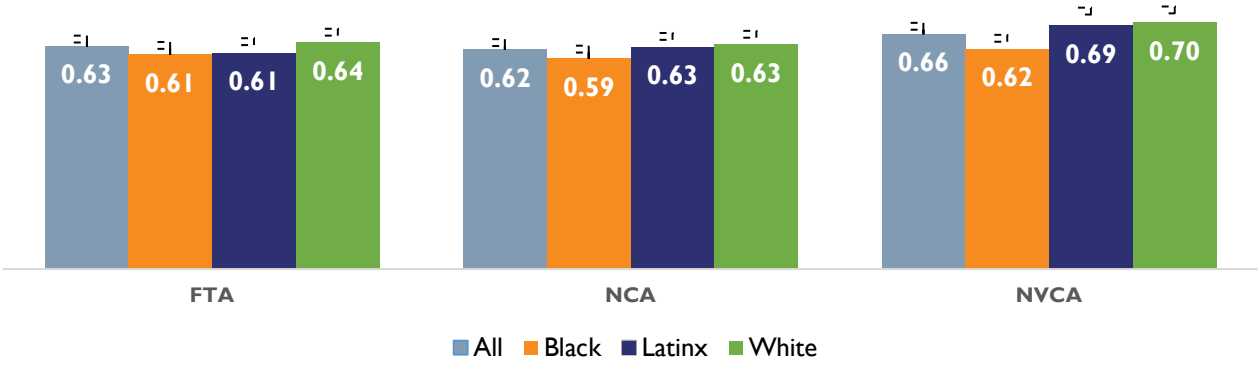
Overall, we find that the many of the factors used to calculate the PSA scores are correlated with the pretrial outcomes the scores are meant to predict.

B. Validation of the PSA: Differential Predictivity

In this section we explore differences in predictivity of the PSA risk scores by race/ethnicity and sex. We use the AUC-ROC values to assess the predictive utility of the PSA risk scores by group. There are no significant differences in the predictivity of the FTA risk score for Black, Latinx, and White individuals (Figure 3). There are, however, statistically significant differences in the values for Black versus White individuals for the two safety scales. The AUC-ROC for the NCA scale is 0.59 for Black individuals compared to 0.63 for White individuals. The NCA risk scale is considered to be a fair predictor for both groups. The difference in the predictivity of the NVCA scale between Black and White individuals is larger and statistically significantly (0.62 compared to 0.70). The NVCA scale is only a fair predictor for Black individuals, compared to a good predictor for all other groups. These

results indicate that the PSA NCA and NVCA risk scales are less predictive of new arrest for Black individuals compared to White individuals.

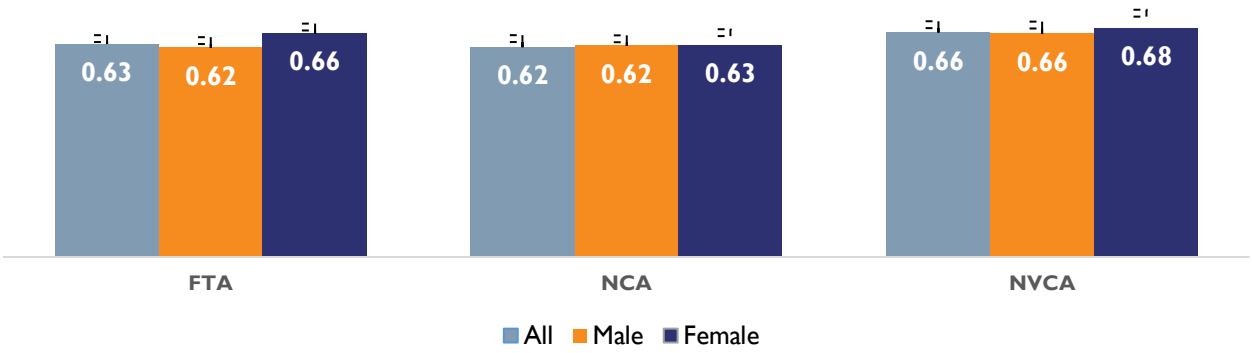
FIGURE 3. AUC-ROC Values by Race/Ethnicity



Note: FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity. The NVCA AUC-ROC models use the weighted risk scale. Using the binary NVCA flag reduces the predictivity: 0.57 overall, 0.56 for Black individuals, 0.56 for Latinx individuals, and 0.59 for White individuals. Benchmarks: AUC < 0.54 = no predictive validity; 0.55-0.63 fair, but not strong evidence of, predictivity; 0.64-0.70 moderately predictive; and AUC > 0.71 strong predictivity (Demarais and Singh, 2013).

The differences in the AUC-ROC values between sexes are smaller (Figure 4). The AUC-ROC value for females on the FTA scale is larger than males (0.66 compared to 0.62) and this difference is statistically significant. There are no significant differences in the scales for new criminal or new violent criminal activity between sexes. These results indicate that the PSA FTA scale is less predictive of appearance for males compared to females.

FIGURE 4. AUC-ROC Values by Sex

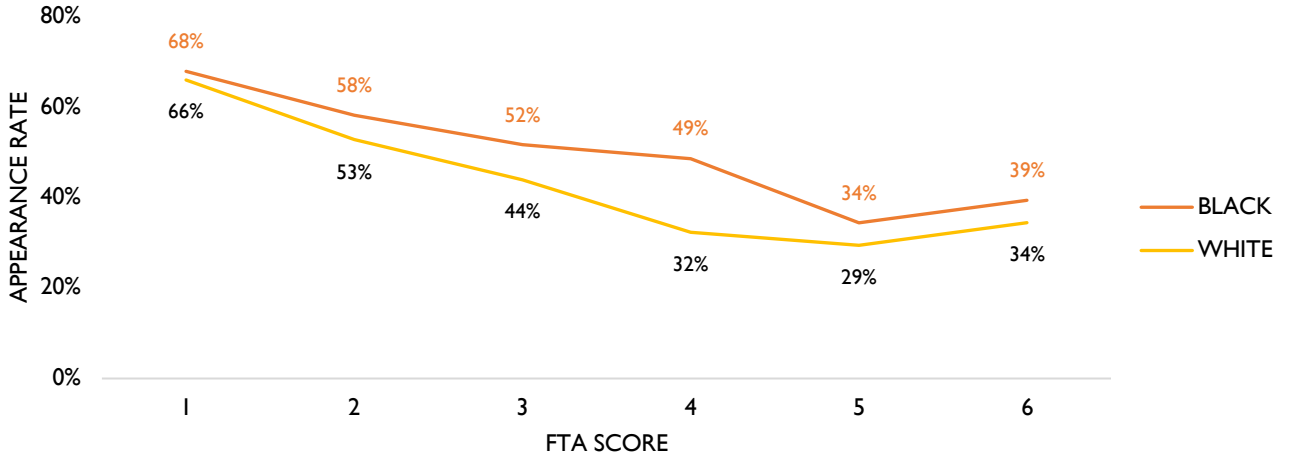


Note: FTA: failure to appear. NCA: new criminal activity. NVCA: new violent criminal activity. The NVCA AUC-ROC models use the weighted risk scale. Using the binary NVCA flag reduces the predictivity: 0.57 for males and females. Benchmarks: AUC < 0.54 = no predictive validity; 0.55-0.63 fair, but not strong evidence of, predictivity; 0.64-0.70 moderately predictive; and AUC > 0.71 strong predictivity (Demarais and Singh, 2013).

Appearance

Next, we assess the observed appearance rate at each risk score by race/ethnicity and sex to see if there are differences. Figure 5 shows the observed appearance rate by FTA score, disaggregated by race. Black individuals appear for court at a slightly higher rate than their White counterparts at all risk levels.

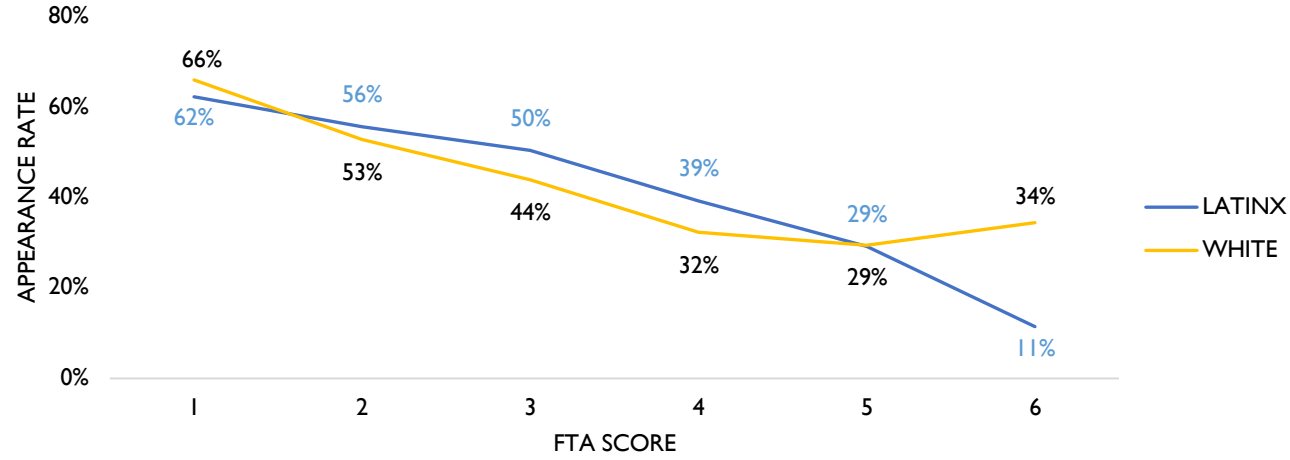
FIGURE 5. Appearance Rates by FTA Score and Race



Note: FTA: failure to appear. See Table A-3 for the sample size of risk score and race/ethnicity combination.

