FRESNO
PSA VALIDATION
August 2025 Report

Marijana Kotlaja, Ph.D.
Senior Associate

Rebecca Stone, Ph.D.
Senior Associate

Kevin Kuehmeier
Senior Associate

Brian Lovins, Ph.D.
President

JSP
JUSTICE SYSTEM PARTNERS

Table of Contents

Introduction

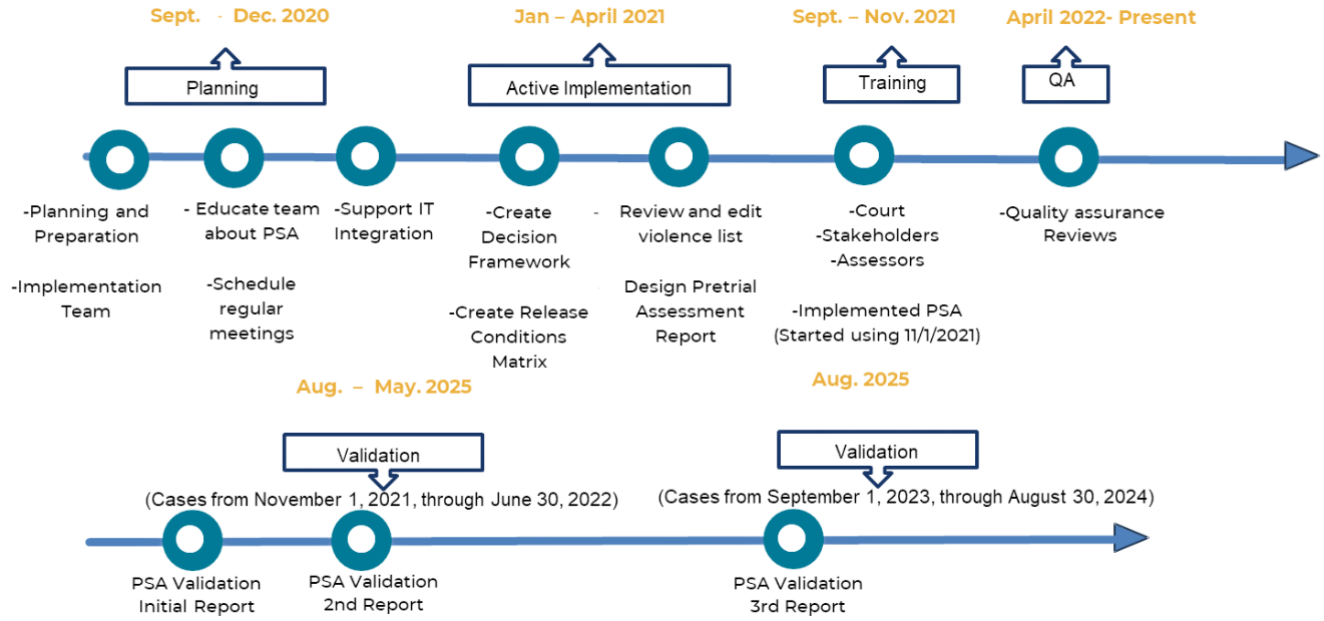
Fresno County, California, is in California's central valley region, where it is surrounded by national parkland and large-scale agricultural areas. Fresno County is demographically diverse, with a population that is approximately 53% Hispanic or Latino, 30% white (non-Hispanic), 8% Asian, and 5% Black (U.S. Census Bureau, 2020). The county also experiences significant socioeconomic challenges, with a poverty rate of 20.5%, notably higher than the California state average. Risk assessments instruments like the Public Safety Assessment (PSA) must be regularly validated against the local population to ensure they achieve their intended purpose: diverting low-risk individuals from unnecessary incarceration. Therefore, ongoing evaluation of the PSA's validity across different demographic groups in diverse jurisdictions like Fresno County is essential to ensure that efforts to improve pretrial justice do not inadvertently reinforce systemic inequities.

The Fresno pretrial division is located within the Fresno County Probation Department. The purpose of the pretrial office is to promote due process to those accused of crime(s), assist the Fresno Superior Court in making release decisions, maintain the integrity of the judicial process by ensuring individuals appear for court, and protect victims, witnesses, and the community. Pretrial Services uses the PSA to assist judicial officers in identifying persons who are unlikely to accrue new arrests and have a high likelihood of appearing for all court appointments, and therefore can be released into the community pending resolution of their Court case(s). The Court then decides whether the person is suitable for Pretrial Release and if any conditions should be imposed.

This report will assess the predictive validity of the PSA across two defined time periods. The combined validation period provides critical insights into the PSA's performance after the system had been fully operational for nearly two years, allowing for assessment of long-term effectiveness, identification of any emerging trends or patterns, and evaluation of how well the tool continues to predict pretrial outcomes. The temporal gap between validation periods also enables examination of system stability and consistency over time, ensuring that the PSA maintains its predictive validity as case volumes and operational practices evolve.

The diagram below provides a description of the PSA implementation and validation timeline. The first phase focused on foundational planning activies, including the formation of the implementation team, educating team members about the PSA framework, scheduling regular coordination meetings, and establishing IT integration support structures. During the active implementation phase, the team developed core operational components including a decision framework, a release conditions matrix, a list of offenses considered "violent crimes" for the purpose of scoring the PSA, and the design of the pretrial assessment report format. Court personnel, stakeholders and assessors were then trained in the PSA system and implemented it on November 1, 2021. Quality assurance began in April of 2022 and is ongoing to ensure PSA effectiveness and accuracy.

Figure 1. Timeline of Fresno PSA Validation



The first and second PSA validations for Fresno County examined cases processed from November 1, 2021 through June 30, 2022, which resulted in 3,408 cases. This report will utilize the cases from this initial validation period and add cases from a later validation period, September 1, 2023 through August 20, 2024.

PSA Instrument Description

The PSA is administered through a series of administrative data reviews (based on previous criminal history, failure to appear, and factors associated with the current offense) in which three scales are produced: Failure to Appear (FTA), New Criminal Arrest (NCA), and New Violent Criminal Activity (NVCA). Each scale is comprised of individual items scored and summed and then plotted on the scale.

The **Failure to Appear (FTA) scale** is designed to predict the likelihood that people will return to court (versus "fail to appear") for pretrial hearings. The **FTA Scale** is comprised of five individual items that total seven points. Three of the items (Pending charge at time of the offense, prior conviction, and prior failure to appear older than two years) are scored "0" or "1" while the fourth item (Prior failure to appear pretrial within the past two years) is scored a "0" for no priors, "2" for one prior FTA, and "4" for two or more prior FTAs within the past two years. Once the raw score is tabulated, it is then applied to the FTA Scale and a subsequent "score" is produced.

The **New Criminal Arrest (NCA) scale** provides decisionmakers with information regarding the likelihood that an individual will be arrested for a new crime while on pretrial release. The **NCA scale** is also plotted on a six-point scale that is comprised of up to 13 points for

the individual raw score. The first item, age at current arrest is scored as a 0 for individuals who were 23 or older while those that are 22 or younger at the time of their arrest are scored a "2". Individuals are scored as a "1" for any prior conviction (I.e. misdemeanor or felony) and scored a 0 if they have no prior convictions. If a person has a pending charge at the time of the new arrest, they are scored a "3" while those with a prior sentence to incarceration is scored a "2". The sixth item, prior violent conviction is scored as a "0" for no priors, a "1" for one or two violent prior convictions, and a "2" for three or more prior violent convictions. And the seventh and final item for the NCA scale is based on the number of prior failures to appear within the past two years. If the person has no prior FTA in the past year, a score of "0" is given, one prior FTA results in a score of "1", while two or more prior FTAs is scored as a "2". Once tabulated and summed, the final score is plotted on the NCA scale and a subsequent "scale score" is produced.

In addition to the FTA and NCA scales, the PSA also provides a score associated with **New Violent Criminal Activity (NVCA)**. The NVCA raw score ranges from "0" to "7" across five items. Current violent offense and ≤20 years of age, pending charge at the time of the offense, and prior conviction (for any reason) are all scored 0 (no) or 1 (yes). If the person's current offense is violent, they receive a score of 2 for this item. And for the last item, prior violent conviction, a person with no prior violent convictions is scored a "0", a person with one or two prior violent convictions is scored a "1", and a person with three or more violent convictions is scored a "2". Unlike the other scales, the **NVCA is designed as a violence flag,** with people who score a "3" or less identified as not having the violence flag and those scoring "4" or above are identified with a violence flag.

In the table below, blue boxes indicate that the risk factor item is included in that scale. For example, the FTA scale includes items 3, 5a, 7, and 8. The item scores for the FTA and NCA scales are weighted to produce an individual scale score of 1-6. The NVCA items are weighted to form a yes/no outcome.

Figure 2. PSA Scale Items

	Scales		
	FTA	**NCA**	**NVCA**
Risk Factor			
1. Age at current offense		■	
2. Current violent offense			■
2a. Current violent offense and ≤20 years			■
3. Pending charge at time of offense	■	■	■
4. Prior misdemeanor conviction		■	
5. Prior felony conviction		■	
5a. Prior conviction	■		■
6. Prior violent conviction (incl. #)		■	■
7. Prior FTA pretrial in past 2 years (incl. #)	■	■	

8. Prior FTA pretrial older than 2 years			
9. Prior sentence to incarceration			
*Boxes shaded **blue** indicate that the item is included in the scale; boxes shaded gray indicate that the item is **not** included in the scale.*			

Validation Method

Data Analysis

Sample

This validation analysis encompasses two distinct time periods with different assessment criteria due to operational constraints.

