THE UNIVERSITY OF
NEW MEXICO

A Revalidation Study of Bernalillo County's Public Safety Assessment

July 2024

Prepared by:
Alex Severson, Ph.D.
Elise Ferguson, M.A.

Prepared for:
New Mexico Administrative Office
of the Courts and Bernalillo County

UNIVERSITY OF NEW MEXICO, INSTITUTE FOR SOCIAL RESEARCH
CENTER FOR APPLIED RESEARCH AND ANALYSIS

Table of Contents

Introduction

The Public Safety Assessment (PSA) is an evidence-based judicial decision-making tool developed by Arnold Ventures designed to assist judges in making release decisions during the pretrial phase (AdvancingPretrial.org, 2020). Arnold Ventures originally constructed the PSA using data from approximately 750,000 cases from 300 jurisdictions to develop a scored set of risk factors predictive of an individual's likelihood of failing to appear in court (FTA) and engaging in new criminal activity (NCA) during the pretrial period. The PSA also flags individuals who present an elevated risk of committing a violent crime during the pretrial period, generating a New Violent Criminal Activity (NVCA) flag. The PSA has been validated on over half a million cases nationally and has been revalidated in locations as diverse as Kentucky (DeMichele et al., 2018), California (Skog & Lacoe, 2021), and Texas (Greiner et al., 2020).

The PSA was implemented in Bernalillo County, New Mexico in June 2017 exclusively for felony cases, making it unique relative to other jurisdictions that typically use the PSA to evaluate risk for both misdemeanants and felons. In 2021, the University of New Mexico Center for Applied Analysis (CARA) completed a validation study of the PSA in Bernalillo County using a sample of 10,289 cases spanning June 2017 to March 2020 (Ferguson et al., 2021). This study found that scores on the PSA in Bernalillo County were predictive of pretrial failure rates and that the PSA possessed fair to good levels of predictive validity.

While we completed the validation study of the PSA in Bernalillo County in 2021, the National Association of Pretrial Services Agencies (NAPSA) recommends revalidating pretrial risk assessments every two to three years (NAPSA, 2020). To this end, the goal of the present study was to revalidate the PSA in Bernalillo County using a sample of 22,387 cases spanning July 1, 2017, through June 30, 2023.

In what follows, we review the factors included within the PSA and describe how scores on the PSA relate to release recommendations. We then review key findings of our 2021 validation study. We then discuss sample construction, our primary empirical results, limitations to the present study, summarize recommendations and our overall findings. For a deeper overview of the background research on the PSA – including the historical background of its implementation in Bernalillo County and an overview of the existing validation research on the PSA, we refer interested readers to our 2021 validation report linked here.

A Brief Primer on the PSA

Pretrial services officers score the PSA by reviewing an individual's criminal history, current cases, and age to create an FTA score (Range: 1-6), an NCA score (Range: 1-6), and a binary flag for NVCA (Range: 0-1). In Table 1, we present how the nine PSA factors map onto the three pretrial outcomes (AdvancingPretrial.org, 2024). An "X" indicates an increase in the individual's likelihood of that outcome, based on the risk factor. For instance, if the individual's current offense is violent (#2 below), it increases that individual's likelihood of committing an NVCA during their pretrial period.

The PSA consists of two scales designed to estimate an individual's likelihood of FTA and NCA. The combination of the NCA and FTA scale scores generate an overall score on the Decision-Making Framework (DMF), which is a matrix of NCA by FTA scores that sets the PSA's conditions of release (COR) recommendations. Each jurisdiction that utilizes the PSA develops a DMF[1]. Figure 1 shows the

[1] The Second Judicial District Court and Bernalillo County Metropolitan Court updated the DMF used in Bernalillo County in April 2023 in consultation with justice stakeholders, including LOPD, DA, and AOC. Relative to the DMF presented in the 2021 validation report, in April 2023, the AOC modified colors in the release matrix to comply with best practices and revised the language contained in the matrix regarding supervision levels to remove language suggesting a defendant should be detained. The tool was not intended to provide a recommendation for detention or release; rather, once a release decision had been made, the

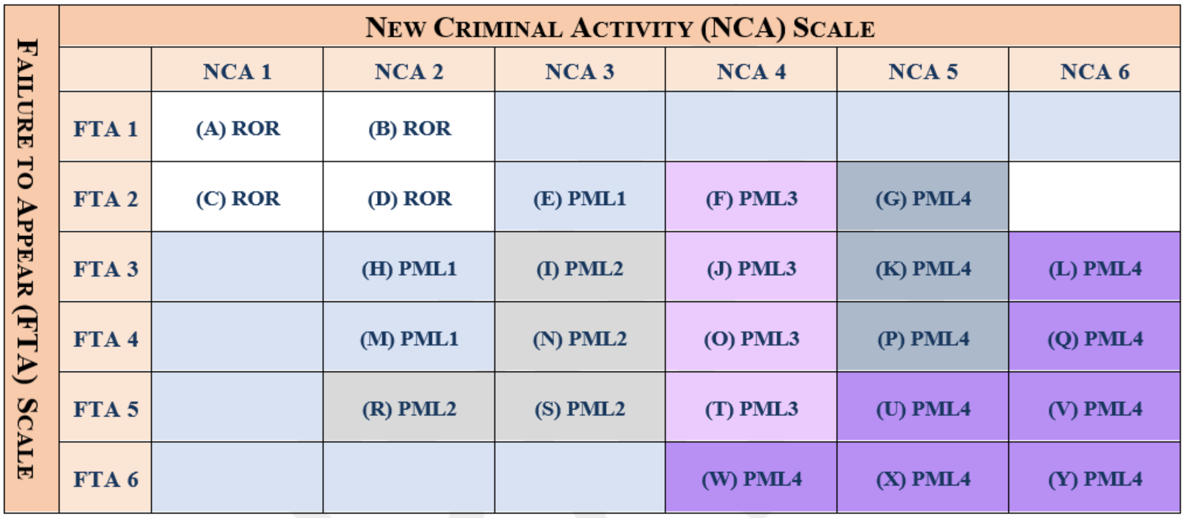
DMF used in Bernalillo County. The release recommendations generated by the DMF vary in terms of the scope of supervision and pretrial services required of the individual from ROR (Release on Own Recognizance), PML 1, PML 2, PML 3, to PML 4 and vary by jurisdiction[2].

Table 1.
PSA Risk Factors and Pretrial Outcomes

Risk Factor	Pretrial Outcome		
	FTA	NCA	NVCA
1. Age at current arrest		X	
2. Current violent offense			X
2A. Current violent offense and 20 years old or younger			X
3. Pending charge at the time of the offense	X	X	X
4. Prior misdemeanor conviction		X	
5. Prior felony conviction		X	
5A. Prior conviction (misdemeanor or felony)	X		X
6. Prior violent conviction		X	X
7. Prior failure to appear in the past two years	X	X	
8. Prior failure to appear older than two years	X		
9. Prior sentence to incarceration		X	

In Figure 1, we present the revised DMF that was developed specifically for Bernalillo County and has been in use since April 2023.

Figure 1.
PSA Decision-Making Framework in Bernalillo County[3]



NEW CRIMINAL ACTIVITY (NCA) SCALE

FAILURE TO APPEAR (FTA) SCALE

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	(A) ROR	(B) ROR				
FTA 2	(C) ROR	(D) ROR	(E) PML1	(F) PML3	(G) PML4	
FTA 3		(H) PML1	(I) PML2	(J) PML3	(K) PML4	(L) PML4
FTA 4		(M) PML1	(N) PML2	(O) PML3	(P) PML4	(Q) PML4
FTA 5		(R) PML2	(S) PML2	(T) PML3	(U) PML4	(V) PML4
FTA 6				(W) PML4	(X) PML4	(Y) PML4

tool is used to make a recommendation for conditions or release and pretrial supervision level. Importantly, the supervision levels themselves were not changed in 2023. The specific change involved changing the "Detain or Release with Maximum Conditions" to PML 4. This change did not change condition of release recommendations or affect supervision levels. The original DMF is included in Appendix D.