Latinx individuals appear for court at higher rates than White individuals, at most risk levels (Figure 6). The appearance rates for an FTA score of six are likely skewed for all three groups due to small sample sizes.

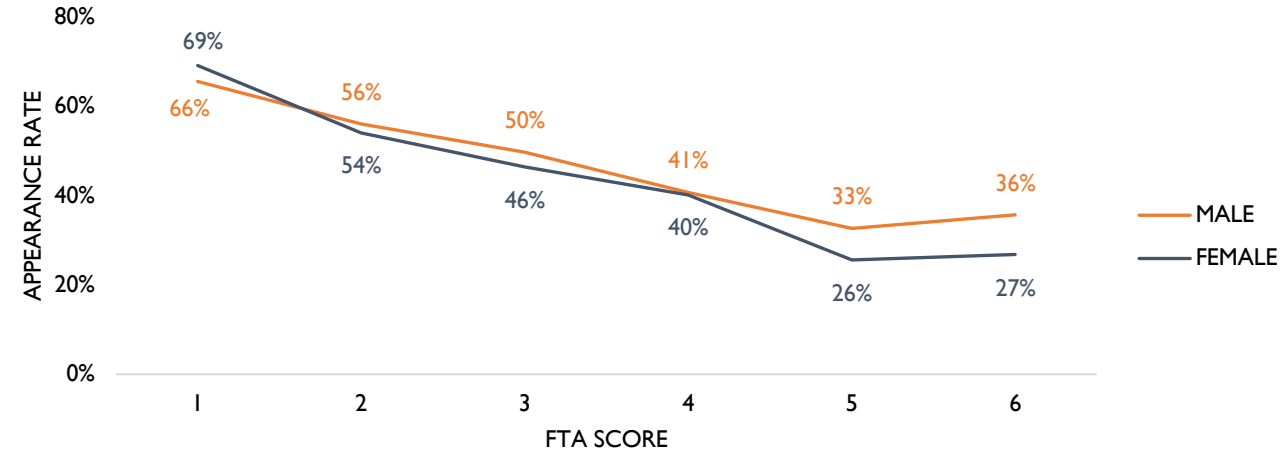
FIGURE 6. Appearance Rates by FTA Score and Ethnicity



Note: FTA: failure to appear. The low appearance rate for Latinx individuals released with an FTA score of six should be interpreted with caution. This is a very small sample (n=33) which is likely skewing the results. See Table A-3 for the sample size of risk score and race/ethnicity combination.

Females and males have similar appearance rates at the lowest FTA score, but females have slightly lower rates at all subsequent scores compared to males (Figure 7). Looking at differences in the underlying risk factors, we see that the appearance rate for females with a pending charge or one FTA in the previous two years is seven-percentage points lower compared to males (38 and 32 percent, respectively, compared to 46 and 39 percent) (Appendix Table A-4).

FIGURE 7. Appearance Rates by FTA Score and Sex

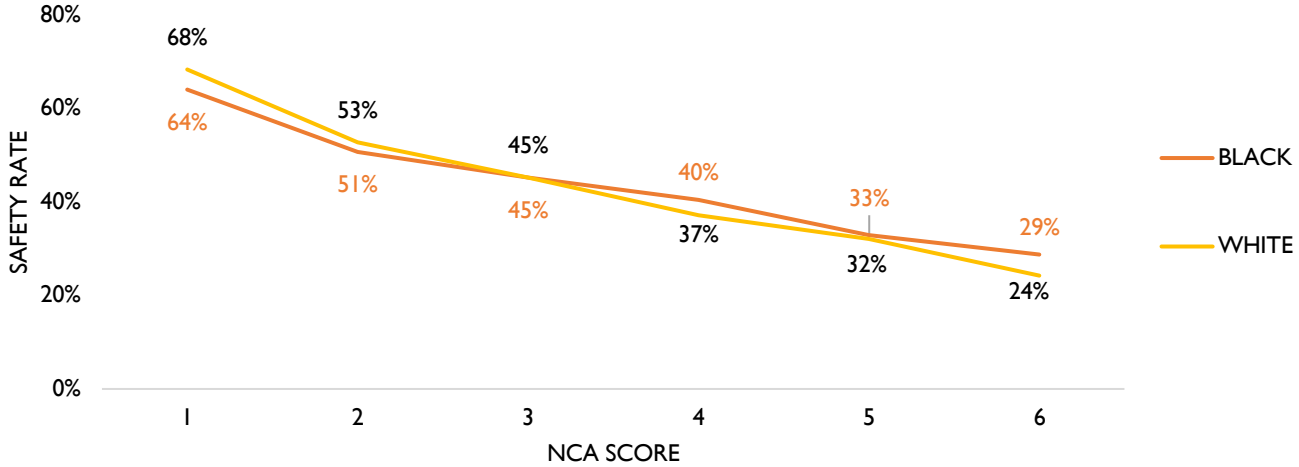


Note: FTA: failure to appear. See Table A-4 for the sample size of risk score and sex combination.

Safety

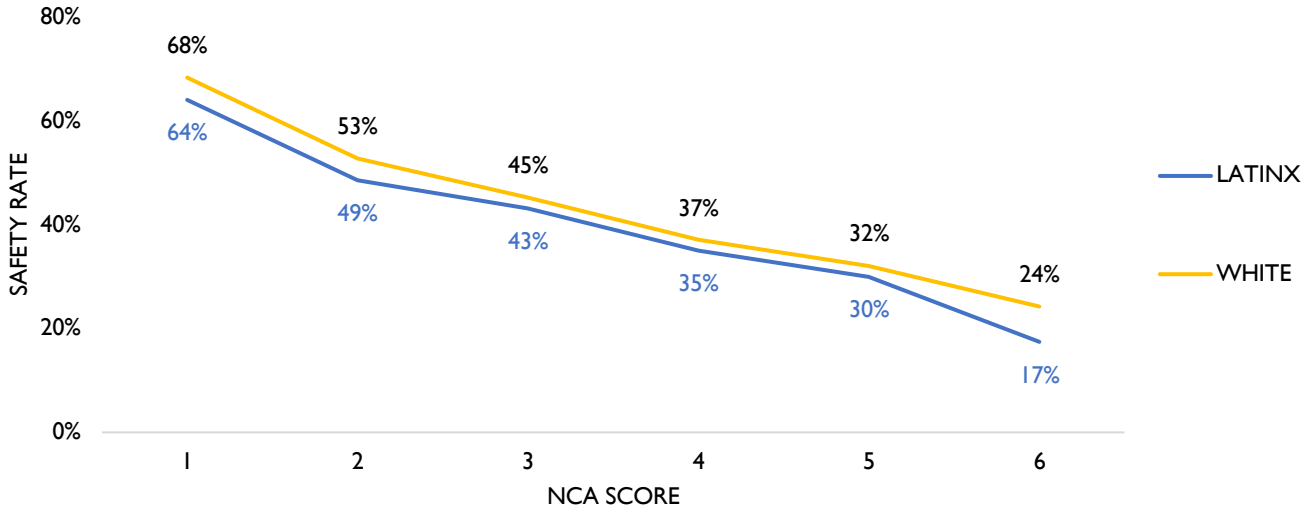
Next, we disaggregate differences in the NCA scale by race/ethnicity and sex. Figure 8 shows the observed safety rate (share of individuals who complete the pretrial period with no new arrests) by NCA score, disaggregated by race, and Figure 9 by ethnicity. At the lowest risk levels, White individuals have a higher safety rate than their Black and Latinx counterparts. Latinx individuals have slightly lower safety rates across all risk levels compared to the other groups, though the values for the highest score (NCA 6) should be interpreted cautiously as the large differences are likely explained by the small sample size. The observed safety rates are fairly consistent when we disaggregate risk factors by race/ethnicity (Appendix table A-5). The largest difference is a 12-percentage point lower safety rate for Latinx individuals with two or more prior violent offenses compared to their Black and White counterparts, though only a very small number of cases have this risk factor present (763 overall, 68 Latinx cases). Additionally, we see White individuals *without* a prior misdemeanor conviction have an 11-percentage point higher safety rate than their Black and Latinx counterparts (60 percent compared to 49 percent).

FIGURE 8. Safety Rates by NCA Score and Race



Note: See Table A-5 for the sample size of risk score and race/ethnicity combination.