Period 1: November 2021 - August 2023

A total of 14,612 PSA assessments were completed during this period. Due to staffing limitations from November 2021 through June 2022, assessment criteria were more restrictive, with PSAs completed only for individuals booked on open charges (including felony warrants) while excluding those with misdemeanor warrants only or those under any form of formal supervision. From July 2022 through August 2023, criteria expanded to include all individuals in Fresno County Jail booked for open charges (including all warrants), excluding only those remanded by the Court, those with Federal holds, and those set for sentencing.

Period 2: September 2023 - August 2024

An additional 12,373 assessments were completed during this period using the expanded criteria established in July 2022.

Validation Sample Criteria

To be included in the validation analysis, cases must meet two requirements: (1) the case must be resolved, and (2) the individual must have been released from jail at any point prior to disposition.

Sample Selection Results

From Period 1's 14,612 assessments, 3,978 cases (27.2%) met the validation criteria. 571 cases (14.3%) were subsequently excluded during sensitivity and diagnostic analyses[1], resulting in 3,407 cases from Period 1. From Period 2's 12,373 assessments, 2,419 (19.5%) cases met the validation criteria and were retained for analysis. The combined validation sample totaled 5,826 cases across both time periods.

Outcome Measures

After cleaning and organizing the data, we conducted a series of analyses examining the relationships between each of the three PSA scales and their relevant outcomes: Failure to Appear (FTA), New Criminal Arrest (NCA), and New Violent Criminal Arrest (NVCA).

[1] Refer to the Initial Validation report for detailed explanation of sensitivity and diagnostic analysis exclusions.

The definition of "violent" crime varies between jurisdictions. In this case, violent crimes are defined in California Penal Code PC 667.5(c).

Table 1. Fresno Validation Outcome Measures, Definitions, and Values				
Outcome	*Abbreviation*	*Definition*	*"0" Value*	*"1" Value*
Failure to Appear	*FTA*	Any failure to appear	Did not fail to appear	One or more failures to appear
New Criminal Arrest	*NCA*	Any new criminal arrest under pretrial release	Did not have a new arrest while released pretrial	Had one or more new arrests while released pretrial
New Violent Criminal Arrest	*NVCA*	One or more new arrests is a violent crime	No new arrests or new arrests do not include a violent crime	One or more new arrests while released pretrial were for a violent crime

For the FTA outcome, the outcome of interest is any failure to appear, where "0" indicates the individual did not fail to appear and "1" indicates the individual had one or more failures to appear. Similarly, the outcome of interest for the NCA scale is defined as any new criminal arrests while under pretrial release, and this was also coded as "0" indicating the individual had no new arrests while on pretrial release and "1" indicates the individual has one or more new arrests while under pretrial release. If the individual had at least one new arrest, a separate variable indicates how many of those new arrests were considered violent crimes (range: 0-4). The New Violent Criminal Arrest (NVCA) outcome is defined as one or more new arrests include a violent arrest. If the individual had no new arrests while under pretrial release, this was coded as "0". If they had new arrests but they were only for nonviolent offenses, this was also coded as "0". If the individual had a new arrest and any of their new arrests was for a violent offense, then it was recoded as "1."

ROC/AUC Analyses

Receiver Operating Characteristics/Area Under the Curve (ROC/AUC) is used to determine the probability that a randomly selected individual from the data who failed (new arrest, new arrest for violent offense, or FTA) had a higher score than a randomly selected individual that did not fail. If the instrument (in this case the PSA) fails to identify any differences between individuals who failed and individuals who did not, then the AUC score is .500.

While there is no exact AUC score that suggests that an assessment is valid, the following is used as a general guide or best practice for interpreting the results of ROC/AUC analyses:

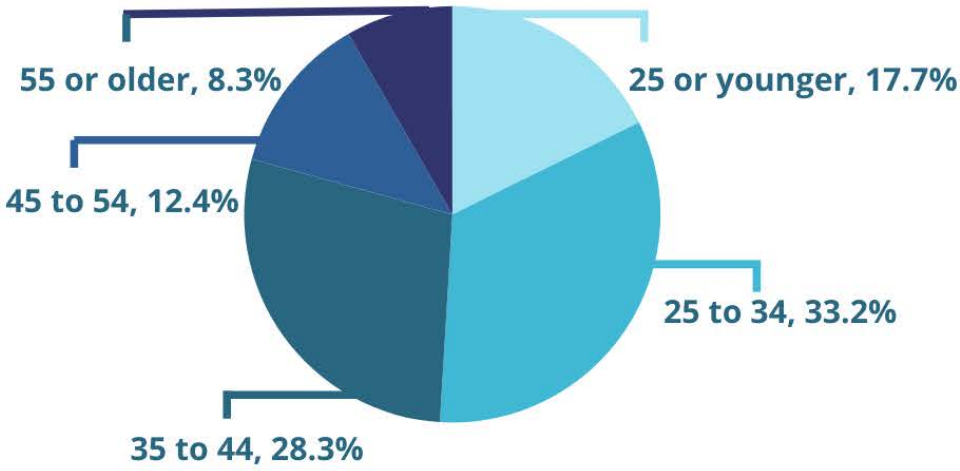
Table 2. ROC/AUC Analyses and Interpretations	
AUC Score	*Interpretation*
.54 or less	No evidence of validity
.55 to .63	Weak evidence
.64 to .70	Moderate evidence
.71 and above	Strong evidence

Sample Descriptives

The descriptive table below demonstrates a predominantly young, male, and Hispanic population. The sample skews toward younger adults, with the 25-34 age group representing the largest segment (33.2%), and over half of all individuals (50.9%) aged 34 or younger. Men comprise 81.4% of the sample compared to 18.6% women. Racially and ethnically, Hispanic or Latino individuals represent the majority at 59.9%, followed by white individuals (19.2%) and Black individuals (16.1%), with all other categories comprising less than 3% of the sample combined.

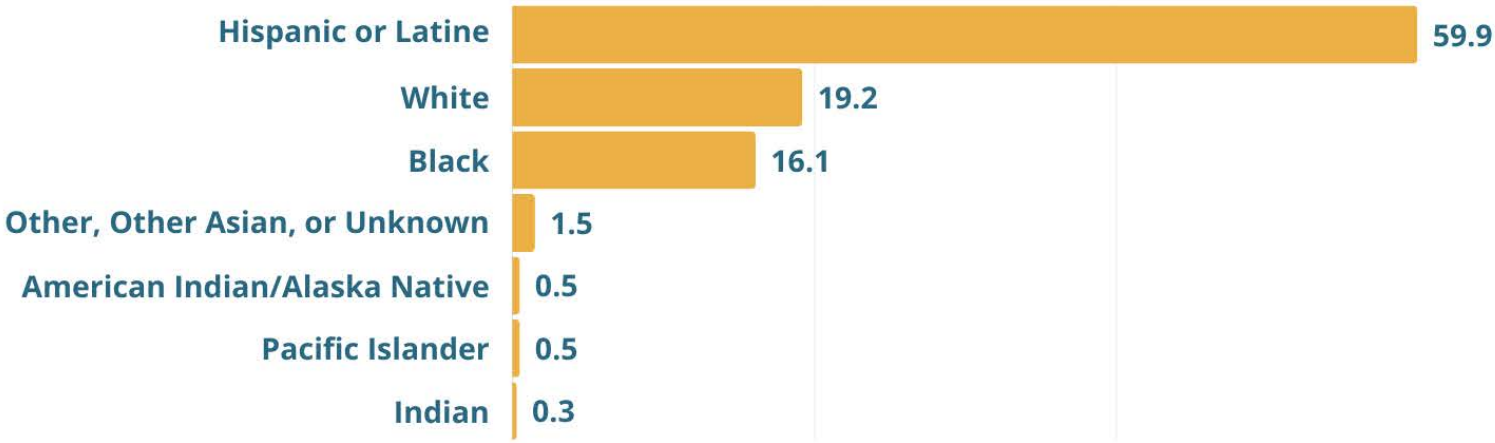
Full Sample Characteristics

Age at Arrest



Gender[2,3]

Race and Ethnicity[4]



[2] We use the term "sex" to present the categorization as captured by the agency (which uses the variable name "Gender" but categories of M, F, and T; it is not clear whether this variable is based on self-reported, perceived, or assigned identity). We recognize the demographic representation of our sample may be different when considering gender identity and could include representation from individuals who identify as non-binary or gender non-conforming. We believe it is important for administrative data to consider the full spectrum of identity to understand the individual experience and disparate outcomes more adequately. This includes disaggregating the "transgender" category into transgender women, transgender men, and transgender nonbinary or gender non-conforming people.

[3] One case was identified as a transgender person. The case could not be included in the validation analysis because a single case is not sufficient for drawing conclusions about the PSA's performance for transgender individuals.