[2] The level of pretrial supervision, or pretrial monitoring level (PML), ranges from Level 1 to Level 4, with increasing degrees of supervision and conditions as PMLs increases.

.

Revalidation Sample Construction

We constructed the revalidation sample using electronic court data from the Bernalillo County Metropolitan Court (BCMC) and the Second Judicial District Court (SJDC). The electronic data included all cases filed in the study time frame. From this data, we selected cases for the sample if:

- The case was opened between July 1, 2017 and June 30, 2023;
- A PSA was administered;
- The defendant was in custody for the Felony First Appearance (FFA) or the Felony Arraignment (FA);
- The case was disposed as of June 30, 2023; and
- The defendant was released into the community between the FFA/FA and the final case disposition.

Between July 1, 2017, and June 30, 2023, 59,642 cases were opened, 30,635 in BCMC and 20,231 in SJDC. For a variety of reasons, we excluded 62% (n = 37,210) from analysis. We describe our exclusionary criteria below and in Table 2.

Table 2.
Exclusions from Electronic Court Data and Outcome Sample

Exclusion Reason	**Total**
Fugitive	1,532
In the BCMC-SJDC File	8,776
No PSA/Not in Custody for PSA	7,554
Not in Custody for FFA/FTA at or Before Felony Arraignment	2,243
Pending Case	3,327
No Exposure Time	9,901
Other (e.g., No Felony Arraignment; No COR)	3,877
Total	37,210

First, we excluded fugitive cases, or cases where the individual was facing charges from related to a fugitive charge from another state or jurisdiction (n =1,532), as these were not eligible for assessment. We categorized cases in the sample by case filings and indictments. For many SJDC cases, there was overlap between the BCMC case and a SJDC case filed during the pretrial period. In Bernalillo County, most felony cases begin in BCMC with few exceptions. Once a case was filed, the prosecution had 60 days from the individual's FFA date and pretrial release date to charge the individual through either a grand jury indictment or through preliminary hearing where a judge may decide there existed evidence sufficient to indict. Once indicted, the BCMC case was linked to the SJDC case, and the case continued in SJDC. For the purposes of the study, we considered the overlapping BCMC and SJDC case as one case. If the indictment occurred after the BCMC pretrial period, the SJDC case was considered separately, with its own pretrial period. We reported the findings from this study in the aggregate rather than separately by court. We excluded 8,776 cases because the SJDC portion of the case was attached to the BCMC portion.

We excluded an additional 7,754 cases because they had no PSA. This occurred most often for SJDC cases. Of the remaining cases, we excluded 2,243 because the defendant was not in custody for the felony first appearance or had an FTA warrant prior to the felony arraignment.

For each case in the sample, we collected and identified the case status as either closed or pending. A case was considered closed if there was a closing event or final disposition, such as a sentence, dismissal, plea bargain, or finding of no probable cause, on or before March 31, 2023. If the case had a closing event, we collected the date of the event as the case close date. If the individual was indicted prior to the disposition of the BCMC case, we considered it a BCMC-SJDC case. If the indictment occurred after the disposition of the BCMC case or did not occur at all, we considered the case BCMC only. We identified a case as pending if there was not a closing event or disposition by March 31, 2023. We included only cases that were both opened and closed between July 1, 2017, and March 31, 2023. Of the 59,642 cases, 5.6% (n = 3,327) were pending. Finally, we excluded 16.6% (n = 9,901) cases where cases were closed but where there was no exposure time in the community[4]. The remaining 22,387 bookings met the criteria of having a PSA and the individual was in custody for the release decision and constitute out outcome sample. There were 15,831 unique individuals within the 22,387 bookings.

Within the outcome sample, 0.6% (n = 136) of individuals had exposure time of less than one day. One percent (n = 229) had exposure time of less than two days. Three percent (n = 799) had exposure time of less than one week. Overall, 80.8% (n = 18,087) of the outcome sample had at least one month of exposure time. The median exposure time was 50.2 days.

Sample Description

In Table 3, we present the distribution of gender and race in the outcome sample. A majority (74.4%; n = 16,648) of individuals in the outcome sample were male, and a plurality (45.4%; n = 10,170) were Hispanic.

Table 3.
Distribution of Gender and Race in Outcome Sample

Category	Count (N = 22,387)	Percent
Gender		
Male	16,648	74.4%
Female	5,599	25.0%
Not Known	140	0.6%
Race		
Hispanic	10,170	45.4%
White	7,169	32.0%
Native American	1,782	8.0%
Black	1,690	7.5%
Other	1,576	7.0%

In Table 4, we present the distribution of individuals' joint FTA and NCA scale scores within the DMF framework for the outcome sample[5]. As mentioned earlier, the DMF was revised in April 2023 and the

[4] Exposure indicates an individual's time spent in the community during the pretrial period of the assessed case. To calculate exposure, we merged jail booking and release data from the Metropolitan Detention Center (MDC) with the individual's court case, selecting the corresponding booking for the assessment and hearing where the release decision was made. If an individual was not released during the pretrial phase or was released to the New Mexico Corrections Department (NMCD), we identified them as having no exposure. Inmates with exposure, in theory however short, had the chance of committing a new crime or having an FTA.

[5] The Bernalillo County DMF was changed by policy in March 2023 to better reflect the intended use case of the tool. Per the AOC, the DMF was not originally intended to provide a recommendation for detention or release; rather, once a release decision had been made, the tool was used to make a recommendation for conditions or release and pretrial supervision level (i.e., Detain/Max Detention was replaced with PML 4). This explains the difference in the matrices we present in Table 2 and 4.

cells labeled Detain/Max Conditions were relabeled as PML 4. Moreover, this change did not change how individuals were supervised.

Table 4.
PSA Condition of Release (COR) Recommendation Categories Based on FTA and NCA Scaled Scores

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	ROR 11.9% (2,661)	ROR 6.8% (1,515)				
FTA 2	ROR 2.4% (540)	ROR 6.0% (1,334)	ROR – PML 1 5.2% (1,155)	ROR – PML 3 2.8% (619)	ROR – PML 4 0.0% (3)	
FTA 3		ROR – PML 1 6.1% (1,369)	ROR – PML 2 9.8% (2,185)	ROR – PML 3 8.3% (1,862)	ROR – PML 4 0.0% (146)	Detain/Max Conditions 0.0% (16)
FTA 4		ROR – PML 1 1.8% (402)	ROR – PML 2 4.1% (929)	ROR – PML 3 4.8% (1,079)	ROR – PML 4 3.9% (862)	Detain/Max Conditions 0.8% (153)
FTA 5		ROR – PML 2 0.2% (54)	ROR – PML 2 1.6% (358)	ROR – PML 3 6.0% (1,335)	Detain/Max Conditions 5.1% (1,150)	Detain/Max Conditions 2.7% (598)
FTA 6				Detain/Max Conditions 1.8% (411)	Detain/Max Conditions 1.8% (402)	Detain/Max Conditions 5.6% (1,249)

In Table 5, we present the collapsed PSA COR recommendation categories for the outcome sample. The largest proportion of individuals (27.0%; n = 6,050) were recommended no pretrial supervision if released. Twenty-two percent of individuals were recommended for ROR – PML 3 (n = 4,895), and 17.9% were recommended for Detain/Max Conditions (as of April 2023 known as PML 4) respectively (n = 3,979).

Table 5.
PSA Collapsed Conditions of Release Recommendation Categories

Release Category	**Count**	**Percent**
ROR	6,050	27.0%
ROR – PML 1	2,926	13.1%
ROR – PML 2	3,526	15.8%
ROR – PML 3	4,895	21.9%
ROR – PML 4	1,011	4.5%
Detain/Max Conditions	3,979	17.8%
Total	22,387	100%

PSA Risk Factors and Scores

In this section, we describe the distribution of FTA, NCA, and NVCA risk factors within the outcome sample and discuss how scores on these risk factors correlated with observed FTA, NCA, and NVCA rates. We also provide information on the charge level for NCA that occurred during the pretrial period for the subset of cases where NCA occurred.