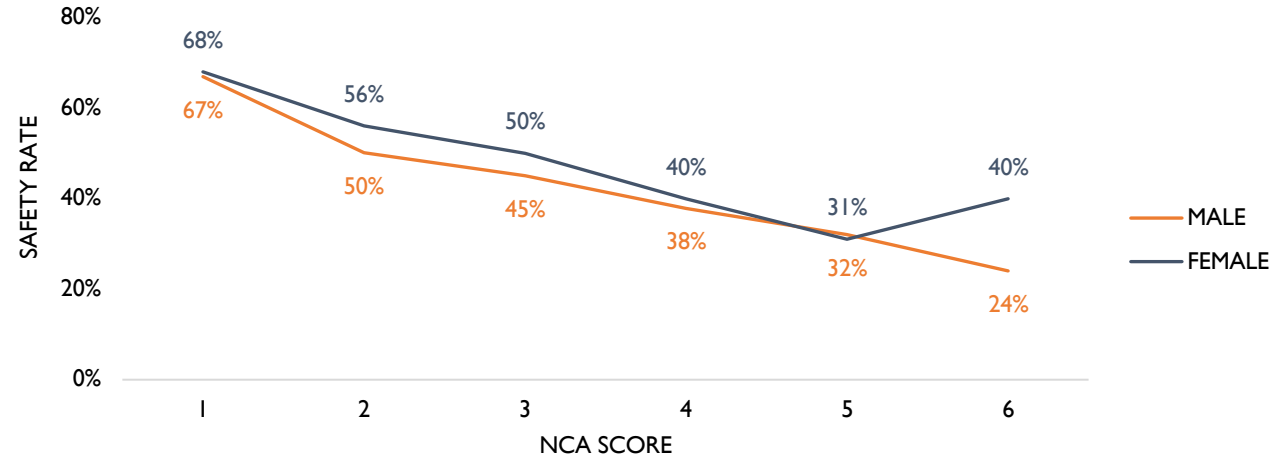
FIGURE 9. Safety Rates by NCA Score and Ethnicity



Note: See Table A-5 for the sample size of risk score and race/ethnicity combination.

Disaggregating the safety rate at each risk score by sex (Figure 10), we see higher safety rates for females compared to males at all points, except among those with a risk score of 5. Looking at the underlying risk factors, we see the largest difference emerge between males and females that are 22 or younger at the age at arrest (Appendix table A-3). These younger males have a twelve-percentage point lower safety rate than similarly aged females in San Francisco (44 percent compared to 56 percent).

FIGURE 10. Safety Rates by NCA Score and Sex

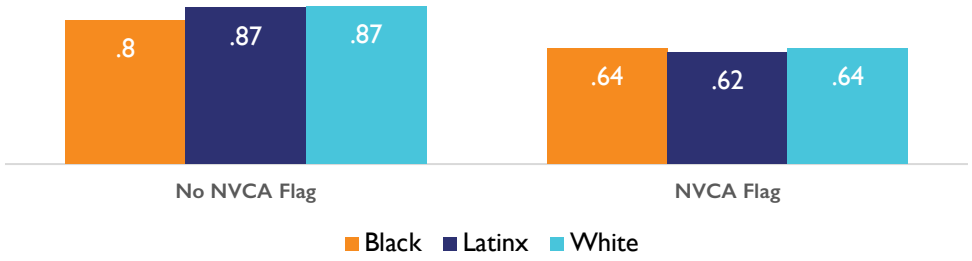


Note: See Table A-4 for the sample size of risk score and sex combination.

Safety (NVCA)

We disaggregate the NVCA flag by race/ethnicity (Figure 11). Latinx individuals have a two-percentage point lower safety rate in the presence of the NVCA flag. Among those not flagged as at risk for new violent criminal activity, Black individuals have seven percentage point lower safety rate than their Latinx and White counterparts.

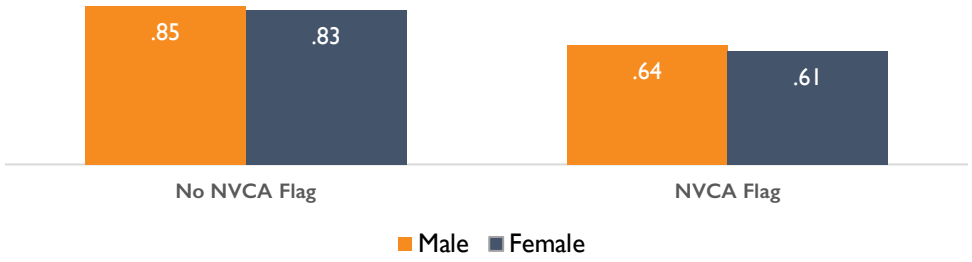
FIGURE 11. Safety Rates by NVCA Score and Race/Ethnicity



Note: Safety rate: share without a new violent arrest. See Table A-5 for the sample size of risk score and race/ethnicity combination.

Disaggregating by sex, the share of females that complete the pretrial period without an arrest for a new violent offense is slightly lower than the share of males (Figure 12). This trend is consistent across those with and without an NVCA flag. While the overall number of individuals with two or more prior violent offenses is fairly small (713 males, 50 females), females in this group are much less likely to complete the pretrial period without a new arrest for a violent offense than males (58 percent compared to 73 percent) (Appendix table A-5).

FIGURE 12. Safety Rates by NVCA Score and Sex



Note: Safety rate: share without a new violent arrest. See Table A-5 for the sample size of risk score and sex combination.

Overall, we observe some differences in the predictivity of the PSA across race/ethnicity and sex. When we consider the individual scales, we observe variation in predictivity by race/ethnicity and sex for both the FTA and NCA scores. We extend this analysis in the next section.

C. Bias & Fairness Assessment

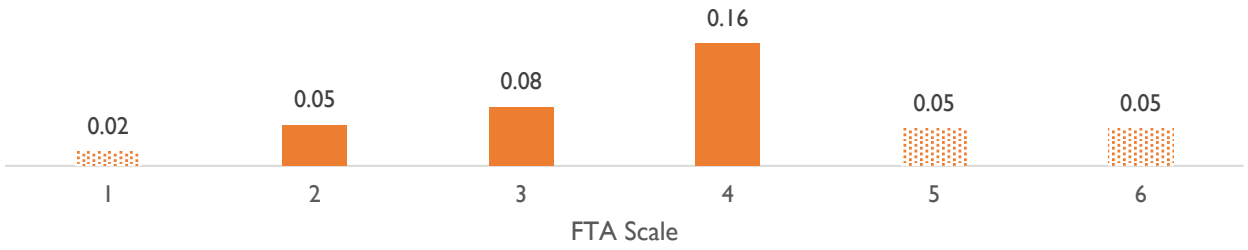
The AUC-ROC values show us the PSA's accuracy is considered to be fair to good, though the NCA and NVCA scales are less predictive of safety rates for Black individuals compared to White individuals. The results also suggest that the FTA scale is less predictive of appearance rates for females compared to males. In this section, we consider the structure of the relationship between race/ethnicity and sex and the risk scores. First, we test for predictive bias by estimating nested moderated regressions to evaluate whether the relationship between the risk score and the outcome are moderated by inclusion of information on race, ethnicity, or sex. Second, we compare differences in the outcome error rates across groups to understand how differential predictivity is occurring in practice. Combined, these measures are not proscriptive, rather they provide additional information to help stakeholders in San Francisco have data-informed discussions about the tradeoff between predictivity and fairness.

Calibration Condition

The figures presented in the previous section show evidence of differences in the relationships between risk scores (both FTA and NCA) and the outcomes they intend to predict, between individuals of different races and sexes. A primary question that remains, is this evidence of predictive bias, or differences in the relationship between the assessed risk level and the outcome? If so, how big is the problem? In this section we conduct nested moderated regressions, also known as the Cleary test, to assess whether a certain risk score has the same meaning for one group compared to another. We first estimate three separate models with only the demographic group of interest, only the assessed risk level, and then both sets of variables. These models show whether each of the demographic and risk score variables are individually significant, such as whether being male is, on its own, predictive of appearance rate. The final model includes the assessed risk level and interactions between each risk level and the demographic variable. This model estimates whether pretrial success is dependent on how the risk factor and demographic group function together, which is evidence of predictive bias.

The figures below present differences in the outcome measure between demographic groups at each assessed risk level, which are estimated using the fully interacted regression model (see Appendix A-1 for full results). From these, we can observe the magnitude of the difference in the outcome between the groups, and whether the differences between the groups are statistically significant. Figure 13 presents the differences in appearance rate between Black and White individuals with the same assessed FTA scores. There are statistically significant differences in appearance rates among individuals with an assessed FTA score of 2, 3, and 4. Within each of these assessed risk levels, Black individuals have higher appearance rates than White individuals, on average, ranging from 5 percentage points to 16 percentage points.

FIGURE 13. Difference in Appearance Rate by FTA Score for Black individuals, relative to White individuals



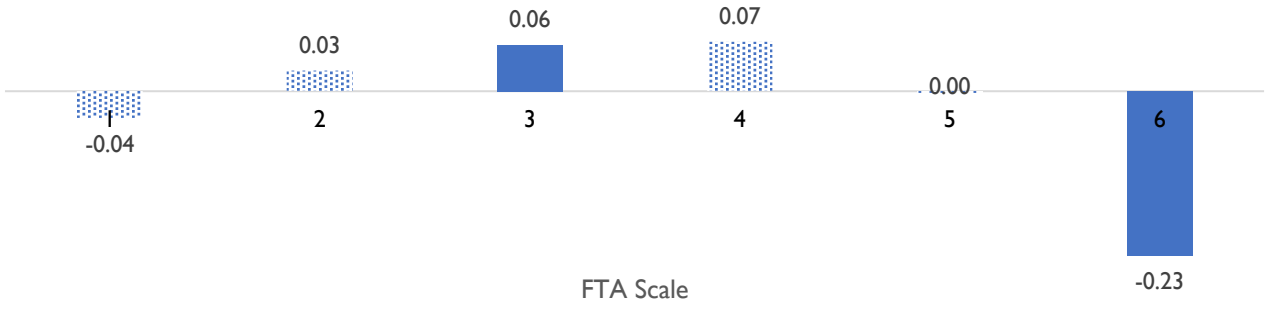
Note: FTA: failure to appear. The figure plots the coefficients on the interaction terms between FTA score and the indicator for Black. See Table A-6 for full regression results. Solid bars represent statistically significant differences for Black individuals relative to White individuals.