[4] We report race as the mutually exclusive categories captured by the agency. We cannot determine individuals in the sample who identify as bi- or multiracial, or who identify in other ways than what is captured by the agency. We recognize self-reported racial identity is critical for accurately reporting the true demographic profile of the sample, the individual's experience, and any disparities.

PSA Assessment Descriptives

As previously described, the PSA is comprised of 3 measures: Failure to Appear (FTA), New Criminal Arrest (NCA), and New Violent Criminal Arrest (NVCA). The tables and charts below show the distribution of scale scores across the full validation sample of 5,827.

Scores for the items included in the FTA scale are summed and weighted into an FTA scale score, which can range from 1 (lowest likelihood of FTA) to 6 (highest likelihood of FTA).

Figure 3. FTA Scale Scores, Proportion of Total

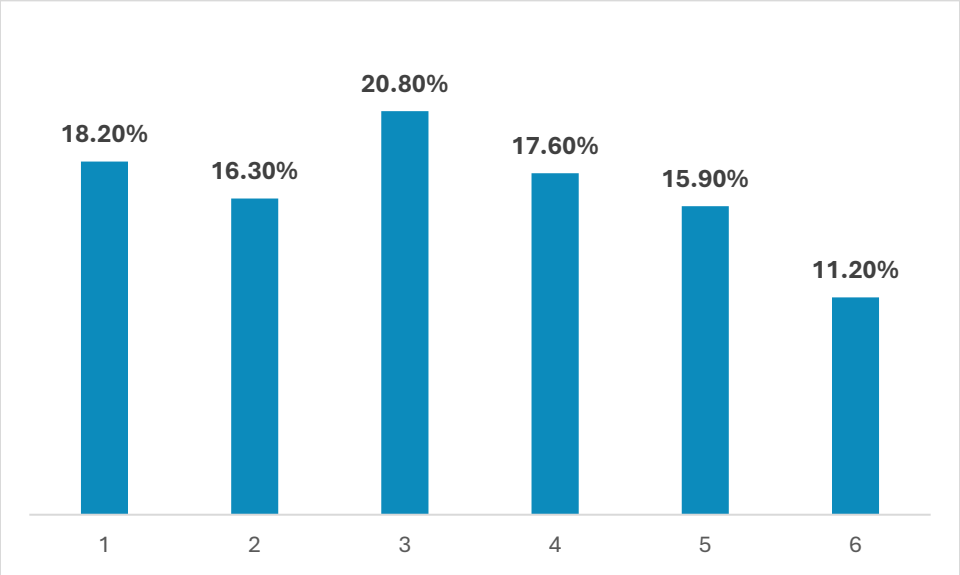


Table 3, FTA Scale Score for Sample		
	f	%
1	1059	18.2%
2	951	16.3%
3	1211	20.8%
4	1025	17.6%
5	927	15.9%
6	653	11.2%
Total	5,826	100%

Scores for the items included in the NCA scale are summed and weighted into an NCA scale score, which can range from 1 (lowest likelihood of NCA) to 6 (highest likelihood of NCA).

Figure 4. NCA Scale Scores, Proportion of Total

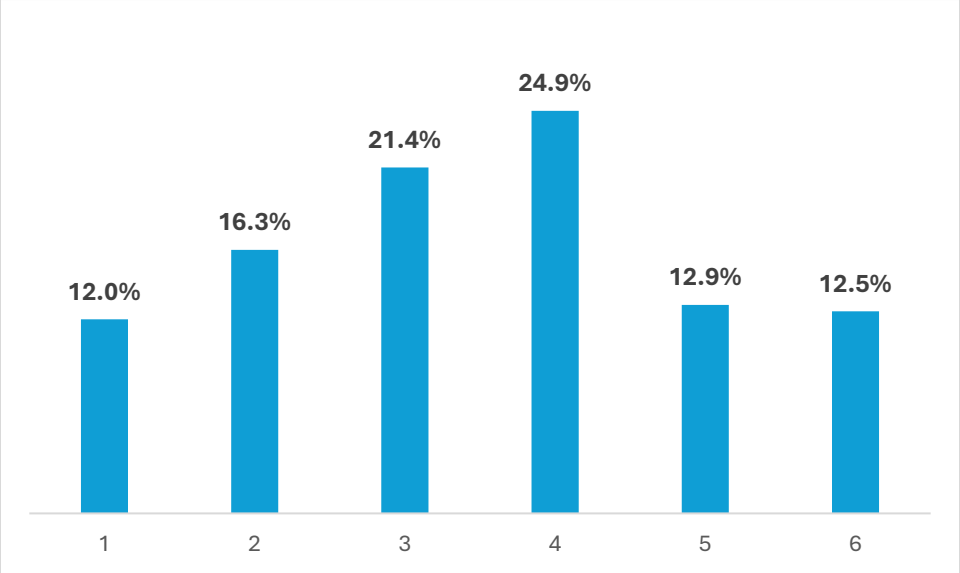


Table 4, NCA Scale Score for Sample		
	f	%
1	670	11.5%
2	868	14.9%
3	1288	22.1%
4	1480	25.4%
5	804	13.8%
6	716	12.3%
Total	5826	100%

The New Violent Criminal Arrest (NVCA) scale differs from the FTA and NCA scales in that it acts as a "flag," where raw scores of 0-3 receive a "No" (coded here as 0) and raw scores of 4 and above receive a "Yes" (coded as 1). This results in only two possible values in Figure 3.

Figure 5. NVCA Flag

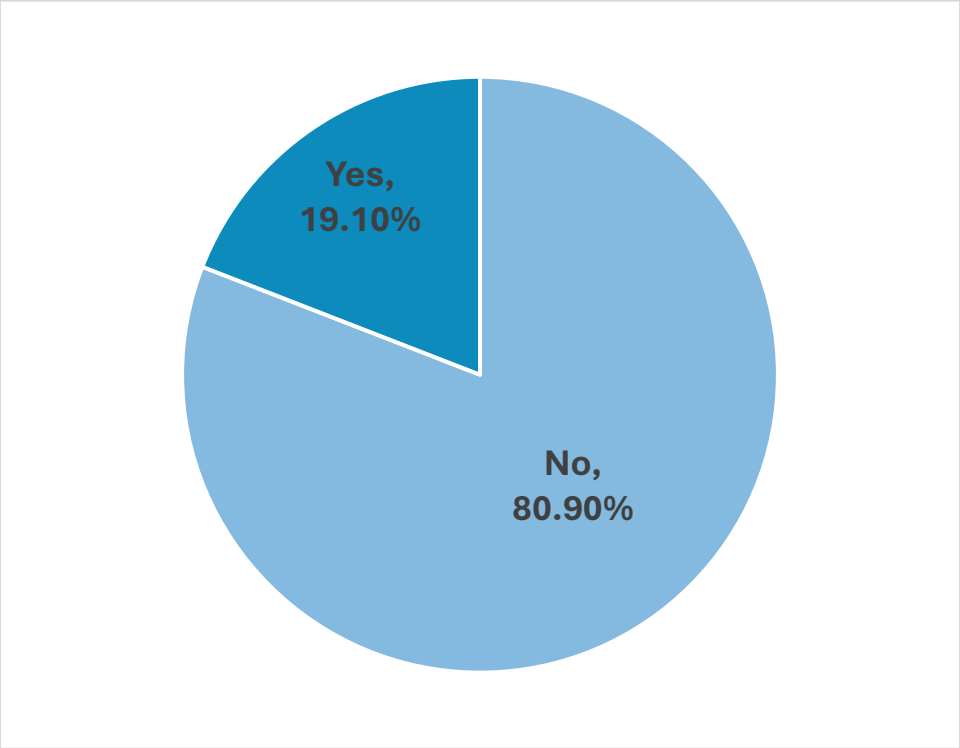


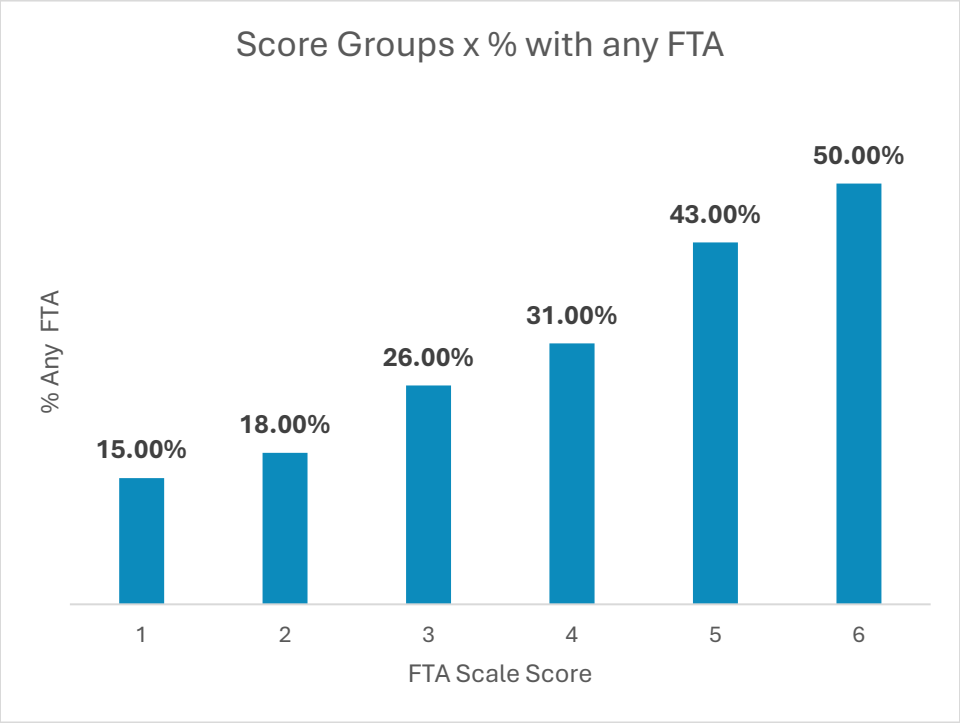
Table 5, NVCA Flag (Yes, No) for Sample		
	f	%
No	4714	80.90%
Yes	1112	19.10%
Total	5826	100%

PSA Scale Scores and Case Outcomes Analyses

Failure to Appear (FTA) Scores and Outcomes

To analyze whether the FTA scale score predicts any FTA outcome, we conducted an ANOVA test to compare the mean proportions of any FTA for each FTA scale score group.