Failure to Appear and Risk Factors

In Table 6, we present the distribution of risk factors that determined an individual's FTA score on the PSA. Most individuals did not have a pending charge at time of arrest (67.3%; n = 15,075), and most had a prior misdemeanor or felony conviction (69.3%; n = 15,519). While most individuals did not have an

FTA within the two years preceding their source case date (63.9%; n = 14,303), a majority (54.5%; n = 12,203) had at least one prior FTA two or more years preceding the source case.

Table 6.
FTA Risk Factors

Factor	Response	Count	Percent
Pending Charge at Time of Arrest	No	15,075	67.3%
	Yes	7,312	32.7%
Any Prior Conviction (MD or Felony)	No	6,868	30.7%
	Yes	15,519	69.3%
Prior FTA in Past Two Years	0	14,303	63.9%
	1	3,545	15.8%
	2+	4,539	20.3%
Prior FTA > Two Years Old	No	10,184	45.5%
	Yes	12,203	54.5%

In Table 7, we present Pearson's r correlation coefficients[6] alongside standard errors for the four factors used to calculate the FTA score with observed FTA outcomes. It is important to evaluate the relationship between risk factors and failure outcomes to understand whether and to what degree the different factors relate to the likelihood of pretrial failure. From Table 8, the two factors most correlated with observed FTA are (1) the scaled FTA score (r = 0.251) followed by (2) an individual having received an FTA within the past two years (r = 0.223). Given various guidelines defining correlation strength as 0.50 – 1.0 = "Strong Correlation", 0.3 – 0.5 = "Moderate Correlation", 0.1 – 0.3 = "Weak Correlation", and < 0.10 as "Negligible Correlations", all correlations observed could be described as "Weak" (Dancey and Reidy, 2007; Akoglu 2017).

Table 7.
Correlation Coefficients for FTA Risk Factors and Outcomes[7]

Factor	Correlation Coefficient	Standard Error
Pending Charge at Time of Arrest	0.145*	+/- 0.007
Any Prior Conviction (MD or Felony)	0.132*	+/- 0.007
Prior FTA in Past Two Years	0.223*	+/- 0.006
Prior FTA > Two Years Old	0.151*	+/- 0.007
FTA Score	0.251*	+/- 0.006

In Table 8, we present the distribution of scores on the FTA scale on the PSA for the outcome sample. The largest FTA Score category was FTA - 3, comprising 24.9% (n = 5,578) of the sample.

[6] Pearson's r coefficients rely on a few assumptions (e.g., normally distributed variables; linearity) that are not satisfied for all variables within the FTA predictor set. For example, the Pearson correlations assumes one variable is dichotomous and the other variable is continuous. However, some of the FTA factors are dichotomous (e.g., Pending Charge; Any Prior Conviction; FTA > Two Years Old) and the outcome (i.e., whether FTA occurred) is dichotomous. For these reasons, we report results of Pearson's r or phi correlation coefficients, the latter of which are used for comparison of two binary variables.

[7] Asterisks (*) indicate p-value < 0.001.

Table 8.
FTA Score for Outcome Measures Sample

FTA Score	Count	Percent
FTA 1	4,176	18.7%
FTA 2	3,651	16.3%
FTA 3	5,578	24.9%
FTA 4	3,425	15.3%
FTA 5	3,495	15.6%
FTA 6	2,062	9.2%
Total	22,387	100%

New Criminal Activity and Risk Factors

In Table 9, we present the frequency table of the factors used to calculate the NCA score. As the NCA scale is comprised of some of the same factors used for the FTA scale (e.g., Pending Charge; Prior FTA in Past Two Years), we only comment on different factors here. A majority (87.7%; n = 19,635) of the outcome sample was 23 years of age or older at the time of the source case. Most had a prior misdemeanor conviction (64.5%; n = 14,433), whereas most did not have a prior felony conviction (61.4%; n = 13,740). A majority of the sample did not have a prior violent conviction (69.7%; n = 15,607), and most had not served a prior sentence to incarceration (55.8%; n = 12,493).

Table 9.
NCA Risk Factors

Factor	Response	Count	Percent
Age at Current Arrest	22 or Younger	2,752	12.3%
	23 or Older	19,635	87.7%
Pending Charge at Time of Arrest	No	15,072	67.3%
	Yes	7,312	32.7%
Prior Misdemeanor Conviction	No	7,954	35.5%
	Yes	14,433	64.5%
Prior Felony Conviction	No	13,740	61.4%
	Yes	8,647	38.6%
Prior Violent Conviction Count	0	15,607	69.7%
	1-2	5,380	24.0%
	3+	1,400	6.2%
Prior FTA in Past Two Years	0	14,303	63.9%
	1	3,545	15.8%
	2+	4,539	20.3%
Prior Sentence to Incarceration	No	12,493	55.8%
	Yes	9,894	44.2%

In Table 10, we present Pearson’s r correlation coefficients for the seven factors used to calculate the NCA and the overall NCA score. From Table 10, we observe that most factors on the NCA scale were either weakly positively correlated with NCA (e.g., Pending Charge; Prior FTA in Past Two Years; Prior Sentence to Incarceration; NCA Score) or were negligibly positively correlated with NCA (e.g., Prior Misdemeanor Conviction; Prior Felony Conviction). One factor – Age at Current Arrest – was negatively and negligibly correlated with NCA and not significantly different from a correlation of 0.0, a finding which, while consistent with the weakness of the correlation we observed in our 2021 validation report, is noteworthy, as generally risk assessment tools should only include items found to be predictive of the outcome of interest (Duwe & Kim, 2016; Georgiou, 2019). We will return to this point in the

Recommendations and *Conclusion* section of the report as the lack of predictive of power of these two factors embedded within the PSA across two validation studies suggests a potential need to rescore (i.e., "renorm") this factor for use on the local population and to conduct further exploratory analyses of the relationship between age and failure rates.

Table 10.

Correlation Coefficients for NCA Risk Factors and Outcomes[8]

Factor	Correlation Coefficient	Standard Error
Age at Current Arrest	-0.016	+/-0.007
Pending Charge at Time of Arrest	0.128*	+/-0.007
Prior Misdemeanor Conviction	0.097*	+/-0.007
Prior Felony Conviction	0.065*	+/-0.007
Prior FTA in Past Two Years	0.162*	+/-0.007
Prior Sentence to Incarceration	0.115*	+/-0.007
NCA Score	0.187*	+/-0.007

In Table 11, we present the distribution of scores on the NCA scale on the PSA for the outcome sample. The largest NCA Score category was NCA - 4, comprising 23.7% (n = 5,306) of all individuals within the outcome sample.

Table 11.

NCA Score for Outcome Measures Sample

NCA Score	Count	Percent
NCA 1	3,201	14.3%
NCA 2	4,674	20.9%
NCA 3	4,627	20.7%
NCA 4	5,306	23.7%
NCA 5	2,563	11.4%
NCA 6	2,016	9.0%
Total	22,387	100%

New Violent Criminal Activity and Risk Factors

In Table 12, we present the frequency table of risk factors used to calculate the NVCA score which, in turn, was converted into the dichotomous NVCA flag. As the NVCA scale includes factors already discussed for both the FTA and NVCA scale (e.g., Pending Charge; Any Prior Conviction; Prior Violent Convictions), we only comment on unique factors here. Sixty percent (n = 13,487) of individuals in the outcome sample did not have a source case charge that was violent. Ninety-seven percent (n = 21,718) of individuals in the outcome sample did not have a source case charge that was violent and were over the age of 20 years old.

[8] Asterisks (*) indicate p-value < 0.001.

Table 12.
NVCA Risk Factors

Factor	Response	Count	Percent
Current Violent Offense	No	13,487	60.2%
	Yes	8,900	39.8%
Current Violent Offense and < 20 Years Old	No	21,718	97.0%
	Yes	669	3.0%
Pending Charge	No	15,075	67.3%
	Yes	7,312	32.7%
Any Prior Conviction (MD or Felony)	No	6,869	30.7%
	Yes	15,518	69.3%
Prior Violent Convictions	0	15,607	69.7%
	1-2	5,380	24.0%
	3+	1,400	6.3%

In Table 13, we present Pearson's r correlation coefficients for the five factors used to calculate the NVCA score and the overall NVCA flag with observed NVCA[9]. From Table 13, we observe that most factors on the NVCA scale were negligibly positively correlated with NVCA with the factor most correlative with NVCA being the binary NVCA Flag (chi-correlation = 0.082). Like the findings we presented in the NCA subsection, the factor which included age (i.e., being convicted of a violent offense *and* being less than 20 years old) was not significantly correlated with observed NVCA rates.