We use a likelihood ratio test to assess whether the model fit is better once we add the interaction terms. We find a positive and statistically significant difference between the models (Chi-sq: 11.79, p-value: 0.038) which, combined with the coefficients on the interactions, is indicative of predictive bias. This suggests that race moderates the strength of the relationship between the FTA score and the outcome (appearing at all court hearings). We run the same test between model two and model three to test whether there are different intercepts for Black compared to White individuals and find evidence of intercept bias (Chi-sq: 31.39, p-value: 0.000). In other words, the relationship between FTA and the outcome appears in part to be driven by differences in appearance rates between the groups. Together, these results suggest that the FTA scale is not calibrated by race.

The model for Latinx/White reflected only two statistically significant differences in appearance rate (Figure 14), and the estimate for FTA risk score of 6 is likely driven by a small sample size. Likelihood

ratio tests show no evidence of intercept bias, but some evidence of slope bias (Chi-sq: 15.164, p-value: 0.01). Again, this finding is likely driven by the estimate for risk score 6. The model of appearance rates by sex reflects no statistically significant differences and the likelihood ratio test finds no evidence of intercept or slope bias. This suggests that the FTA scale is calibrated by sex.

FIGURE 14. Difference in Appearance Rate by FTA Score for Latinx individuals, relative to White individuals

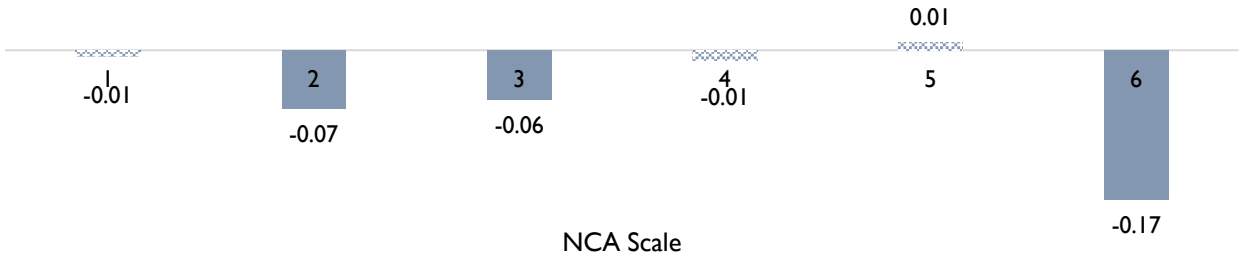


Note: FTA: failure to appear. The figure plots the coefficients on the interaction terms between FTA score and the indicator for Latinx. See Table A-7 for full regression results. Solid bars represent statistically significant differences for Latinx individuals relative to White individuals.

Turning to the new criminal activity scale, the fully interacted models show no statistically significant differences in safety rates within the NCA risk levels for Black and White individuals (Appendix Table A-9), or for Latinx and White individuals (Appendix Table A-10). The likelihood ratio test finds evidence of intercept bias between Latinx and White individuals (Chi-sq: 5.223, p-value: 0.022). This suggests that the NCA scale is calibrated by race and ethnicity, though there may be some underlying differences in safety rates for Latinx individuals relative to White individuals.

In contrast, we observe differences in the new criminal activity scale between males and females. Figure 15 presents the differences in the safety rate for males relative to females, and shows that for NCA risk levels 2, 3, and 6, males have lower safety rates than females despite being assessed as having similar risk of new arrest during the pretrial period. The likelihood ratio test indicates that the differences are driven by intercept bias (Chi-sq: 8.698, p-value: 0.003), or different underlying safety rates between the groups, not slope bias. This suggests that the NCA scale is not calibrated by sex.

FIGURE 15. Difference in Safety Rate by NCA Score and Sex

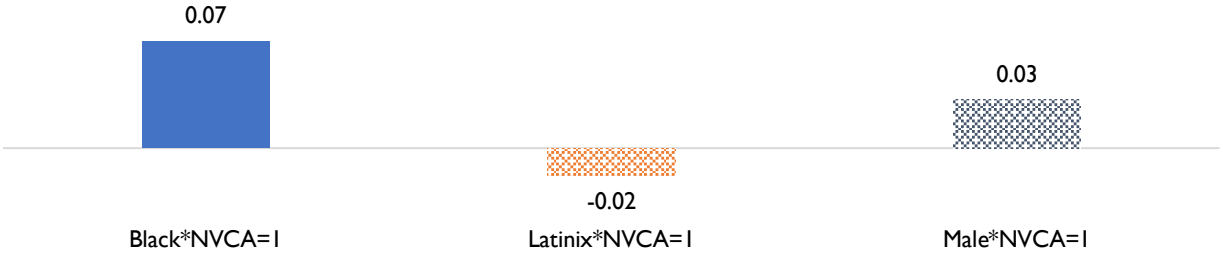


Note: NCA: new criminal activity. The figure plots the coefficients on the interaction terms between NCA score and the indicator for male. See Table A-11 for full regression results. Solid bars represent statistically significant differences for males relative to females.

Finally, we consider the relationship between the new violent criminal activity flag and the safety rate for new arrests for violent offenses, by race, ethnicity, and sex. The new violent criminal activity flag takes the value of one or zero, therefore the models include only one interaction term. Figure 16 summarizes the primary results: we observe a statistically significant difference in the safety rate

(arrests for violent offenses) between Black and White individuals, indicating that Black individuals assessed as at risk of new violent criminal activity are more likely to complete the pretrial period *without* a new violent arrest than White individuals assessed as at risk of new violent criminal activity. The likelihood ratio test indicates the presence of both intercept bias (Chi-sq: 48.737, p-value: 0.00) and slope bias (Chi-sq: 5.484, p-value: 0.019). The differences in safety rates between Latinx and White individuals, and males and females, are not statistically significant. This is evidence that the NVCA flag is calibrated by ethnicity and sex, but not calibrated by race.

FIGURE 16. Difference in Safety Rate (Violent) by Race, Ethnicity, and Sex



Note: NVCA: new violent criminal activity. The figure plots the coefficients on the interaction terms between NVCA flag and the indicator for Black relative to White individuals, Latinx relative to White individuals, and males relative to females from separate regression models. See Table A-12, A-13, and A-14 for full regression results. Solid bars represent statistically significant differences.

One concern is that models including only PSA scores and demographics omit important information – notably, the actual type of release supervision individuals experience during the pretrial period. The type of pretrial release is determined by the judge. If release types vary systematically by race, ethnicity, or sex among individuals with the same assessed risk level, then differences in the relationship between the risk score and the outcome between these groups may be in part due to release type. However, when we estimate the same interaction models summarized above including controls for release type, we observe nearly identical results (see Appendix A-1 for full results).

Predictive Bias and Fairness: Disparate Impact

To assess disparate impact, we calculate the Positive Predictive Value (PPV), False Positive Rate (FPR), and False Negative Rate (FNR). To calculate these rates, we need to select a cutoff score to define the high-risk population for the FTA and NCA scales. We classify a person as high risk in a specific case if the score is 4 or above, in line with the cutoff for the NVCA flag. As a robustness check, we increase the cutoff value to 5 and include these in the appendix (see Table A-20). While all results in previous sections frame outcomes as positive measures (i.e. share that successfully appeared at all hearings), this section will use negative outcomes (i.e. share that failed to appear at one or more hearings) to facilitate interpretation.

First, we present the positive predictive value (PPV), which tells us how often a positive test (a PSA resulting in a designation of high risk) represents a true positive (actual misconduct during pretrial release) (Table 18). In other words, this shows the odds of pretrial misconduct conditional on being assessed as high risk. Across all three outcome measures, we see a larger number of Black individuals are assessed as high risk compared to White and Latinx individuals. We also see that Black individuals have a lower PPV for both FTA and NCA compared to their Latinx and White counterparts, indicating a smaller share of Black individuals who are classified as high risk subsequently fail to appear or are arrested for a new custodial offense compared to the other groups. The PPV for new violent criminal activity is much lower across all groups, which is reasonable as new violent criminal activity occurs less frequently. Black and White individuals who are assessed as high risk for new violent criminal activity

are actually arrested for a new violent offense in 36% of releases, compared to 38% for Latinx individuals.

TABLE 18. Positive Predictive Value Rate by Race/Ethnicity

	POSITIVE PREDICTIVE VALUE					
	BLACK		LATINX		WHITE	
	PPV	N	PPV	N	PPV	N
FTA	0.59	1,127	0.67	487	0.69	917
NCA	0.63	2,094	0.68	796	0.66	1,246
NVCA	0.36	642	0.37	145	0.35	295

Note: Positive predictive value is the percentage of people classified by the PSA as high risk who had a pretrial failure. N is the group total (denominator).