Figure 6. FTA Scale Validation



The ANOVA results show significant differences between groups ($F = 85.18$, $p < .001$). The effect size ($\eta^2 = .068$) indicates a moderately strong relationship between the FTA scale score and FTA outcome. The ROC/AUC analysis produced a test result of .66 which indicates moderate evidence of validity for the whole sample.

We used a Sidak post hoc analysis to identify between-group differences. Individuals with FTA scores of 1 and 2 do not have significant differences in FTA outcomes ($p = .846$) but do differ significantly from individuals with higher scores (3-6). Groups 3 and 4 show no significant difference from each other ($p = .082$), while all other group comparisons are statistically significant. Individuals in the lower FTA groups (1-2) comprise 34.5% of the total sample but only 19.5% of those with any positive FTA outcomes, demonstrating the scale's ability to identify lower-risk individuals.

The graphic below provides a visual representation of where we see differences between the score groups, with significant differences shaded in dark blue and no difference shaded in light blue.

Table 6, Significant Differences in FTA Outcomes Between FTA Scale Scores

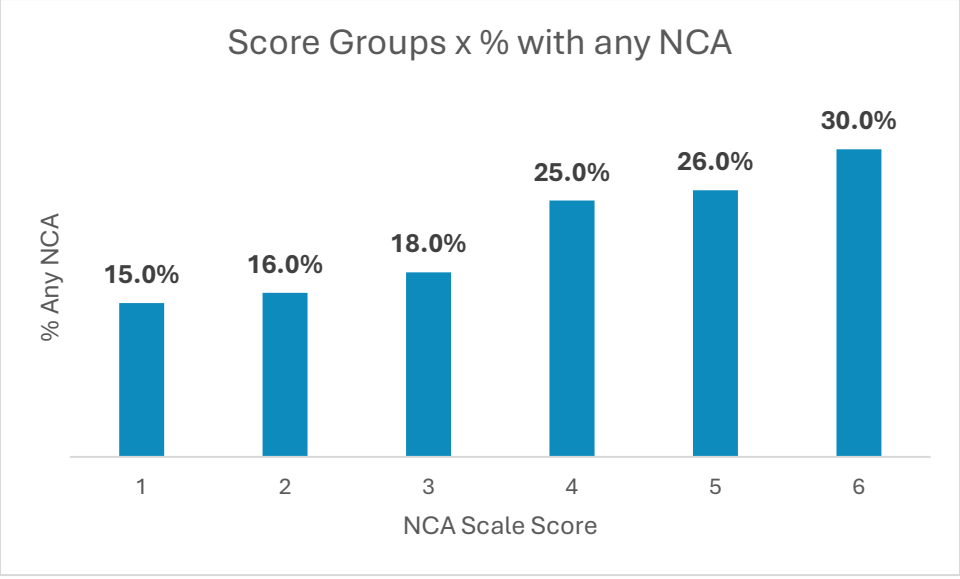
	1	2	3	4	5	6
1						
2						
3	*	*				
4	*	*				
5	*	*	*	*		
6	*	*	*	*	*	

Cells marked with a * indicate that these scale scores have significantly different FTA outcomes. For example, individuals who score a 1 do not have significantly different outcomes from individuals who score a 2. Individuals who score a 1 *do* have significantly different outcomes from those who score a 3, 4, 5, or 6.

New Criminal Arrest (NCA) Scores and Outcomes

The following analysis compares the mean proportion of people at each NCA scale level with any new criminal arrest (NCA) in the observation period.

Figure 7. NCA Scale Validation



The ANOVA results show significant differences between groups ($F = 18.207$, $p < .001$). The effect size ($\eta^2 = .015$) indicates a small relationship between the NCA scale score and any NCA outcome. The ROC/AUC analysis produced a test result of .583, which indicates weak

evidence of validity. This may be due to the small differences in outcomes between some score groups (e.g. score groups 1 and 2, and score groups 4 and 5).

The graphic below provides a visual representation of the Sidak post-hoc analysis, where we see significant differences between lower and higher scores, but failure to distinguish between neighboring scores (e.g., no difference between scores of 1, 2, and 3; no difference between scores of 4 and 5; and no difference between scores of 5 and 6).

Table 7, Significant Differences in NCA Outcomes Between NCA Scale Scores

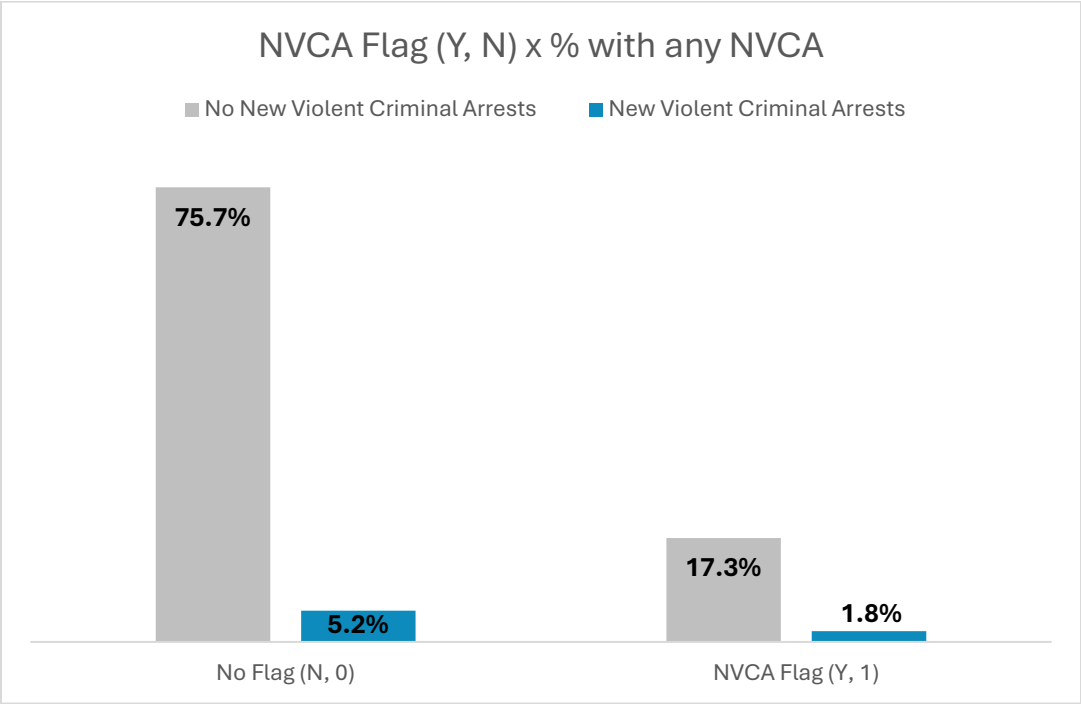
	1	2	3	4	5	6
1						
2						
3						
4	*	*	*			
5	*	*	*			
6	*	*	*	*		

Cells marked with a * indicate that these scale scores have significantly NCA outcomes. The analysis shows clear separation between lower-risk (scores 1-3) and higher-risk (scores 4-6) groups, but limited discrimination within these clusters, indicating weaker overall predictive validity compared to the FTA scale.

New Violent Criminal Arrest (NVCA) Flag and Outcomes

The New Violent Criminal Arrest (NVCA) scale acts as a "flag," where raw scores of 0-3 receive a "No" (coded here as 0) and raw scores of 4 and above receive a "Yes" (coded as 1), and therefore an ANOVA is no longer appropriate. To validate the NVCA flag, we used a chi-square (χ^2) test of independence.