Table 13.
Correlation Coefficients for NVCA Risk Factors and Outcomes[10]

Factor	Pearson's r	Standard Error
Current Violent Offense	0.058*	+/-0.007
Current Violent Offense and <= 20 Years Old	0.012	+/-0.007
Pending Charge	0.049*	+/-0.007
Any Prior Conviction (MD or Felony)	0.032*	+/-0.007
Prior Violent Conviction	0.058*	+/-0.007
NVCA Flag	0.082*	+/-0.007

Table 14 displays the distribution of NVCA flags for the outcome sample. Seventeen percent (n = 3,785) of the outcome sample received the NVCA flag conditional on how they scored on the NVCA risk factors.

[9] The NVCA scores are collapsed into the binary NVCA flag with scores of 1–4 equal to no flag and scores of 5–6 equal to a violent flag.
[10] Asterisks (*) indicate p-value < 0.001.

Table 14.
NVCA Flag for Outcome Measures Sample

NCA Score	Count	Percent
No NVCA Flag	18,602	83.1%
NVCA Flag	3,785	16.9%
Total	22,387	100%

PSA Outcome Measures

In this section, we discuss the distribution of observed FTA, NCA, and NVCA rates for each release category within the DMF, show how observed FTA, NCA, and NVCA rates correlated with individuals' scores on the FTA, NCA, and NVCA scales conditional on judge adherence to the COR recommendations, and present descriptive statistics describing the charges of the subset of individuals within the outcome sample who engaged in NCA (e.g., the severity of the NCA charges in relation to the source charge; the qualitative type of charge received).

We present overall rates of FTA, NCA, and NVCA for the outcome sample in Table 15 below[11]. Table 6 shows FTA, NCA, and NVCA rates that parallel those found in the original validation study. The outcome sample FTA rate was 25.1% (n = 5,629), the NCA rate was 19.2% (n = 4,291), and the NVCA rate was 4.8% (n = 1,072), whereas the rates we reported in our 2021 study were an FTA rate of 22.9%, an NCA rate of 19.0%, and an NVCA rate of 4.7%.

When we split the data to explore whether there were significant differences in the FTA, NCA, and NVCA rates by period (before, during, and after Covid-19[12]), per a chi-square test, we found evidence of significantly different FTA rates between individuals with case opening dates which occurred during the pre-Covid, Covid, and post-Covid time periods, $\chi^2(2, N = 380) = 315.6, p < 0.00$. Importantly, in March 2022, the New Mexico Supreme Court issued order No. 22-8500-017, which implemented mandatory status hearings for defendants not in custody. In the period between 2017 and 2021, there were only 21 status hearings. In the period between 2022 and 2023, there were 2,492 status hearings. The 2,492 status hearings which occurred following the New Mexico Supreme Court order comprised 30.8% of all FTA cases for the sample period. The addition of status hearings is likely the primary driver of the increase in failure rates observed during and after Covid as increased hearings increased opportunities for failure. Post-hoc comparisons indicated failure to appear rates differed significantly across the periods: before Covid-19 [FTA Rate: 19.7%], during Covid-19 [FTA Rate: 29.7%], and after Covid-19 [FTA Rate: 33.5%][13][14]. We did not observe significant difference in NCA or NVCA rates by period.

[11] It is important to note that overall FTA rates may be biased downwards because not everyone in the sample had the possibility of failure (e.g., not everyone in the outcome sample had a subsequent court appearance and opportunity to fail and thus, their FTA rate would always be 0.0%). Including these individuals within the sample likely underestimates to the "true" FTA rate, though the broad rate we present here is technically an accurate estimate of overall FTA prevalence.

[12] We used the following dates to define the pre-Covid, during-Covid, and post-Covid time periods. We defined the pre-Covid period as any case with an opening date occurring before March 15, 2020, which represented the start of U.S. state lockdowns per the CDC. We defined a case as during-Covid period as any case with an opening date occurring between March 15, 2020 and May 5, 2023 when the WHO declared Covid was no longer a global public health emergency on May 5, 2023. Relatedly, May 11, 2023 was the date the U.S. federal Public Health Emergency for Covid-19 Warning expired. Accordingly, we defined a case as during the post-Covid period as any case with an opening date occurring after May 5, 2023.

[13] We also present descriptive statistics on failure rates by court, fiscal year, and time to failure in Appendix A.

[14] In March 2022, the New Mexico Supreme Court issued order No. 22-8500-017, which implemented mandatory status hearings for defendants not in custody. Preliminary review indicates that FTA for these hearings is likely the primary cause of the increase during and after Covid.

Table 15.
Overall FTA, NCA, and NVCA Rates

Outcome	Count	Frequency	2021 Validation Frequencies
FTA	5,629	25.1%	22.9%
NCA	4,291	19.2%	19.0%
NVCA	1,072	4.8%	4.7%

Failure to Appear Rates

The overall FTA rate of the outcome sample was 25.1%. In Table 16, we present the FTA rates for each DMF release category. As scores on the NCA and FTA scales increased, FTA rates increased.

Table 16.
FTA Rate by PSA COR Recommendation Category[15]

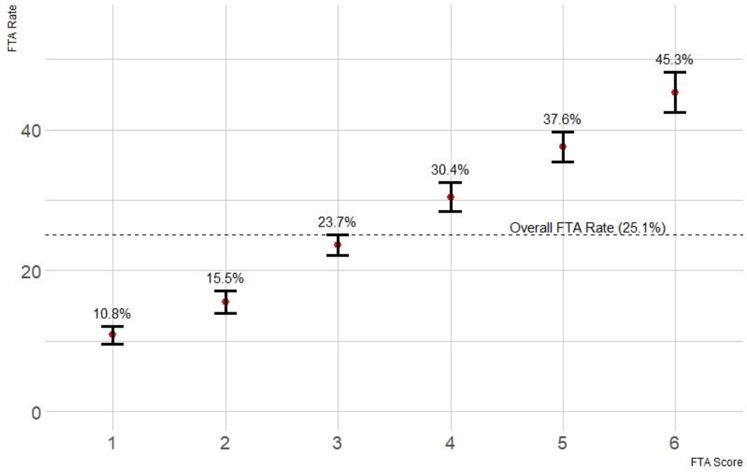
	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	ROR 10.9% (291)	ROR 10.7% (162)				
FTA 2	ROR 17.8% (96)	ROR 10.9% (145)	ROR – PML 1 17.9% (207)	ROR – PML 3 18.9% (117)	ROR – PML 4 66.7% (2)	
FTA 3		ROR – PML 1 16.8% (230)	ROR – PML 2 25.9% (566)	ROR – PML 3 25.9% (483)	ROR – PML 4 24.7% (36)	Detain/Max Conditions 31.3% (5)
FTA 4		ROR – PML 1 28.9% (116)	ROR – PML 2 30.4% (282)	ROR – PML 3 31.4% (339)	ROR – PML 4 29.9% (258)	Detain/Max Conditions 30.7% (47)
FTA 5		ROR – PML 2 31.5% (17)	ROR – PML 2 39.7% (142)	ROR – PML 3 36.8% (491)	Detain/Max Conditions 37.0% (426)	Detain/Max Conditions 39.6% (237)
FTA 6				Detain/Max Conditions 43.8% (180)	Detain/Max Conditions 46.3% (186)	Detain/Max Conditions 45.5 % (568)

In Figure 2, we show variance in the observed FTA rate by scores on the FTA scale with associated 99% confidence intervals[16]. Results suggest a positive relationship between scores on the FTA scale and observed FTA rates as well as statistically significant increases in FTA rates across all increasing score categories. For example, individuals who scored a one on the FTA scale failed to appear approximately 11% of the time, whereas individuals who scored a six on the FTA scale failed to appear approximately 45% of the time. Thus, the FTA scale correlates with observed FTA, suggestive of the scale's predictive validity.