The false positive rate (FPR) is the odds of completing the pretrial period without misconduct, given a "high risk" assessment. The FPR is an outcome error rate and shows the frequency with which the tool over-estimates the risk of an individual. Black and White individuals have comparable FPRs for failure to appear: 23% of Black individuals and 22% of White individuals who complete the pretrial period without an FTA were assessed as high risk for an FTA. (Table 19). The rates begin to diverge, however, when looking at the two new criminal activity measures. Of the Black individuals that were not arrested on a new offense during pretrial release, 46% were assessed as high risk, compared to 24% for Latinx and 33% for White individuals. The FPRs are substantially lower for new violent criminal activity, but the rate for Black individuals (14%) is more than triple that of Latinx (4%) and more than 50% higher than White individuals (8%).

TABLE 19. False Positive Rate by Race/Ethnicity

	FALSE POSITIVE					
	BLACK		LATINX		WHITE	
	FP	N	FP	N	FP	N
FTA	0.23	2,062	0.13	1,257	0.22	1,325
NCA	0.46	1,675	0.24	1,071	0.33	1,275
NVCA	0.14	3,002	0.04	2,052	0.08	2,430

Note: False positive rate is the percentage of people who did not have a pretrial failure but had been classified as high risk by the PSA. N is the total number that did not have a pretrial failure.

The false negative rate (FNR) is the second outcome error rate measured. The FNR tells us the share of individuals who did have a pretrial failure that were assessed as low risk. It is the inverse of the FPR, essentially indicating the share whose risk was under-assessed. Of the 1,841 Black individuals that failed to appear at one or more hearings, 64% were assessed as low risk by the PSA (compared to 72% for Latinx and 59% for White individuals) (Table 20).

TABLE 20. False Negative Rate by Race/Ethnicity

	FALSE NEGATIVE					
	BLACK		LATINX		WHITE	
	FN	N	FN	N	FN	N
FTA	0.64	1,837	0.72	1,144	0.59	1,539
NCA	0.40	2,224	0.59	1,330	0.48	1,589
NVCA	0.74	897	0.85	349	0.76	434

Note: False negative rate is the percentage of people had a pretrial failure but had not been classified as high risk by the PSA. N is the total number that did have a pretrial failure.

IV. Policy Considerations

The validation of the San Francisco PSA concludes that the PSA risk scales are fair to good predictors of actual rates of failure to appear in court, new criminal activity during the pretrial period, and new violent criminal activity. These results meet the minimum level typically required for a tool to be considered sufficiently predictive.

The analysis also provides evidence of predictive bias in the FTA, NCA, and NVCA scales. The results of the moderator regressions show the FTA scale is calibrated by sex, but not calibrated by race and ethnicity. Further, the NCA scale is calibrated by race and ethnicity, but not calibrated by sex. Finally, the NVCA flag is not calibrated by race, but is calibrated by ethnicity and sex. Overall, these results suggest that for each of the risk scales, the relationship between the risk score and the outcome is different by demographic group.

The PSA workgroup should consider the following:

Review PSA risk scales: Reviewing the PSA risk scales with AV and other jurisdictions with a recent validation. Indications of predictive bias are concerning, and modifications should be explored to minimize differential predictions by race/ethnicity and sex. Working closely with the tool developer will ensure that any modifications to the scales do not invalidate the tool altogether.

Low-touch interventions to increase appearance rates. SF might consider interventions to increase appearance rates overall, and particularly among females released prior to trial. While males and females have comparable appearance rates overall (51 percent compared to 50 percent), females assessed as higher risk on the FTA risk scale have lower appearance rates than their male counterparts (Figure 7). Analyzing the factors that affect appearance rates for females and males might illuminate potential points of intervention. Descriptively, we see a higher appearance rate for individuals released to SF Pretrial's supervision in the COVID-19 when court was held virtually. In addition, courts may have waived appearances and stayed bench warrants at higher rates during this period, which may have contributed to improved appearance rates. Determining what factors contributed to the improvement in appearance rates could provide helpful information about improvements to court processes after the pandemic.[18]

Re-validation plan: Establish a plan to re-validate the tool every three years, as required by SB 36. If changes are made to the risk scales or DMF, CPL recommends a pre-validation and a follow-up study approximately one year into implementation to ensure the changes have not led to differential predictions by race, ethnicity, or sex.

[18] Based on regular analysis of appearance and safety rates of SF Pretrial clients, CPL found a seven-percentage point increase in the appearance rate during the COVID-19 pandemic compared to the prior year (69 percent compared to 76%).

Works Cited

Alexander, Michelle. 20202. The New Jim Crow: Mass Incarceration in the Age of Colorblindness. 10th anniversary edition. New York: The New Press.

Bonilla-Silva, Eduardo. 1997. "Rethinking Racism: Toward a Structural Interpretation." American Sociological Review 62 (3): 465–80. https://doi.org/10.2307/2657316.

DeMichele, Matthew et. al. (2018). The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168452.

Desmarais, Sarah L. and Lowder, Evan M. (2019) Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys. Retrieved from: https://www.safetyandjusticechallenge.org/wp-content/uploads/2019/02/Pretrial-Risk-Assessment-Primer-February-2019.pdf.

Desmarais, Sarah L. Singh, Jay P. (2013). Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States. Retrieved from: https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf

Greiner, James, Stubenberg, Matthew, and Halen, Ryan (2020). Validation of the PSA in Harris County, TX. Retrieved from: http://a2jlab.org/wp-content/uploads/2020/11/Validation-of-the-PSA-in-Harris-County-TX.pdf.

Greiner, James, Stubenberg, Matthew, and Halen, Ryan (2020). Validation of the PSA in McClean County, Illinois. Retrieved from: http://a2jlab.org/wp-content/uploads/2020/11/Validation-of-the-PSA-in-McLean-County-IL.pdf.

Hamilton Project (2016). Rates of Drug Use and Sales by Race. Retrieved from: https://www.hamiltonproject.org/charts/rates_of_drug_use_and_sales_by_race_rates_of_drug_related_criminal_justice

Hess, James and Turner, Susan (2021). Validation of the PSA in Los Angeles County. Center for Evidence Based Corrections, University of California at Irvine.

Lowenkamp, Christopher, DeMichele, Matthew, and Warren, Lauren Klein (2020). Replication and Extension of the Lucas County PSA Project. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727443.

Lowenkamp, Christopher (2016). Investigating the PSA-court and pretrial interventions: San Francisco.

Massey, Douglas S., and Nancy A. Denton. 1998. American Apartheid. Cambridge, MA: Harvard University Press

Mayson, Sandra G (2018). Bias In, Bias Out. Yale Law Journal 2218. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004.

Miller, Matthew, Mauroff, David, Barron, Cristina, and Broughton, Bob (2021). An Evolving Pretrial Landscape: San Francisco's Strategic Response to *In re Humphrey*. Forthcoming article.

VanNostrand, Marie. (2015). Measuring and Managing Pretrial Risk: Improving Public Safety, Fairness, and Cost Effectiveness. Retrieved from: https://www.sog.unc.edu/sites/www.sog.unc.edu/files/course_materials/NC%20Association%20of%20District%20Court%20Judges%20Conference%20VanNostrand.pdf.

U.S. Bureau of Justice Statistics, Prisoners in 2016, 8 tbl.6 (Jan. 2018).

Appendix A-1: Additional Results

Pretrial Releases and Detentions

Table A-1 summarizes the results of four models that test the relationship between race/ethnicity, DMF recommendation, and PSA risk score and detention for the full pretrial period.

TABLE A-1: Factors Associated with Detention

	(1)	(2)	(3)	(4)
Person Identified as Black	0.0336**	0.00580	0.0138	0.0243
	(0.0106)	(0.0103)	(0.0104)	(0.0140)
Person Identified as Latinx	-0.0447***	-0.00783	-0.00580	-0.0225*
	(0.016)	(0.0109)	(0.0110)	(0.0112)
Black*DMF				-0.000026
				(0.00724)
Latinx*DMF				0.0116
				(0.00790)
OR-Minimum		0.0842***		
		(0.0101)		
ACM		0.161***		
		(0.0114)		
RNR		0.335***		
		(0.0893)		
FTA Score			0.0538***	
			(0.0479)	
NCA Score			0.0225***	
			(0.00484)	
NVCA Flag			0.208***	
			(0.0126)	
Constant	0.262***	0.0735***	-0.0174	0.0735***
	(0.00814)	(0.0828)	(0.0122)	(0.0902)
DMF Controls		X		X
PSA Controls			X	
DMF Interaction				X

Notes: Sample size is 12,480, full sample restricted to include individuals identified as Black, Latinx, or White. *p<0.05 ** p<0.01 *** p<0.001.

Validation of the PSA: Overall Performance

Table A-2 summarizes the appearance and safety rates reported for jurisdictions that have published their local PSA validation reports. Comparisons between jurisdictions should be done cautiously. Differences in data sources will undoubtedly impact results. Jurisdictions that are only using county or regional data will not report any new arrests that occur elsewhere in the state, likely overestimating the two safety rate measures. Differences in the risk level of the pretrial population will also impact the results. San Francisco detains very few people pretrial, compared to many jurisdictions around the country, and thus persons released likely have a higher risk profile than those elsewhere.