Figure 8. NVCA Scale Validation

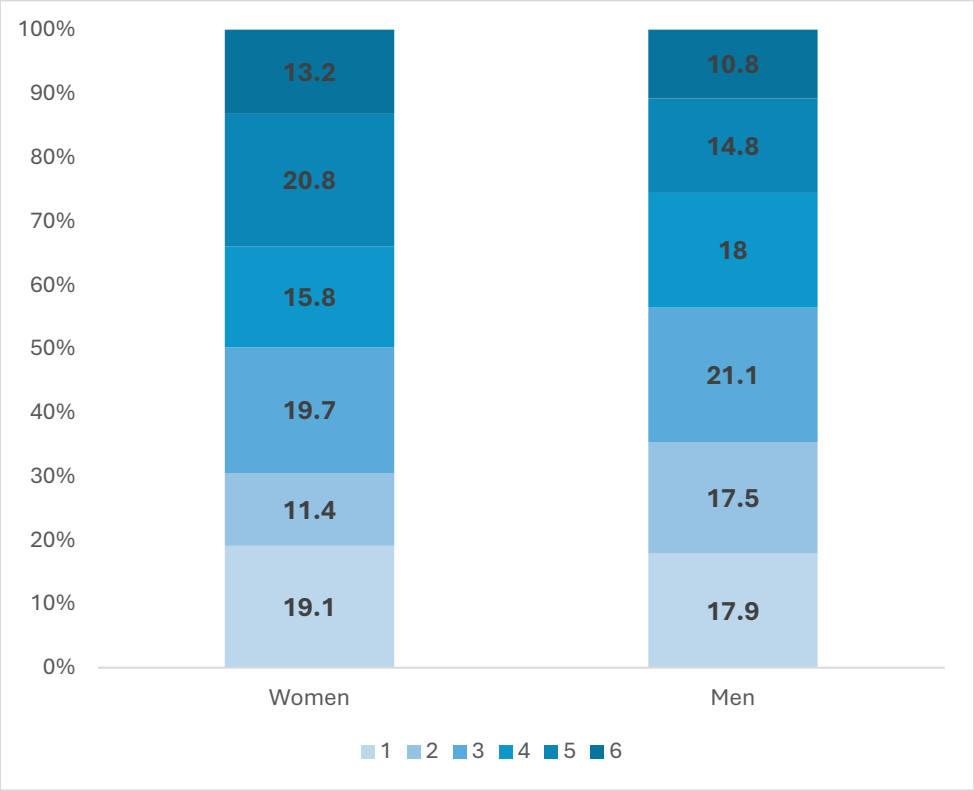


The chi-square analysis shows a significant association between NVCA flag status and new violent crime outcomes ($\chi^2 = 10.668$, $p = .001$). The effect size (Cramer's V = .043) indicates a small relationship between the NVCA flag and new violent crime arrest outcomes. This indicates that there is a statistically significant difference in the likelihood of new violent criminal arrests depending on whether the individual received the NVCA flag, but that those who received the flag (which should indicate higher risk of NVCA) actually had *lower* likelihood of NVCA. However, the overall difference is very small, and the number of new violent criminal offenses is also small. In practice, this may not translate to meaningful differences in supervision practices. This finding indicates that the NVCA flag is not functioning as intended for the Fresno County pretrial population.

Subgroup Analyses

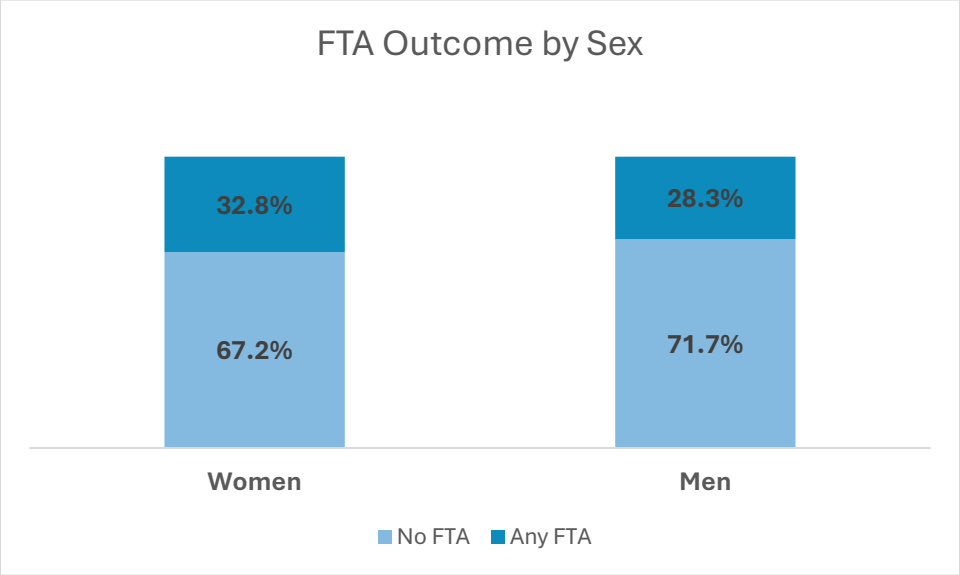
Sex

Figure 9. FTA Scale Score by Sex



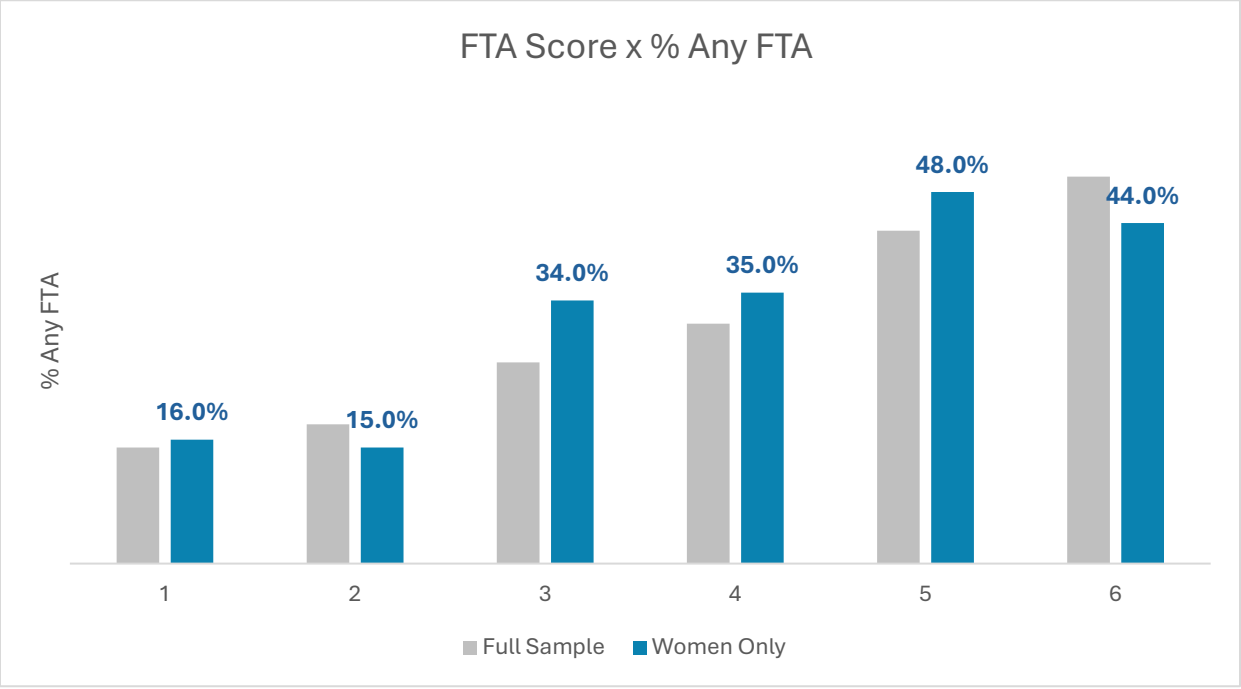
Proportionally, more women than men score a 1, 5 or 6 on the FTA scale ($\chi 2 = 52.56$, $p < .001$). Men are more likely than women to score a 2, 3, or 4. Some gender-neutral assessments are known to overclassify women, so the gender discrepancy at score 5 is worth further investigation to understand why women might be more likely to receive these higher scores and if this truly translates to higher likelihood of FTA.

Figure 10. FTA Scale Score by Sex



Women were slightly more likely to have any FTAs, with 32.6% of women having at least one FTA versus 28.3% of men. This is a statistically significant difference ($\chi 2 = 9.58$, $p = .008$). This suggests that the small differences between women's and men's FTA scale scores do translate into meaningful differences in their actual likelihood of failing to appear for a court hearing.