[15] Percentages are of the total outcome sample, not of the specific cell. Parentheses indicate the sample size of individuals within the cell who failed to appear.

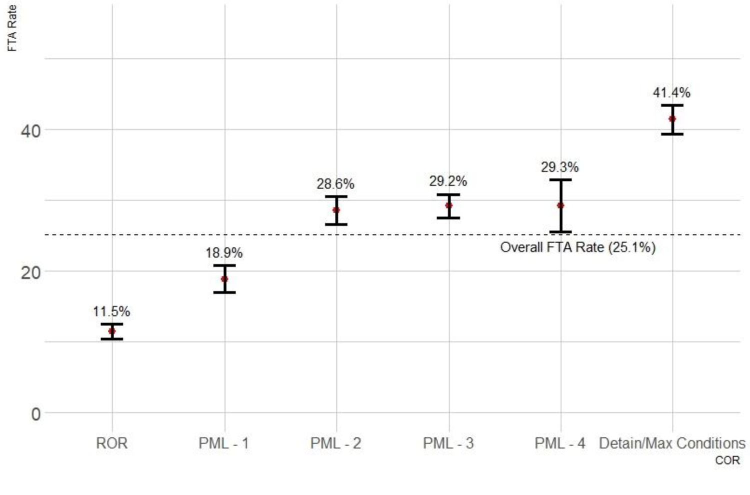
[16] We used 99% confidence intervals instead of the default 95% confidence intervals to generate more conservative estimates considering the large sample size for this study which can artificially bias estimates of statistical-significance upwards. For example, see work by Lin et al., (2013), DeMichele et al., (2020), and Monahan et al. (2017).

Figure 2.
FTA Rate by FTA Score



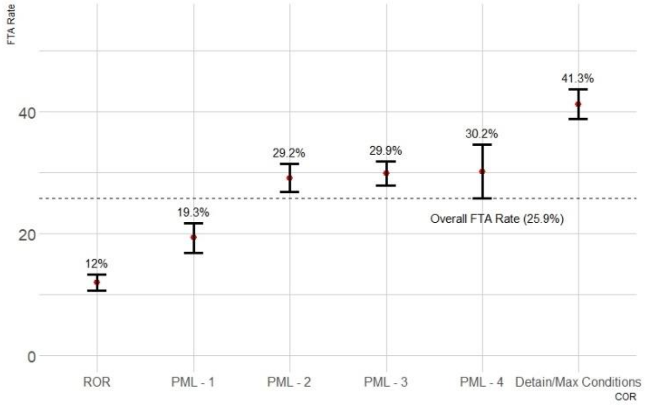
In Figure 3, we present the FTA rate by the collapsed release recommendation category. Interestingly, while FTA rates generally increased conditional on the COR category, there were not statistically significant differences in FTA rates for those in PML – 2 (FTA Rate: 28.6%), PML – 3 (FTA Rate: 29.2%), or PML – 4 (FTA Rate: 29.3%) categories. Those with the Detain/Max Conditions designation had statistically higher FTA rates (FTA Rate: 41.4%) than those in any other release category.

Figure 3.
FTA Rate by Collapsed Recommendation Category



However, looking at the relationship between the release recommendation categories and observed FTA rates is misleading as such an analysis does not account for whether a judge adhered to or deviated from the CORs recommended by the PSA tool. That is, Figure 3 collapses together cases where the PSA's recommended CORs were followed with cases where they were deviated from. To address this issue, in Figure 4, we display the FTA rate by the collapsed recommendation category *only among the subset of* *cases where judges adhered to COR*. Interestingly, we observe a similar mean FTA rate (FTA Rate: 25.9%) relative to the case where deviations from PSA COR recommendations were included in the overall calculation of FTA (FTA Rate: 25.1%).

Figure 4.
FTA Rate by Collapsed Recommendation Category When COR Adhered To



New Criminal Activity Rates

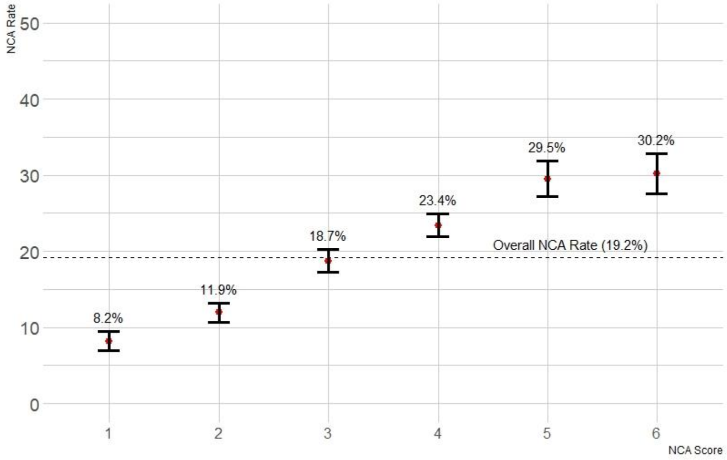
We received data from the BCMC and the SJDC to evaluate whether an NCA occurred during the pretrial period of the case. Violations of city and county ordinances were not considered new criminal activity, and the NCA rate includes both violent and non-violent criminal activity. The overall NCA rate for the outcome sample was 19.2%. In Table 17, we present NCA rates for each recommendation category. As with FTA, as scores within the DMF increased, observed NCA rates increased.

Table 17.
NCA Rates by PSA COR Recommendations

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	ROR 8.3% (220)	ROR 10.9% (165)				
FTA 2	ROR 7.8% (42)	ROR 9.6% (128)	ROR – PML 1 15.9% (184)	ROR – PML 3 20.7% (128)	ROR – PML 4 66.7% (2)	
FTA 3		ROR – PML 1 12.5% (171)	ROR – PML 2 17.4% (381)	ROR – PML 3 19.4% (361)	ROR – PML 4 28.1% (41)	Detain/Max Conditions 25.0% (4)
FTA 4		ROR – PML 1 18.9% (76)	ROR – PML 2 22.1% (205)	ROR – PML 3 25.3% (273)	ROR – PML 4 26.9% (232)	Detain/Max Conditions 29.4% (45)
FTA 5		ROR – PML 2 33.3% (18)	ROR – PML 2 27.1% (97)	ROR – PML 3 27.1% (362)	Detain/Max Conditions 32.0% (368)	Detain/Max Conditions 28.1% (169)
FTA 6				Detain/Max Conditions 28.2% (116)	Detain/Max Conditions 28.1% (113)	Detain/Max Conditions 31.3% (391)

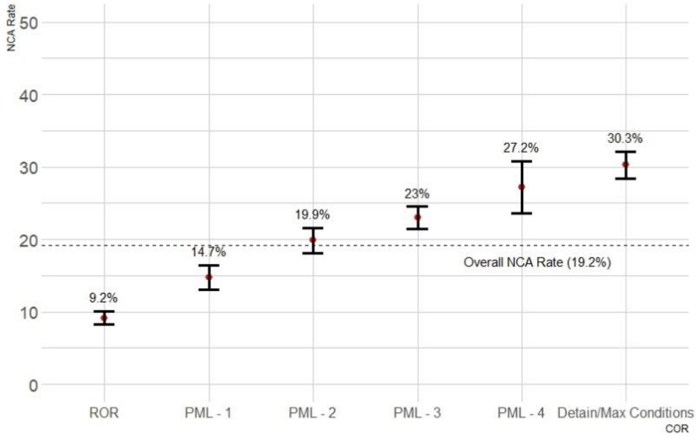
In Figure 5, we show variability in the observed NCA rate by scores on the NCA scale with associated 99% confidence intervals. Results suggest a strictly increasing and positive relationship between scores on the NCA scale and observed NCA rates as well as statistically significant increases in NCA rates across increasing score categories, except for there not being statistically significant differences in NCA rates between those who scored a five or six on the scale. For example, individuals who scored a one on the NCA scale engaged in new criminal activity approximately 8% of the time whereas individuals who scored a six on the NCA scale engaged in new criminal activity approximately 30% of the time. Like the pattern observed with the FTA scale, the NCA scale seems to accurately map onto observed NCA rates, suggestive of the scale's predictive validity.