TABLE A-2: Appearance and Safety Rate Comparison

JURISDICTION	DATA	APPEARANCE	SAFETY	SAFETY (NVCA)
Harris County, TX	County	73%*	83%	96%
McClean County, IL	County	82%**	84%	94%
Lucas County, OH	Regional	75%	85%	96%
Kentucky	State	85%	90%	99%
Los Angeles, CA	State	46%	51%	91%
San Francisco, CA	State	51%	45%	82%

Note: *In Harris County, the study uses two definitions of FTA. A base FTA is counted when a bench warrant is issued in the case from which the PSA originated. Due to data limitations, these are restricted to FTAs in which a warrant is associated with bond forfeiture. An FTA+ has the same definition of base FTA, but relaxes the restriction that the warrant be associated with bond forfeiture. This table presents the FTA+ rate, appearance rate increase to 81% if the base FTA definition is used. **Similar to Harris County, McLean County only counts FTAs if they occur on the court number associated with the PSA.

Validation of the PSA: Differential Predictivity

Tables A-3 – A5 provide the number of persons in the release sample at each FTA, NCA and NVCA score, disaggregated by race/ethnicity and sex.

TABLE A-3: Sample Size for FTA Score, Disaggregated by Race/Ethnicity and Sex

	ALL	BLACK	LATINX	WHITE	MALE	FEMALE
1	2,102	592	742	494	1,705	397
2	2,879	1,030	811	836	2,497	382
3	2,215	1,150	361	617	1,903	312
4	1,205	504	257	380	1,001	204
5	1,150	468	195	418	951	199
6	330	155	35	119	274	56
N	*9,881*	*3,899*	*2,401*	*2,864*	*8,331*	*1,550*

TABLE A-4: Sample Size for NCA Score, Disaggregated by Race/Ethnicity and Sex

	ALL	BLACK	LATINX	WHITE	MALE	FEMALE
1	1,294	276	435	377	1017	277
2	1,872	676	584	466	1546	326
3	2,362	853	586	775	1985	377
4	2,612	1252	533	711	2297	315
5	1,067	491	177	333	887	180
6	674	351	86	202	599	75
N	*9,881*	*3,899*	*2,401*	*2,864*	*8,331*	*1,550*

TABLE A-5: Sample Size for NVCA Flag, Disaggregated by Race/Ethnicity and Sex

	ALL	BLACK	LATINX	WHITE	MALE	FEMALE
No	1,145	3,362	2,257	2,557	7,346	1,390
Yes	8,736	642	143	304	986	160
N	*9,881*	*3,899*	*2,401*	*2,864*	*8,331*	*1,550*

Table A-6 provides the calibration results comparing the FTA scale for Black to White individuals. We find evidence of intercept bias (chi-sq = 24.678, p-value = 0.00), but no indication of slope bias (chi-sq = 3.067, p-value = 0.08). In the interaction model neither the coefficient on Black nor the interaction term with FTA are statistically significant.

Calibration condition

Tables A-6 through A-14 present full results from the nested regression models. In each table, column 1 presents the relationship between the outcome and an indicator for the demographic group. Column 2 presents the relationship between the scale values and the outcome. Column three includes both the demographic group indicator and the scale values. Column 4 includes interactions between the demographic group indicator and the scales values, and column 5 include controls for pretrial release type.

TABLE A-6: Calibration Results: PSA Failure to Appear Scale (Black compared to White)

Appearance rate	(1)	(2)	(3)	(4)	(5)
Black	0.0662***		0.0676***		
	(0.0123)		(0.0121)		
FTA=2		-0.113***	-0.113***	-0.132***	-0.121***
		(0.0187)	(0.0186)	(0.0277)	(0.0276)
FTA=3		-0.181***	-0.188***	-0.221***	-0.202***
		(0.0188)	(0.0188)	(0.0294)	(0.0298)
FTA=4		-0.254***	-0.256***	-0.336***	-0.321***
		(0.0221)	(0.0221)	(0.0333)	(0.0337)
FTA=5		-0.350***	-0.349***	-0.366***	-0.363***
		(0.0221)	(0.0221)	(0.0324)	(0.0330)
FTA=6		-0.298***	-0.299***	-0.315***	-0.318***
		(0.0330)	(0.0330)	(0.0498)	(0.0501)
Black*FTA=1				0.0191	0.0256
				(0.0297)	(0.0295)
Black*FTA=2				0.0540*	0.0433
				(0.0227)	(0.0226)
Black*FTA=3				0.0773**	0.0639**
				(0.0243)	(0.0242)
Black*FTA=4				0.162***	0.143***
				(0.0331)	(0.0330)
Black*FTA=5				0.0498	0.0510
				(0.0328)	(0.0326)
Black*FTA=6				0.0490	0.0517
				(0.0594)	(0.0590)
Constant	0.463***	0.670***	0.634***	0.660***	0.618***
	(0.00932)	(0.0148)	(0.0162)	(0.0219)	(0.0251)
Observations	6763	6763	6763	6763	6761
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: evidence of intercept bias (chi-sq = 31.392, p-value = 0.000) and slope bias (chi-sq = 11.786, p-value = 0.038).

TABLE A-7: Calibration Results of PSA FTA Scale (Latinx compared to White)

Appearance rate	(1)	(2)	(3)	(4)	(5)
Latinx	0.0609***		0.0157		
	(0.0138)		(0.0137)		
FTA=2		-0.0959***	-0.0943***	-0.132***	-0.128***
		(0.0183)	(0.0183)	(0.0276)	(0.0274)
FTA=3		-0.174***	-0.171***	-0.221***	-0.218***
		(0.0208)	(0.0210)	(0.0293)	(0.0297)
FTA=4		-0.286***	-0.283***	-0.336***	-0.341***
		(0.0237)	(0.0239)	(0.0331)	(0.0336)
FTA=5		-0.344***	-0.339***	-0.366***	-0.391***
		(0.0240)	(0.0243)	(0.0323)	(0.0330)
FTA=6		-0.345***	-0.339***	-0.315***	-0.347***
		(0.0415)	(0.0418)	(0.0496)	(0.0499)
Latinx*FTA=1				-0.0373	-0.0322
				(0.0282)	(0.0279)
Latinx*FTA=2				0.0286	0.0243
				(0.0239)	(0.0237)
Latinx*FTA=3				0.0649*	0.0492
				(0.0322)	(0.0319)
Latinx*FTA=4				0.0693	0.0654
				(0.0392)	(0.0388)
Latinx*FTA=5				-0.00195	0.0112
				(0.0421)	(0.0417)
Latinx*FTA=6				-0.230*	-0.227*
				(0.0934)	(0.0924)
Constant	0.463***	0.638***	0.628***	0.660***	0.638***
	(0.00933)	(0.0138)	(0.0161)	(0.0218)	(0.0260)
Observations	5265	5265	5265	5265	5263
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: no evidence of intercept bias (chi-sq = 1.31, p-value = 0.252); evidence of slope bias (chi-sq = 15.164, p-value = 0.01).

TABLE A-8. Calibration Results of PSA FTA Risk Scale (Male compared to Female)

Appearance rate	(1)	(2)	(3)	(4)	(5)
Male	0.0146		0.0159		
	(0.0138)		(0.0135)		
FTA=2		-0.105***	-0.106***	-0.151***	-0.158***
		(0.0140)	(0.0140)	(0.0349)	(0.0347)
FTA=3		-0.170***	-0.171***	-0.228***	-0.229***
		(0.0148)	(0.0149)	(0.0369)	(0.0368)
FTA=4		-0.257***	-0.257***	-0.291***	-0.300***
		(0.0176)	(0.0176)	(0.0420)	(0.0420)
FTA=5		-0.349***	-0.349***	-0.436***	-0.450***
		(0.0179)	(0.0179)	(0.0423)	(0.0423)
FTA=6		-0.321***	-0.322***	-0.425***	-0.442***
		(0.0289)	(0.0289)	(0.0696)	(0.0694)
Male*FTA=1				-0.0358	-0.0459
				(0.0272)	(0.0270)
Male*FTA=2				0.0192	0.0178
				(0.0268)	(0.0265)
Male*FTA=3				0.0339	0.0262
				(0.0298)	(0.0295)
Male*FTA=4				0.00563	-0.000708
				(0.0374)	(0.0372)
Male*FTA=5				0.0707	0.0705
				(0.0380)	(0.0377)
Male*FTA=6				0.0898	0.0850
				(0.0715)	(0.0708)
Constant	0.500***	0.664***	0.651***	0.693***	0.674***
	(0.0127)	(0.0106)	(0.0153)	(0.0245)	(0.0266)
Observations	9881	9881	9881	9881	9877
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: no evidence of intercept bias (chi-sq = 1.386, p-value = 0.239) or slope bias (chi-sq = 7.242, p-value = 0.203).