Figure 11. Subgroup Analysis: Full Sample vs Women Only

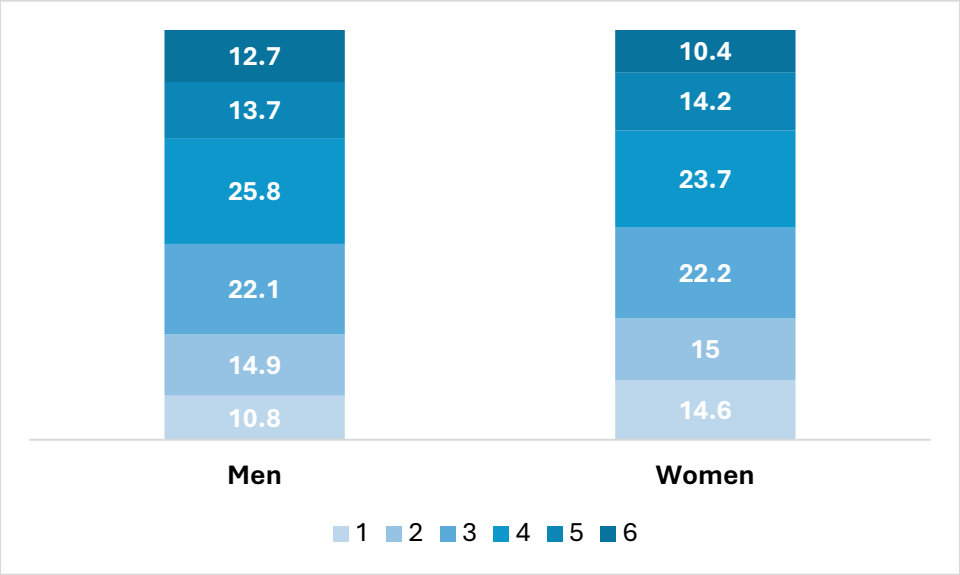


The subgroup analysis for women in the sample (n = 1,082) shows that women generally had slightly higher FTA likelihood than the full sample at most score levels, with the exception of scores 2 and 6 where women had slightly lower rates. The effect size (η^2 = .070) for women alone is slightly larger than that for the whole sample (η^2 = .068), indicating the FTA scale may have slightly stronger discriminatory power among women.

The pattern doesn't follow the expected progression as clearly for women as it does for the full sample, particularly at the higher end of the scale where women's FTA rates plateau rather than continuing to increase.

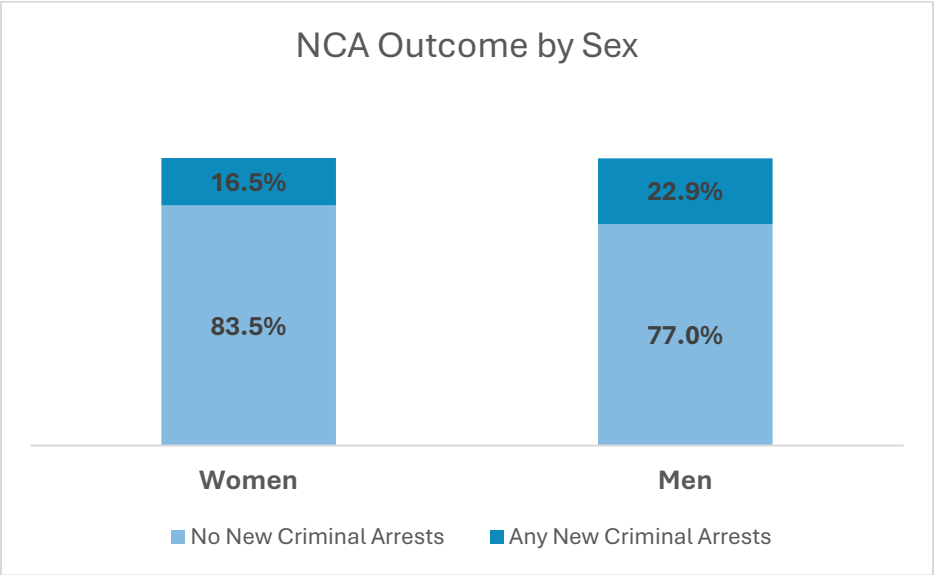
NCA Scores and Outcomes by Sex

Figure 12. NCA Scale Score by Sex



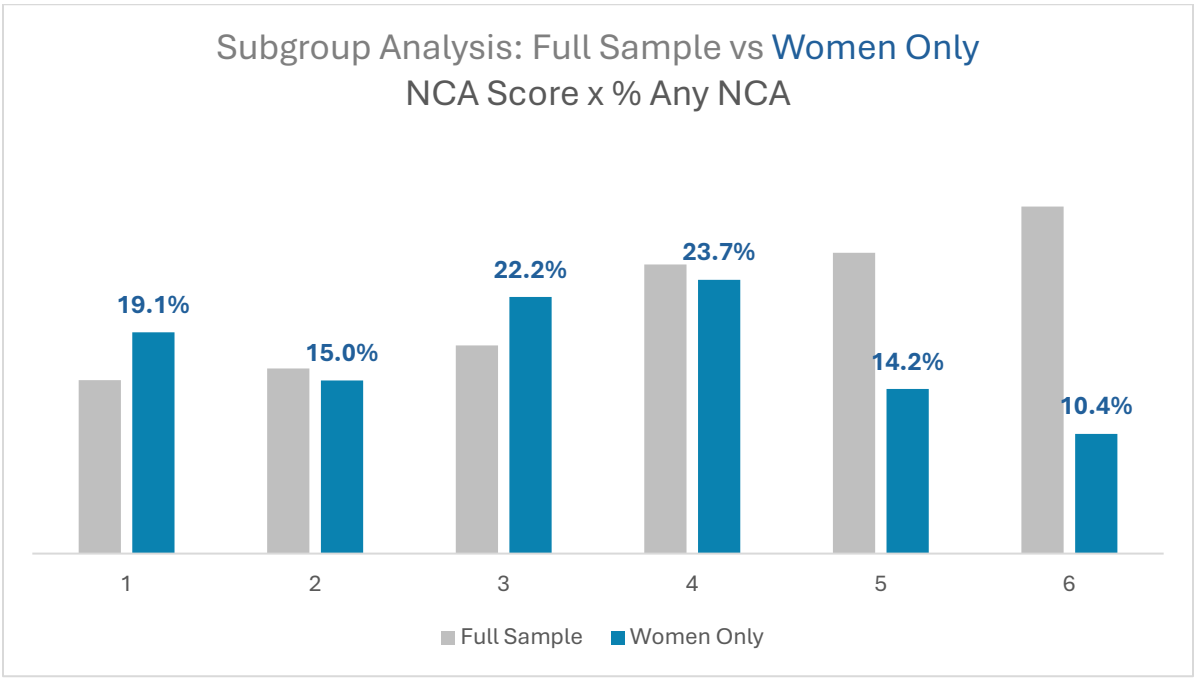
As shown above, there are significant (χ^2 = 23.43, $p < .01$) differences between sexes in NCA scale scores. Women were more likely to score on the lower end of the NCA scale (3 or below), and men were more likely to score on the higher end (4 or above). The most notable differences occur at the extreme ends of the scale, where women are more likely than men to be scored 1 (14.6% vs 10.8%, respectively) and men are more likely than women to be scored 6 (12.7% vs 10.4%, respectively).

Figure 13. NCA Outcome by Sex



The chi-square analysis demonstrates a highly significant association between gender and new criminal arrests ($\chi^2 = 21.202$, $p < .001$). Men were nearly 40% more likely to reoffend than women during the pretrial period. While approximately three-quarters of men (77.0%) avoided new arrests, an even higher proportion of women (83.5%) successfully remained arrest-free while awaiting trial.

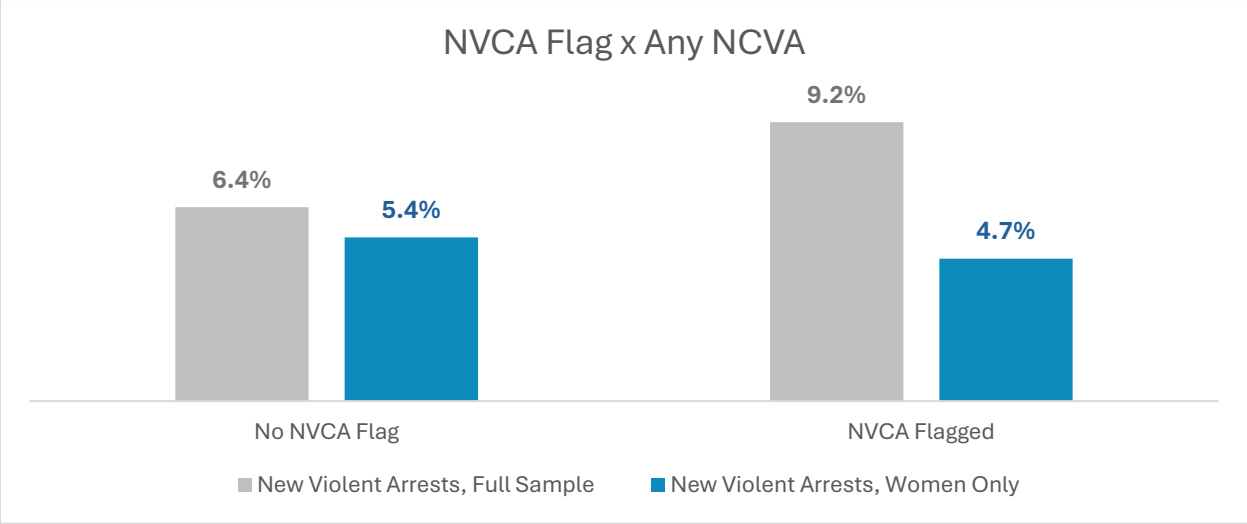
Figure 14. NCA Outcome by Sex



The subgroup analysis for the NCA scale for women shows that, compared to the full sample, women have higher likelihood of any NCA at every scale score except scores 5 and 6, where likelihood of any NCA is much lower. The Sidak post hoc analysis shows that only one significant difference exists: score 5 differs significantly from score 2 ($p = .003$). The effect size (η^2) for this subgroup analysis is .013, which is similar to the whole-sample analysis ($\eta^2 = .015$). The ROC/AUC analysis produced a score of .56, which is similar to the whole-sample analysis (.583) and suggests weak evidence of the scale's validity.