Figure 5.
NCA Rate by NCA Score



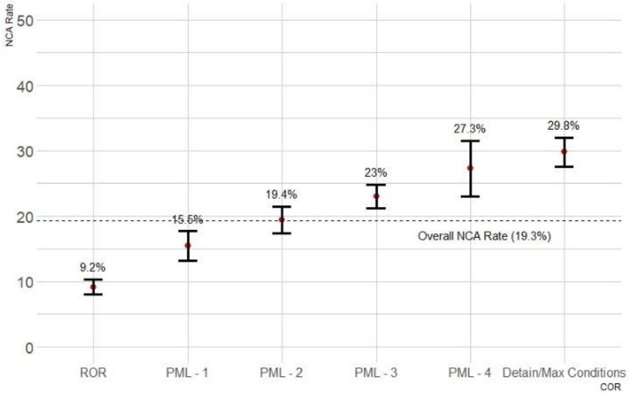
In Figure 6, we present the NCA rate by the collapsed release recommendation category. As CORs became more restrictive, NCA rates increased. For example, individuals who were recommended to be released on their own recognizance without restrictions had an NCA rate of approximately 9%, whereas those who were detained or had maximum restriction conditions recommended had an NCA rate of approximately 30%.

Figure 6.
NCA Rate by Collapsed PSA COR Recommendation Category



However, the same adherence issue is present with the visualization in Figure 6 as was with FTA. That is, Figure 6 collapses together cases where the PSA's recommended CORs were followed with cases where they were deviated from. To address this issue, in Figure 7, we display the NCA rate by the collapsed recommendation category *only among subset of cases where judges adhered to the PSA's recommended COR*.

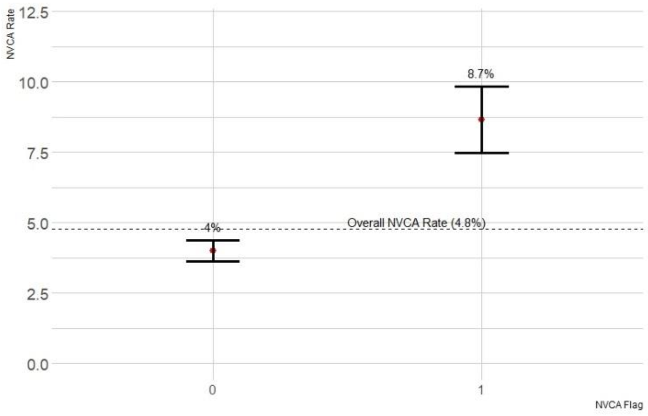
Figure 7.
NCA Rate by Collapsed Recommendation Category When COR Adhered To



New Violent Criminal Activity Flag

Unlike the FTA and NCA scales, the propensity to engage in new violent criminal activity was measured using a dichotomous flag (0 = No NVCA Risk; 1 = NVCA Risk). The observed NVCA rate within the outcome sample was 4.8%. In Figure 8, we display the NVCA rate by the NVCA flag with 99% confidence intervals. Whereas individuals who did not receive the NVCA flag had an observed NVCA rate of 4.0%, individuals with the NVCA flag had an observed NVCA rate of approximately 8.7%.

Figure 8.
NVCA Rate by NVCA Flag



New Criminal Activity and Charge Details

Approximately 19% (n = 4,307) of individuals in the outcome sample engaged in NCA during their pretrial period. It is important to explore the characteristics of the NCA to observe both (1) the type of NCA engaged in and (2) the severity of the NCA engaged in in relation to the source charge. To this end, a 2022 National Academies report on the limitations to existing measurements of recidivism states, "Limiting analyses to simple, binary outcomes (whether someone did or did not engage in criminal behavior following release) without disaggregating by measures of severity or other salient correlates is an approach that should be avoided" and recommends that researchers "supplement binary recidivism measures with measures of desistance from crime such as the frequency and seriousness of offense and length of time until a new offense" (National Academies, 2022). For this reason, we report on the severity of new offenses in relation to the source charge in this section.

To develop the NCA charge information, we collected and assigned a charge level and category for an individual's top three related charges within the pretrial period. We selected the highest charge from those three and classified these as either a first-degree felony (F1), a second-degree felony (F2), a third-degree felony (F3), a fourth-degree felony (F4), misdemeanor (MD), or petty misdemeanor (PM). We then classified charge categories into one of six crime type categories: violent, drug, property, DWI, and public order or other crimes. We present the nexus of NCA highest charge level with charge category in Table 17.

Table 18 shows that 47% (n = 2,029) of individuals in the outcome sample who committed an NCA within the pretrial period committed a fourth-degree felony and 39% (n = 1,681) of individuals in the outcome sample who committed an NCA within the pretrial period engaged in property crimes. Interestingly, 39% of individuals in the sample had a less serious NCA charge than their source charge (i.e., 28% committed a misdemeanor as their NCA offense; 11% committed a petty misdemeanor.)

Table 18.
NCA Highest Charge Level and Category

	Violent	Drug	Property	DWI	Public Order	Total	% of All NCAs
F1	18	4	0	0	0	22	1%
F2	90	74	14	0	4	182	4%
F3	223	41	133	1	5	403	9%
F4	340	789	820	4	76	2,029	47%
MD	301	140	509	34	234	1,218	28%
PM	102	18	205	14	114	453	11%
Total	1,074	1,066	1,681	53	433	4,307	100%
% of All NCAs	25%	25%	39%	1%	10%	100%	--

In Table 19, we show the relationship between the collapsed PSA release recommendation category and NCA charge severity.

Table 19.
NCA Charge Level by Collapsed Recommendation Category

	F1	F2	F3	F4	MD	PM	Total
ROR	7	26	66	216	177	63	555
ROR – PML 1	2	18	37	201	128	46	432
ROR – PML 2	4	30	79	345	177	70	705
ROR – PML 3	2	52	98	569	299	109	1,129
ROR – PML 4	2	16	24	133	74	26	275
Detain/Max	5	40	99	565	363	139	1,211
Total	22	182	403	2,029	1,218	453	4,307

In Table 20, we show how the severity of the NCA within the pretrial period varies in relation to the severity of the source case charge level. Table 20 suggests that for those individuals who committed an NCA in the pretrial period, a majority (50%; n = 2,135) committed a new criminal activity that was less severe than the source case charge for which they were charged.

Table 20.
Source Charge Levels and NCA Charge Level Comparisons

Source Charge Level	NCA Charge Lower		NCA Charge Same		NCA Charge Higher		Total
	Count	Percent	Count	Percent	Count	Percent	
F1	16	100%	0	0%	0	0%	16
F2	287	92%	21	7%	3	1%	311
F3	599	83%	86	12%	41	6%	726
F4	1,233	38%	1,572	49%	428	13%	3,233
Total	2,135	50%	1,679	39%	472	11%	4,286

Following recent work by Moore, Guerin, and Ferguson (2024), we present NCA rates disaggregated by charge severity by PSA NCA score in Figure 9 specifically among the subset of individuals who had an

NCA within the pretrial period and in Figure 10 for the full revalidation sample. Figures 9 and 10 show the NCA rate by NCA score for first-degree to fourth-degree felonies, misdemeanors, and petty misdemeanors.

Figure 9 shows that approximately 40% of all NCAs, independent of score category, were for misdemeanors or petty misdemeanors. For example, defendants with NCA scores of 5 and 6 had NCA rates of 29.5% and 30.2% respectively, yet, from Figure 10 these defendants only rates of felony rearrest of 17.7% and 18.1% respectively. Of new felony charges, the vast majority were fourth-degree felonies. Even for defendants labeled by the PSA as being at the highest risk for NCA (i.e., individuals with NCA score 6), rearrest for high-level felonies was rare (i.e., less than 3% for first-degree through third-degree felonies).