TABLE A-9. Calibration Results: PSA NCA Risk Scale (Black compared to White)

	(1)	(2)	(3)	(4)	(5)
Black	-0.0156		0.00600		
	(0.0122)		(0.0121)		
NCA=2		-0.150***	-0.151***	-0.156***	-0.141***
		(0.0238)	(0.0239)	(0.0336)	(0.0335)
NCA=3		-0.213***	-0.214***	-0.231***	-0.206***
		(0.0225)	(0.0225)	(0.0305)	(0.0306)
NCA=4		-0.274***	-0.275***	-0.313***	-0.267***
		(0.0219)	(0.0221)	(0.0309)	(0.0315)
NCA=5		-0.340***	-0.341***	-0.363***	-0.325***
		(0.0254)	(0.0255)	(0.0365)	(0.0370)
NCA=6		-0.395***	-0.396***	-0.442***	-0.404***
		(0.0281)	(0.0282)	(0.0423)	(0.0429)
Black*NCA=1				-0.0430	-0.0348
				(0.0385)	(0.0384)
Black*NCA=2				-0.0205	-0.0274
				(0.0292)	(0.0291)
Black*NCA=3				-0.000383	-0.00438
				(0.0241)	(0.0240)
Black*NCA=4				0.0328	0.0275
				(0.0228)	(0.0227)
Black*NCA=5				0.00862	0.00154
				(0.0345)	(0.0344)
Black*NCA=6				0.0452	0.0431
				(0.0429)	(0.0427)
Constant	0.445***	0.666***	0.664***	0.684***	0.593***
	(0.00927)	(0.0190)	(0.0197)	(0.0250)	(0.0278)
Observations	6763	6763	6763	6763	6761
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: no evidence of intercept bias (chi-sq = 0.248, p-value = 0.619) or slope bias (chi-sq = 4.753, p-value = 0.447).

TABLE A-10: Calibration Results of PSA NCA Scale (Latinx compared to White)

	(1)	(2)	(3)	(4)	(5)
Latinx	0.000883		-0.0309*		
	(0.0138)		(0.0135)		
NCA=2		-0.157***	-0.156***	-0.156***	-0.146***
		(0.0226)	(0.0226)	(0.0335)	(0.0335)
NCA=3		-0.218***	-0.221***	-0.231***	-0.216***
		(0.0214)	(0.0215)	(0.0303)	(0.0305)
NCA=4		-0.299***	-0.302***	-0.313***	-0.291***
		(0.0218)	(0.0218)	(0.0308)	(0.0315)
NCA=5		-0.348***	-0.353***	-0.363***	-0.350***
		(0.0273)	(0.0274)	(0.0363)	(0.0370)
NCA=6		-0.439***	-0.446***	-0.442***	-0.434***
		(0.0331)	(0.0333)	(0.0421)	(0.0430)
Latinx*NCA=1				-0.0430	-0.0445
				(0.0340)	(0.0339)
Latinx*NCA=2				-0.0416	-0.0477
				(0.0300)	(0.0300)
Latinx*NCA=3				-0.0212	-0.0291
				(0.0264)	(0.0264)
Latinx*NCA=4				-0.0205	-0.0256
				(0.0277)	(0.0276)
Latinx*NCA=5				-0.0219	-0.0184
				(0.0449)	(0.0450)
Latinx*NCA=6				-0.0682	-0.0635
				(0.0622)	(0.0621)
Constant	0.445***	0.661***	0.678***	0.684***	0.630***
	(0.00929)	(0.0170)	(0.0184)	(0.0249)	(0.0286)
Observations	5265	5265	5265	5265	5263
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: evidence of intercept bias (chi-sq = 5.223, p-value = 0.022); no evidence of slope bias (chi-sq = 0.932, p-value = 0.968).

TABLE A-11: Calibration Results of PSA NCA Scale (Male compared to Female)

	(1)	(2)	(3)	(4)	(5)
Male	-0.0619***		-0.0396**		
	(0.0137)		(0.0134)		
NCA=2		-0.158***	-0.157***	-0.108**	-0.106**
		(0.0175)	(0.0175)	(0.0395)	(0.0395)
NCA=3		-0.223***	-0.220***	-0.180***	-0.166***
		(0.0167)	(0.0167)	(0.0383)	(0.0383)
NCA=4		-0.293***	-0.290***	-0.288***	-0.267***
		(0.0165)	(0.0165)	(0.0398)	(0.0401)
NCA=5		-0.354***	-0.353***	-0.368***	-0.352***
		(0.0200)	(0.0200)	(0.0463)	(0.0465)
NCA=6		-0.413***	-0.409***	-0.265***	-0.250***
		(0.0230)	(0.0230)	(0.0630)	(0.0632)
Male*NCA=1				-0.00712	-0.0121
				(0.0328)	(0.0327)
Male*NCA=2				-0.0673*	-0.0684*
				(0.0295)	(0.0294)
Male*NCA=3				-0.0574*	-0.0615*
				(0.0272)	(0.0271)
Male*NCA=4				-0.0122	-0.0110
				(0.0291)	(0.0290)
Male*NCA=5				0.00907	0.00821
				(0.0395)	(0.0395)
Male*NCA=6				-0.173**	-0.173**
				(0.0592)	(0.0591)
Constant	0.498***	0.673***	0.704***	0.679***	0.616***
	(0.0126)	(0.0135)	(0.0171)	(0.0291)	(0.0309)
Observations	9881	9881	9881	9881	9877
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release (including transfer to another court, court own-release, etc.). Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: evidence of intercept bias (chi-sq = 8.698, p-value = 0.003); no evidence of slope bias (chi-sq = 9.771, p-value = 0.082).

TABLE A-12. Calibration Results of PSA New Violent Criminal Activity Flag (Black compared to White)

	(1)	(2)	(3)	(4)	(5)
Black	-0.0775***		-0.0675***	-0.0760***	-0.0757***
	(0.00973)		(0.00966)	(0.0103)	(0.0103)
NVCA=1		-0.187***	-0.178***	-0.224***	-0.213***
		(0.0138)	(0.0138)	(0.0240)	(0.0241)
Black*NVCA=1				0.0688*	0.0651*
				(0.0294)	(0.0293)
Constant	0.830***	0.829***	0.867***	0.872***	0.853***
	(0.0107)	(0.00514)	(0.00744)	(0.00771)	(0.0109)
Observations	6761	6763	6763	6763	6761
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release. Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: evidence of intercept bias (chi-sq = 48.737, p-value = 0.00) and slope bias (chi-sq = 5.484, p-value = 0.019).

TABLE A-13: Calibration Results of PSA NVCA Flag (Latinx compared to White)

	(1)	(2)	(3)	(4)	(5)
Latinx	-0.00233		-0.00362	-0.00231	-0.00860
	(0.00989)		(0.00972)	(0.0101)	(0.0102)
RANVCA Dummy		-0.230***	-0.230***	-0.224***	-0.212***
		(0.0174)	(0.0175)	(0.0215)	(0.0216)
Latinx*NVCA=1				-0.0176	-0.0162
				(0.0369)	(0.0369)
Constant	0.825***	0.870***	0.872***	0.872***	0.849***
	(0.0105)	(0.00504)	(0.00679)	(0.00691)	(0.0106)
Observations	5263	5265	5265	5265	5263
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release. Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: no evidence of intercept bias (chi-sq = 0.139, p-value = 0.709) or slope bias (chi-sq = 00.226, p-value = 0.634).

TABLE A-14: Calibration Results of PSA NVCA Flag (Male compared to Female)

	(1)	(2)	(3)	(4)	(5)
Male	0.0138		0.0170	0.0137	0.0151
	(0.0106)		(0.0105)	(0.0111)	(0.0111)
RANVCA Dummy		-0.203***	-0.203***	-0.230***	-0.218***
		(0.0119)	(0.0119)	(0.0318)	(0.0319)
Male*NVCA=1				0.0313	0.0292
				(0.0343)	(0.0343)
Constant	0.808***	0.843***	0.828***	0.831***	0.808***
	(0.00977)	(0.00406)	(0.00971)	(0.0102)	(0.0121)
Observations	9881	9881	9881	9881	9877
Pretrial release controls	No	No	No	No	Yes

Note: Model 5 includes controls for release type: bail, own-release (no active supervision), own-release (minimal supervision), Assertive Case Management/electronic monitoring, and other pretrial release. Standard errors in parentheses. * p<0.05 ** p<0.01 *** p<0.001. Likelihood ratio test: no evidence of intercept bias (chi-sq = 2.619, p-value = 0.106) or slope bias (chi-sq = 0.83, p-value = 0.362).

Predictive Bias and Fairness

Tables A-12 – A14 provide the outcome error rates for the NCA and FTA scales if we used "5" as a cut-off instead of "4."

TABLE A-12 Positive Predictive Value Rate by Race/Ethnicity

	POSITIVE PREDICTIVE VALUE					
	BLACK		LATINX		WHITE	
	PPV	N	PPV	N	PPV	N
FTA	0.64	623	0.73	230	0.69	537
NCA	0.69	842	0.74	263	0.71	535

Note: Positive predictive value is the percentage of people classified by the PSA as high risk who had a pretrial failure. N is the group total (denominator)

TABLE A-13. False Positive Rate by Race/Ethnicity

	FALSE POSITIVE					
	BLACK		LATINX		WHITE	
	FP	N	FP	N	FP	N
FTA	0.11	2,062	0.05	1,257	0.12	1,325
NCA	0.16	1,675	0.06	1,071	0.12	1,275

Note: False positive rate is the percentage of people who did not have a pretrial failure but had been classified as high risk by the PSA. N is the total number that did not have a pretrial failure.