NVCA Flag and Outcomes by Sex

Figure 15. Subgroup Analysis: Full Sample vs Women Only



The subgroup analysis revealed that women who received the NVCA flag were not more likely to have a new violent crime arrest (4.7%) compared to women who were not flagged (5.4%). This finding differs from the full sample results and suggests the NVCA flag may not function as effectively for identifying women at higher risk for new violent arrests. The result is not statistically significant ($\chi^2 = 0.126$, $p = .722$) with a weak effect size (Cramer's $V = .011$).

The NVCA flag does not demonstrate significant predictive power within the women-only subgroup. The slightly lower rate of new violent crime arrests among flagged women when compared to women who were not flagged (4.7% vs 5.4%, respectively) suggests that factors contributing to violent recidivism may operate differently for women, or that the NVCA flag criteria may be less applicable to for this group.

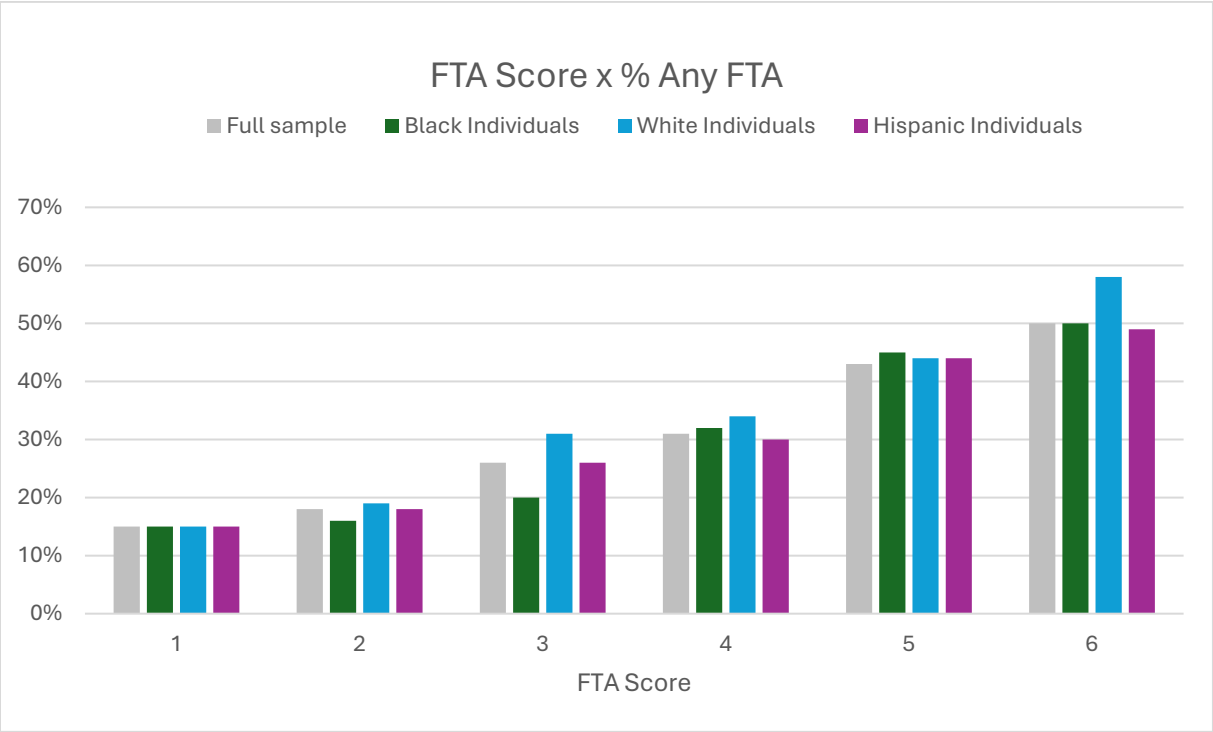
Race

The race measure in this data set includes several categories with very few cases. To avoid overgeneralizing from small samples, we have performed subgroup analyses for only the

three largest groups: Black[2] individuals (n = 936), white individuals (n = 1,118), and Hispanic individuals (n = 3,490).

FTA Scores and Outcomes by Race

Figure 16. Race/Ethnicity Subgroup Analysis



The "big picture" of the subgroup analysis reveals distinct racial patterns in FTA risk across score levels. At score 1, all racial groups show identical FTA rates (15%). However, clear differences emerge at higher score levels.

At moderate FTA scores (2-4), white individuals consistently demonstrate the highest rates of any FTA, with particularly notable differences at scores 2 and 3 (19% vs 16% for Black individuals at score 2; 31% vs 20% for Black individuals at score 3). Black individuals show the lowest FTA rates at scores 2-3, suggesting the scale may be less sensitive for this population at moderate risk levels. At higher FTA scores (5-6), the pattern shifts. At score 5,

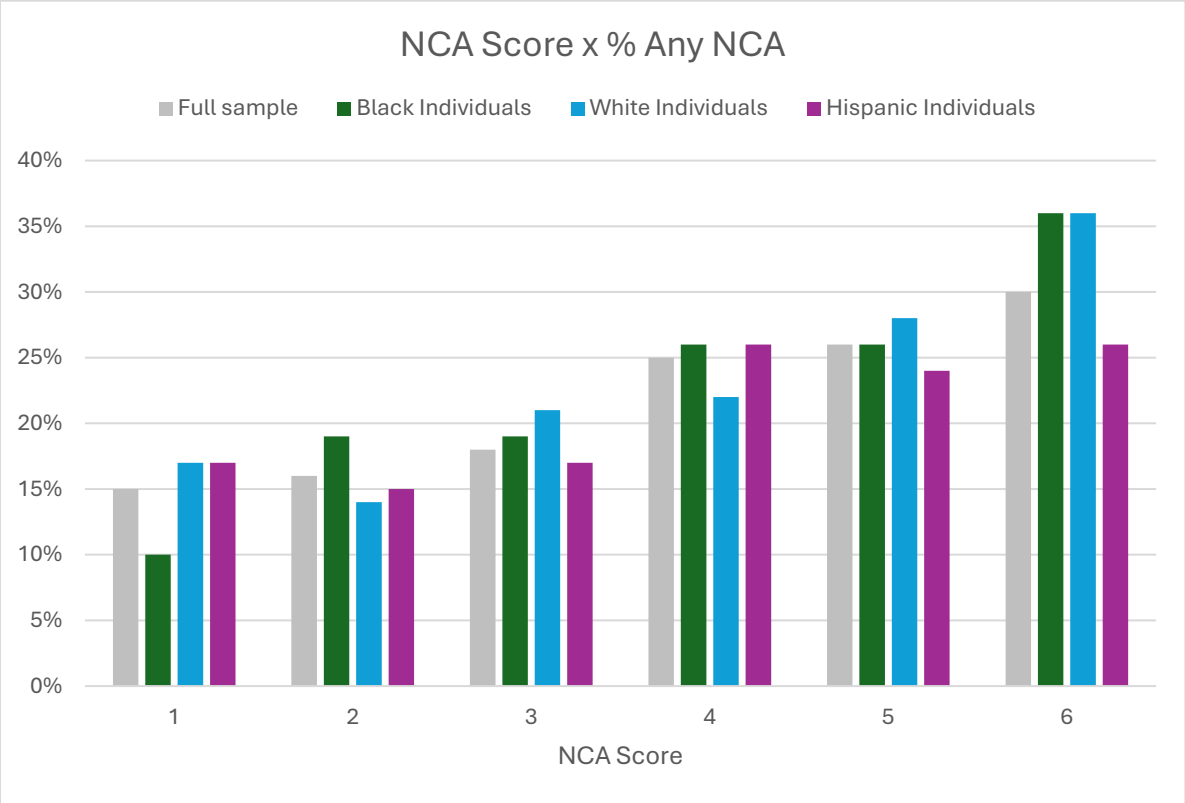
[2] In this document, we refer to "Black" as anyone belonging to the African diaspora. Additionally, throughout this document in line with Crenshaw (1988:1332), we capitalize "Black" as Black individuals constitute a specific cultural group and as such, require denotation as a proper noun. Those of the African diaspora have a set of shared cultures and experiences. We do not capitalize white, as white people are not a single cultural group. Crenshaw, Kimberlé (1988). "Race, Reform and Retrenchment: Transformation and Legitimation in Anti Discrimination Law." *Harvard Law Review*.

Black individuals show the highest FTA rate (45%), slightly exceeding white (44%) and Hispanic (44%) individuals. At the highest score level (6), white individuals demonstrate the highest FTA rate (58%), compared to Black (50%) and Hispanic (49%) individuals.

Hispanic individuals generally track closely with the full sample rates across most score levels, showing consistent but moderate increases from 15% to 49%. This suggests the FTA scale functions most similarly for Hispanic individuals compared to the overall validation sample, while showing more pronounced variation in effectiveness for Black and white subgroups at different risk levels.

NCA Scores and NCA Outcome by Race

Figure 17. Race/Ethnicity Subgroup Analysis



The NCA subgroup analysis reveals complex and shifting racial patterns that differ significantly from the FTA analysis. Hispanic individuals do NOT consistently demonstrate higher rates across all score levels; instead, the racial group with the highest NCA rates varies considerably by score level.