Figure 9.
NCA Rates by Severity According to PSA NCA Score Among Those who Had an NCA

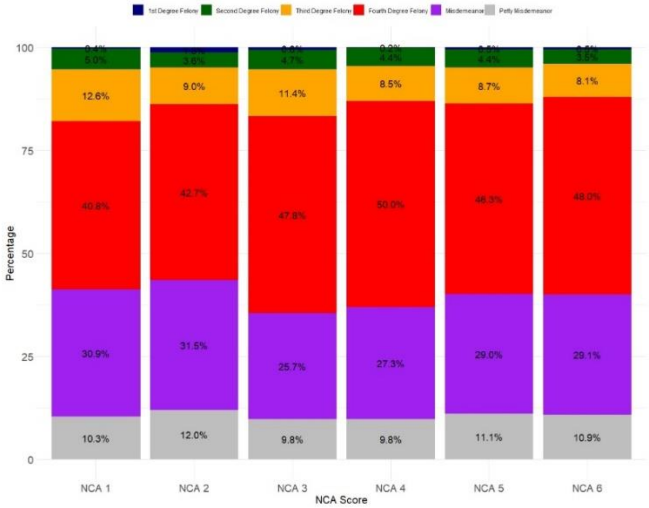
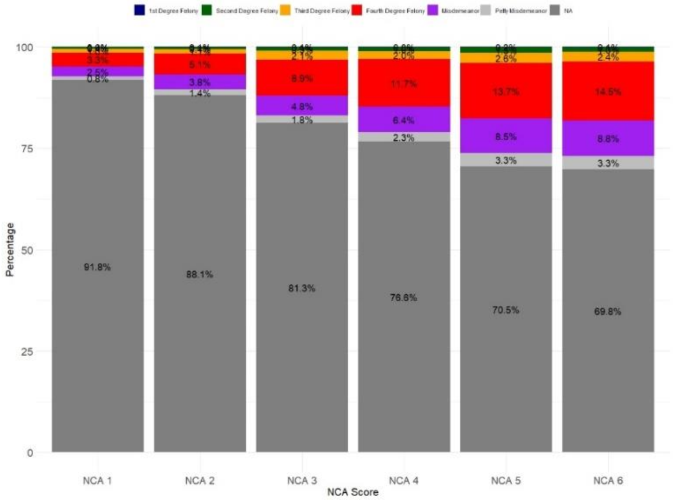


Figure 10.
NCA Rates by Severity According to PSA NCA Score for Full Sample



Predictive Validity of the PSA

One way we evaluated the predictive validity of the PSA was using Area Under the Curve (AUC) estimates, a common accuracy metric used in risk assessment diagnostics. The AUC score, which ranges from 0.0 to 1.0, measures how well the PSA distinguishes between successful and unsuccessful pretrial outcomes, with a score of 0.50 indicating that the PSA performs no better than random chance, and values above 0.50 indicating improved predictive accuracy. For example, an AUC of 0.65, means that when drawing two random cases from the data set, one of which had the pretrial failure and the other did not, between 65% of the time the case that had the pretrial failure would have a higher score on the PSA than a randomly selected individual who did not fail. Consistent with prior work, we default to the use of Desmarais and Singh's 2013 guidelines for interpreting what the AUC scores signal about predictive validity where scores of < 0.55 indicate "poor" predictive validity, scores of 0.55 – 0.63 indicate "fair" predictive validity, scores of 0.64-0.71 indicate "good" predictive validity, and scores of 0.71 – 1.00 indicate "excellent" predictive validity.

Overall Predictive Validity - AUCs

We present AUC ROC estimates for the overall PSA score as well as the FTA, NCA, and NVCA scale and flag at predicting FTA, NCA, and NVCA outcomes alongside 99% confidence intervals in Table 21. Using Desmarais and Singh's benchmarks as a reference point, results suggest that the overall PSA score had "good" predictive performance predicting both FTA and NCA and "fair" predictive performance predicting NVCA. These results are consistent with the AUC results obtained in our original validation study which reported FTA, NCA, and NVCA AUCs of 0.64,0.64, and 0.57 respectively (Ferguson et al., 2021). Substantively, our findings indicate that when drawing two random cases from the outcome dataset, one of which had the pretrial failure and the other did not, between 59% and 67% of the time the case with the pretrial failure would have a higher score on the PSA than a randomly selected individual who did not fail across all failure types.

We note briefly why the NVCA flag may have lower predictive validity than FTA and NCA scales. First, NVCA was coded as a dichotomous flag whereas FTA and NCA were coded as scores. Collapsing more information from the NVCA score into a dichotomous flag removes potentially useful granular predictive information from the model which increases the potential for uncertainty in predictive validity estimates (i.e., the reason why the NVCA scale has higher AUC than the flag). Additionally, AUCs for NVCA may be lower than AUCs for FTA and NCA due to a statistical issue called class imbalance which can emerge in cases in which events are rarer. With class imbalance, there are fewer positive examples compared to negative examples (i.e., higher volumes of successes than failures observed). As a result, the model generating the AUC estimates has less exposure to positive instances during training which can lead to a less accurate, statistically-noisier estimation of the true underlying distribution of the positive class (i.e., why the confidence intervals for NVCA are consistently larger than for NCA and FTA).

Table 21.
Area Under the Curve Receiver Operator Characteristics

Outcome - Scale	**AUC Score with 99% CIs (n = 22,387)**
FTA – PSA Score	0.668 [0.657 – 0.678]
FTA – FTA Scale	0.664 [0.653 – 0.674]
NCA – PSA Score	0.643 [0.631 – 0.654]
NCA – FTA Scale	0.636 [0.623 – 0.647]
NVCA – PSA Score	0.587 [0.564 – 0.609]
NVCA – NVCA Scale	0.623 [0.601 – 0.645]
NVCA – NVCA Flag	0.571 [0.553 – 0.590]

Overall Predictive Validity – Odds Ratios

Another approach commonly used in the literature to evaluate model performance is to use logistic regression to estimate odds ratios (ORs) to evaluate whether there are significantly different odds of failure outcomes conditioning on relevant variables. While there are some statistical limitations to the use of logistic regression in these contexts (e.g., among other things, logistic regression requires that the independent variables are linearly related to the log odds of the dependent variable), multiple visual diagnostic tests and model-fit statistics revealed the linearity assumption was not violated across multivariate logistic models for FTA and NCA outcomes[17].

In Table 22, we present the ORs from six multivariate logistic regressions which predicted FTA, NCA and NVCA outcomes as a function of FTA, NCA, and NVCA scale scores or collapsed PSA scale scores, detention length, exposure length, race, and sex, while controlling for time by including fiscal-year fixed effects. Substantively, after adjusting for other factors not embedded within the PSA, an OR of 1.50 for FTA indicates that there was a 50% increase in the odds of failing to appear at court for each one-point increase in the score on the six-point FTA scale. Similarly, an OR of 1.33 for NCA indicates that there was a 33% increase in the odds of new criminal activity for each one-point increase in the score on the NCA scale. An OR of 2.06 for NVCA indicates that there was a 106% increase in the odds of new violent criminal activity when a defendant had the NVCA flag relative to when they did not. For each one unit increase in the six-point Collapsed PSA Scale which corresponds to the DMF, there was a 40% increase in odds of failing to appear at court, a 28% increase in the odds of engaging in new criminal activity, and a 17% increase in the odds of engaging in new violent criminal activity.