TABLE 21. False Negative Rate by Race/Ethnicity

	FALSE NEGATIVE					
	BLACK		LATINX		WHITE	
	FN	N	FN	N	FN	N
FTA	0.78	1,837	0.85	1,144	0.76	1,539
NCA	0.74	2,224	0.85	1,330	0.76	1,589

Note: False negative rate is the percentage of people had a pretrial failure but had not been classified as high risk by the PSA. N is the total number that did have a pretrial failure.

Appendix A-2: Data Sources

This validation study uses longitudinal data from four sources:

- **San Francisco District Attorney:** covers all adult criminal cases in San Francisco since 2008. This dataset provides information on arrest date and charge for most misdemeanor and felony arrests; filed charge; case disposition (conviction, dismissal, etc.); failures to appear; and general demographic information. We utilize this data set to categorize the booked offense and to measure the appearance rate.
- **San Francisco Sheriff's Office:** provides information from all bookings and releases from the County Jail since 2010. This data is used to measure in-custody time, releases on bail, and to determine sex, race and ethnicity.
- **San Francisco Pretrial Diversion Project (SF Pretrial):** beginning in April 2016, this dataset includes the weighted score for each pretrial risk factor and the release recommendation. The dataset also includes the release decision at pre-arraignment and arraignment for individuals who were presented, including the date of the event, the judge, and the judicial decision. We use this data to determine the recommended level of supervision for each individual record – using both the raw PSA risk scores and the DMF recommendation; release decision at pre-arraignment or arraignment; and terminations for failing to appear at a court hearing.
- **California Department of Criminal Justice's Automated Criminal History System:** these data include information on individual-level arrests, charges, and case dispositions, with offense and status (infraction, misdemeanor, felony); sentence duration and location; and date and county of arrests and dispositions. We use this data to measure the safety rate for any new offense and any new violent offense.

Appendix A-3: Empirical Strategy

This appendix provides detailed information about the empirical strategies employed to answer our primary research questions. It is organized into three sections: Predictivity of the PSA; Assessment of the Implementation of the PSA; and Assessment of Predictive Fairness and Bias.

Predictivity of the PSA

We evaluate the predictive accuracy of each PSA risk score using Area Under the Curve (AUC) Receiver Operator Characteristics (ROC) estimates. The AUC-ROC value reflects the ability of a linear scale, like the PSA, to distinguish between success and non-successes.[19] We look at the AUC-ROC estimates for the full population and individually for the following sub-groups: (1) male; (2) female; (3) Black; (4) Latinx; and (5) White. Subsequent analyses look separately at relationship between the PSA risk scales and the risk factors.

We begin with the risk scales, hypothesizing that the likelihood of success – appearing at all court hearings and remaining arrest-free during the pretrial period – will decrease as we move up the risk scale. Specifically, we estimate the model in equation 1 separately for each risk scale,

$$(1)\ Y_i = \beta_0 + \beta_1(\text{RiskScale}_i) + \varepsilon_i,$$

where Y_i is the realized outcomes for individual i (e.g. appearance rate) and B_1 represents the coefficient on each level of the risk scale (with one as the omitted category for FTA and NCA).

We then look at the predictivity of each individual PSA risk factor (listed in Table 1) on the specific outcome (FTA, NCA, and NVCA). We run a simple bivariate logistic regression to assess the individual risk factor's correlation with the specified outcome. We estimate the model in equation 2 for each risk factor,

$$(2)\ Y_i = \beta_0 + \beta_1(\text{RiskFactor}_i) + \varepsilon_i$$

Where Y_i is the realized outcomes for individual i for each of the three PSA risk factor groups and B_1 represents the beta coefficient on each individual risk factor included in the overall risk score (denoted by RiskFactor_i). These results show whether some risk factors are more or less correlated with the outcomes than others.

Next, we run a multivariate logistic regression to test the combination of all risk factors on the specified outcome. In model 3, we examine which risk factor(s) have the greatest correlation with the outcome measure:

$$(3)\ Y_i = \beta_0 + \beta_1(\text{Pending}_i) + \beta_2(\text{PriorConviction}_i) + \beta_3(\text{PriorFTA}_i) + \beta_4(\text{PriorFTA2}_i) + \varepsilon_i,$$

where Y_i is the overall appearance rate for individual i, Pending is a binary measure if they had a pending charge at arrest, PriorConviction is a binary measure of any prior conviction, PriorFTA is a categorical variable (0 if no FTA in prior 2 years, 2 if one, and 4 if 2 or more), and PriorFTA2 is a binary measure of any FTAs that are older than two years.

[19] An AUC-ROC of 1 would indicate the PSA perfectly predicts the actual outcome in 100% of cases; and AUC-ROC of 0.5 would indicate that the PSA performs no better than a coin toss. A meta-analysis by Demarais and Singh (2013) and Damarais, Johnson, and Singh (2016) suggests that AUC-ROC values less than 0.54 indicate poor predictivity, 0.55 to 0.63 are fair, 0.64 to 0.71 indicate good predictivity, and anything higher than 0.71 is considered excellent.

Assessment of Predictive Fairness and Bias

To assess predictive utility and predictive fairness, we use the calibration condition, which measures the extent to which "…a given score will have the same meaning regardless of group membership" (Mohan, Skeem, and Lowenkamp, 2016: 193). We then compare the different regression lines using likelihood ratio tests for each subgroup, notably the starting point (intercept) and trajectory (slope) to assess predictive bias. Statistically significant differences in the regression lines between groups is an indicator of predictive bias.

The moderator regression technique involves four models run separately for each demographic subgroup of interest: (1) male; (2) female; (3) Black; (4) Latinx and (5) White. The four equations are summarized below, using likelihood of appearing at all hearings as the sample outcome.

First, we estimate the outcome using a binary indicator of the demographic subgroup in equation 4:

$$(3)\quad Y_{is} = \beta_0 + \beta_1(\mathrm{Subgroup}_s) + \varepsilon_{is},$$

where Y_{is} is the share of individuals that appear to all hearings during the pretrial period and B_1 is the beta coefficient on an indicator variable for the demographic subgroup of interest (i.e. 1 if male, 0 if not) (denoted by $\mathrm{Subgroup}_s$). The model presents the correlation between subgroup membership and the outcome variables, without controlling for any of the other factors that may contribute to the outcome.

Second, we run a fitted model with PSA risk score for each outcome separately and without demographic controls (equation 5).

$$(4)\quad Y_i = \beta_0 + \beta_1(\mathrm{RiskScore}_i) + \varepsilon_i,$$

where Y_i is the share of cases that appear to all hearings during the pretrial period and B_1 represents the regression coefficient on each of the six RA-FTA risk scores (denoted by $\mathrm{RiskScore}_i$). This model provides the correlation between the risk score and the outcome it is trying to predict.

Third, we use a multivariate model to jointly estimate the correlations between the demographic subgroup and PSA risk score (equation 6):

$$(6)\ Y_{is} = \beta_0 + \beta_1(\mathrm{RiskScore}_i) + \beta_2(\mathrm{Subgroup}_s) + \varepsilon_{is}$$

where Y_{is} is the share of cases that appear to all hearings during the pretrial period for each demographic subgroup and B_1 represents each of the six RA-FTA risk scores (denoted by $\mathrm{RiskScore}_i$), and B_2 is the coefficient on an indicator variable for each subgroup s (denoted by $\mathrm{Subgroup}_s$). We compare the coefficients of this model to the results of equations 4 and 5 to measure how simple correlations change with the inclusion of controls.

Our final multivariate model estimates both the relationship between the demographic subgroup indicators, PSA risk score, and their interaction (equation 7):

$$(7)\ Y_{is} = \beta_0 + \beta_1(\mathrm{RiskScore}_i) + \beta_2(\mathrm{Subgroup}_s) + \beta_3(\mathrm{Subgroup}_s) * (\mathrm{RiskScore}_i) + \varepsilon_{is},$$

where Y_{is} is the share of cases that appear to all hearings during the pretrial period and B_1 represents the coefficient on each of the six RA-FTA risk scores (denoted by $\mathrm{RiskScore}_i$), B_2 is an indicator variable for each subgroup (denoted by $\mathrm{Subgroup}_s$), and B_3 is the interaction of subgroup and PSA risk score. The interaction term tests to what extent the likelihood of a pretrial success (i.e. appearing at all

hearings) is a function of the demographic subgroup and the PSA risk score together – i.e. are PSA scores more or less predictive of outcomes for different subgroups?

Finally, we present subgroup mean differences per weighted risk score and PSA risk score to help the PSA Workgroup assess disparate impact. We present the Positive Predictive Value (PPV) which is the share of persons that are defined as high risk in each respective category and who did have new (violent) criminal activity or who failed to appear. Secondly, we present the False Positive Rate (FPR) for this same group – the share that were rated high risk but did not actually have new (violent) criminal activity or fail to appear. Lastly, we present the False Negative Rate (FNR) -- the share that were did have a pretrial failure and were assessed as low risk These rates will be calculated for the full population and then separately for Males, Females, Blacks, Latinx and Whites.

In order to measure the PPV, FPR, and FNR for the NCA and FTA Risk scores, we transform the scale into a binary measure: 1-4 as "not high risk" and 5-6 as "high risk." We also test whether our results are sensitive to this selected cut-off. These measures, on their own, are not indicators of test bias. Rather these differences may be an inequitable consequence of the tool and should be presented to the PSA Workgroup.