At lower NCA scores (1-2), the pattern is mixed: Hispanic and white individuals show identical rates at score 1 (17% each, compared to 10% for Black individuals), but Black individuals demonstrate the highest rate at score 2 (19%), with Hispanic (15%) and white

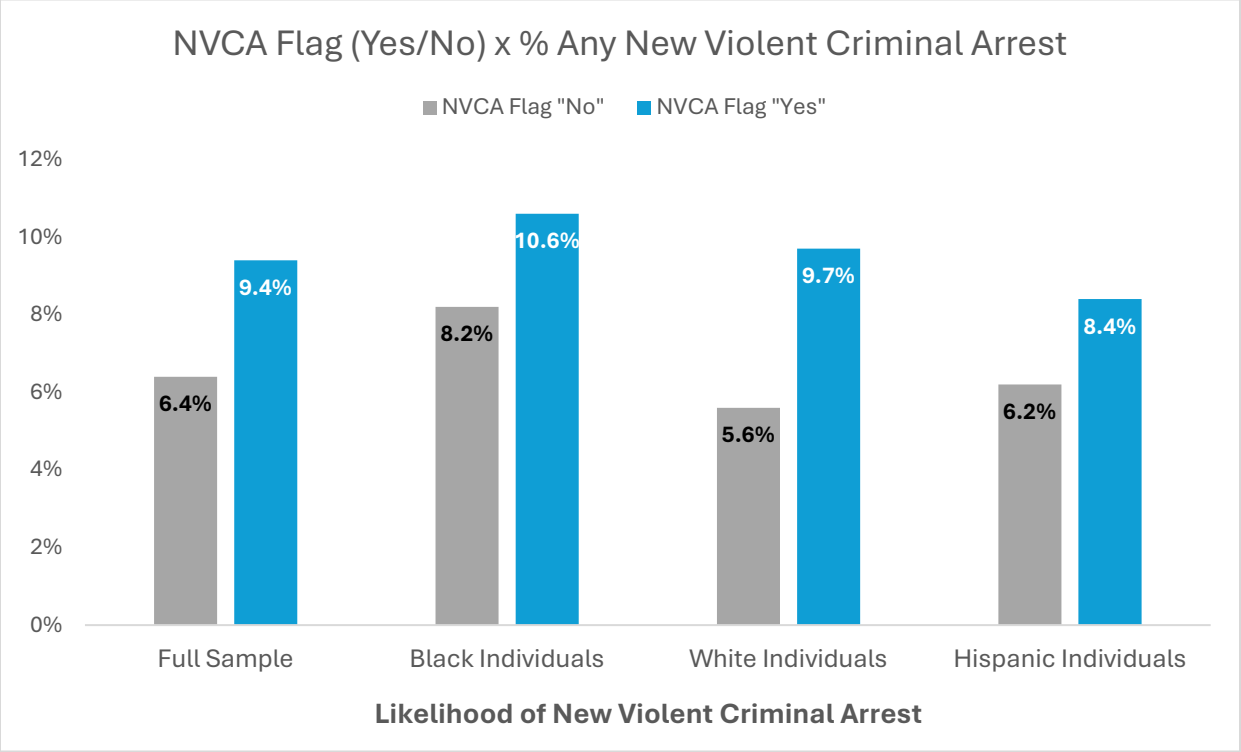
(14%) individuals showing lower, similar rates. At moderate scores (3-4), white individuals show the highest rates at score 3 (21%), while Black and Hispanic individuals are tied at score 4 (26% each), both exceeding white individuals (22%). This inconsistent pattern suggests variable predictive performance across racial groups depending on risk level.

At higher scores (5-6), Black and white individuals converge at the highest rates (36% each at score 6), while Hispanic individuals show the lowest rates (26% at score 6). This represents a complete reversal from lower score levels.

Unlike the FTA analysis where patterns were more consistent within score ranges, the NCA scale shows unstable predictive patterns across racial groups. Black and white individuals show remarkably similar performance only at the highest risk level (score 6), while demonstrating substantial differences at lower scores. The lack of consistent patterns suggests the NCA scale may have differential validity across racial groups, potentially indicating that any observable difference across racial groups reflects systematic measurement issues rather than true risk differences.

NVCA Flag and NVCA Outcome by Race

Figure 18. Race Subgroup Analysis



In the full sample, 6.4% of individuals flagged "No" for NVCA had a new violent criminal arrest during the observation period compared to 9.4% of those flagged. This difference is

statistically significant (χ^2 = 10.668, p = .001, Cramer's V = .043), indicating the NVCA flag functions as intended in the overall sample. Of Black individuals not flagged, 8.2% had a new violent criminal arrest, compared to 10.6% of those flagged. While this 2.4% difference is in the expected direction, it is not statistically significant (χ^2 = 1.284, p = .257, Cramer's V = .037), suggesting the NVCA flag has limited predictive validity for Black individuals.

There is a small but significant difference in NVCA outcomes for Hispanic individuals where 6.2% of those not flagged had a new violent criminal arrest versus 8.4% of those flagged. This 2.2% difference is statistically significant (χ^2 = 3.995, p = .046, Cramer's V = .034), indicating modest predictive validity for this subgroup. The biggest difference is observed for white individuals where those flagged are nearly twice as likely to have a new violent criminal arrest (9.7% compared to 5.6% of those not flagged. (χ^2 = 4.437, p = .035, Cramer's V = .063). Thus, the NVCA flag performs best for white individuals, shows modest effectiveness for Hispanic individuals, but demonstrates limited predictive validity for Black individuals, which may risk misclassification within this subgroup.

Key Findings

Overall, this analysis finds that the PSA performs moderately well for predicting pretrial outcomes for this sample, but there are several notable exceptions.

The FTA scale shows the strongest predictive validity, with moderate evidence of effectiveness (AUC = .66, η^2 = .068) and clear discrimination between risk levels. However, the scale exhibits clustering patterns where neighboring scores often fail to show significant differences—particularly between scores 1-2 and between scores 5-6. The NCA scale demonstrates weaker overall performance with only small effect sizes (η^2 = .015) and weak evidence of validity (AUC = .583). The scale effectively separates individuals into two broad risk clusters—lower-risk (scores 1-3) and higher-risk (scores 4-6)—but provides limited discrimination *within* these clusters. The FTA scale also reveals concerning racial patterns, with white individuals consistently showing the highest failure-to-appear rates at most score levels, reaching 58% at score 6 compared to 50% for Black individuals and 49% for Hispanic individuals. This differential performance across racial groups suggests the scale's risk factors may not capture risk equally across populations.

Several concerns emerge regarding the NVCA flag performance. The flag does not function as intended in the full sample, with flagged individuals showing lower rates of new violent arrests than those not flagged. This represents a fundamental breakdown in the tool's predictive validity for violent re-offense. It also fails to predict the risk for women. Among

women, those flagged had lower rates of new violent arrests when compared to those not flagged (4.7% vs 5.4%, respectively), and this difference is not statistically significant (p = .722). This represents a fundamental breakdown in the tool's core function for nearly one-fifth of the assessed population and raises serious questions about gender bias in risk assessment.

Equally concerning are the racial disparities in NVCA flag effectiveness. The flag performs optimally for white individuals, showing nearly double the risk for flagged individuals when compared to not flagged individuals (5.6% vs 9.7%, p = .035). However, for Black individuals—who comprise 16.1% of the sample—the flag shows no predictive validity whatsoever between flagged and not flagged individuals (8.2% vs 10.6%, p = .257). Given that Black individuals are more likely to receive NVCA flags than the sample average, this suggests systematic overclassification that could contribute to disproportionate detention rates without corresponding public safety benefits.

The FTA scale also reveals concerning racial patterns, with white individuals consistently showing the highest failure-to-appear rates at most score levels, reaching 58% at score 6 compared to 50% for Black individuals and 49% for Hispanic individuals. This differential performance across racial groups suggests the scale's risk factors may not capture risk equally across populations.

In conclusion, the current validation findings largely corroborate and strengthen the patterns observed in the second validation conducted in May 2025, while revealing some developments. Both validations demonstrate remarkably consistent performance for the FTA scale, with the current validation showing an AUC of .66 compared to .67 in the previous validation, confirming moderate predictive validity. Similarly, both studies identified persistent weaknesses in the NCA scale and fundamental problems with the NVCA flag's performance across demographic subgroups. However, the current validation's larger sample size (5,826 vs 3,408 cases) and extended timeframe provides stronger statistical power to confirm demographic disparities that were suggested in the earlier analysis. Most notably, while the second validation indicated problematic NVCA flag performance for Black individuals and women, the current validation reveals that this has deteriorated further—with flagged individuals now actually showing *lower* rates of new violent arrests than non-flagged individuals across multiple subgroups. The consistency of these findings across two independent validation periods, combined with the worsening performance of the NVCA flag, strengthens the urgency for the systematic review and recalibration of assessment criteria, particularly for demographic subgroups where the tool demonstrates persistent bias or inverse predictive relationships. Additionally, future

research should investigate the underlying factors contributing to these disparities to ensure the PSA promotes rather than undermines equitable pretrial justice outcomes.