Table 22.
Odds Ratios of FTA, NCA, and NVCA Scale Scores Controlling for Other Factors[18]

	Dependent variable:					
	FTA	NCA	NVCA	FTA	NCA	NVCA
	(1)	(2)	(3)	(4)	(5)	(6)
FTA Scale	**1.50*****	--	--	--	--	--
	(0.02)	--	--	--	--	--
NCA Scale	--	**1.33*****	--	--	--	--
	--	**(0.02)**	--	--	--	--
NVCA Scale	--	--	**2.06*****	--	--	--
	--	--	**(0.15)**	--	--	--
PSA Score	--	--	--	**1.40*****	**1.28*****	**1.17*****
	--	--	--	**(0.02)**	**(0.01)**	**(0.02)**

[17] First, we created partial residual plots as a visual diagnostic method to evaluate whether expected and observed residuals of the PSA scales deviated from one another and to see whether the fit of the observed residuals was non-linear. For FTA models, results suggested alignment and linearity of the expected and observed residuals. Second, we estimated different model specifications (e.g., squaring and cubing the scaled score term) and compared model fit statistics (e.g., AICs) across the three different specifications. Results suggested negligible improvements to model fit by including these transformations of the score variables within the model (AIC Original - FTA: 13,311.57; AIC Squared – FTA: 13,306.75; AIC Cubed - FTA: 13,307.35) suggesting that the use of logistic regression is appropriate in the present context for FTA. Similarly, cubing and squaring the NCA scale did not result in appreciable improvements in model fit statistics (AIC Original: 18,679.2; AIC Squared – NCA: 18,650.6; AIC Cubed – NCA – 18,649.80).

[18] Reference category for detention length is < 1 week. Reference category for exposure length is <= 2 weeks. Reference category for Race is Hispanic. Reference category for Gender is Male. Reference category for Adherence is Adhered. Reference category for fiscal year is FY 2018. Note that in logistic models, we excluded cases where exposure length was less than <1 week.

Detention Length: 1-2 Weeks	1.17^{*}	1.16^{**}	1.45^{***}	1.21^{**}	1.14^{**}	1.30^{**}
	(0.10)	(0.08)	(0.16)	(0.10)	(0.08)	(0.14)
Detention Length: 2-3 Weeks	1.43^{***}	1.01	1.27	1.50^{***}	0.99	1.07
	(0.17)	(0.10)	(0.21)	(0.18)	(0.10)	(0.18)
Detention Length: 3-4 Weeks	0.79^{*}	1.02	1.10	0.83	0.98	0.94
	(0.10)	(0.11)	(0.20)	(0.10)	(0.10)	(0.18)
Detention Length: 4+ Weeks	0.34^{***}	0.77^{***}	0.96	0.36^{***}	0.76^{***}	0.82
	(0.03)	(0.06)	(0.14)	(0.03)	(0.06)	(0.12)
Exposure Length: 2 – 4 Weeks	2.28	2.37^{***}	2.15^{***}	2.25	2.37^{***}	2.24^{***}
	(1.21)	(0.31)	(0.52)	(1.20)	(0.31)	(0.54)
Exposure Length: 4 – 6 Weeks	8.37^{***}	1.75^{***}	1.66^{**}	8.33^{***}	1.75^{***}	1.75^{**}
	(4.26)	(0.23)	(0.40)	(4.25)	(0.23)	(0.42)
Exposure Length: 6 – 8 Weeks	10.40^{***}	1.31^{**}	1.03	10.29^{***}	1.33^{**}	1.14
	(5.27)	(0.17)	(0.25)	(5.21)	(0.17)	(0.27)
Exposure Length: 8 – 10 Weeks	100.70^{***}	3.25^{***}	2.55^{***}	97.30^{***}	3.25^{***}	2.60^{***}
	(51.09)	(0.45)	(0.64)	(49.34)	(0.45)	(0.66)
Exposure Length: 10+ Weeks	382.19^{***}	3.70^{***}	2.61^{***}	377.34^{***}	3.69^{***}	2.57^{***}
	(192.37)	(0.45)	(0.60)	(189.86)	(0.45)	(0.59)
Race: Black	1.03	1.05	1.21^{*}	1.01	1.06	1.26^{**}
	(0.09)	(0.07)	(0.13)	(0.08)	(0.07)	(0.14)
Race: Native American	1.36^{***}	0.79^{***}	1.07	1.36^{***}	0.79^{***}	1.18
	(0.11)	(0.06)	(0.12)	(0.11)	(0.06)	(0.13)
Race: White	0.99	1.04	0.85^{**}	1.00	1.05	0.86^{**}
	(0.05)	(0.04)	(0.06)	(0.05)	(0.04)	(0.07)
Sex: Female	1.06	0.86^{***}	0.70^{***}	1.15^{***}	0.85^{***}	0.69^{***}
	(0.05)	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)
Adherence: No COR	0.21^{***}	0.74^{***}	0.82^{*}	0.20^{***}	0.76^{***}	1.08
	(0.02)	(0.05)	(0.09)	(0.01)	(0.05)	(0.12)
Adherence: Less Restrictive COR	0.85^{*}	0.85^{**}	0.94	0.96	0.86^{**}	0.90
	(0.07)	(0.06)	(0.13)	(0.08)	(0.06)	(0.12)
Adherence: More Restrictive COR	0.69^{***}	1.01	0.97	0.69^{***}	1.02	1.03
	(0.06)	(0.07)	(0.11)	(0.06)	(0.07)	(0.12)
Fiscal Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Constant	0.0003^{***}	0.04^{***}	0.01^{***}	0.0004^{***}	0.05^{***}	0.02^{***}
	(0.0002)	(0.01)	(0.003)	(0.0002)	(0.01)	(0.005)
Observations	19,959	19,959	19,959	19,959	19,959	19,959
Log Likelihood	-6,632.79	-9,316.60	-3,843.08	-6,677.97	-9,303.51	-3,856.81

Akaike Inf. Crit.	13,311.57	18,679.20	7,732.17	13,401.94	18,653.01	7,759.61
Note:						*p**p***p<0.01

Overall Predictive Validity – Random Forest Models

It is also important to evaluate how the risk factors included within each of the FTA, NCA, and NVCA scales predict outcomes of interest. We tried evaluating this through an examination of bivariate correlations reported in the original validation study. However, one weakness of examining correlations and trying to infer the importance of a given relationship between a risk factor and failure outcome is the issue of confounding, which can mask the relative importance of different variables.

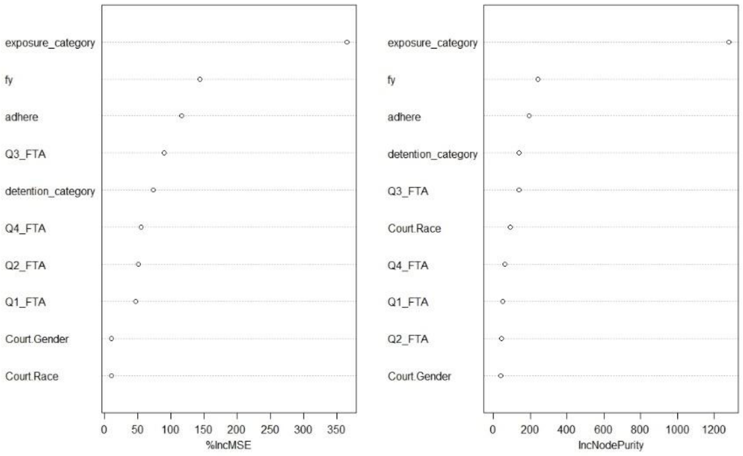
Because of this, we estimated a series of machine-learning random forest models. In short, random forest models work by creating a group of decision trees, each trained on a different random sample of the full sample data and relevant variables. From the random forest models, we produced a set of variable importance plots (VIPs). These variable importance plots (VIPs) can tell one what effect removing a given variable from the regression model would have on increasing the mean squared error (MSE). When removing a variable from a model produces a higher MSE, this means that the variable is “more important” for the prediction of the outcome. In this way, we can better understand the relative importance of the different failure risk factors in predicting the specific outcomes of interest. To our understanding, this is the first validation or revalidation study of the PSA which has explored both (1) the relative importance of each risk factor to predicting failure outcomes and (2) how the risk factors rank in comparison to other potentially predictive factors not included within the PSA scale itself.

Figures 11-14 show the VIPs for FTA, NCA, and NVCA outcomes respectively derived from random forest models which predicted failure outcomes as a function of scores on each scale’s risk factors, adherence, race, gender, detention length, exposure length, and fiscal year of PSA assessment. Specifically, readers should examine the %IncMSE portion of the visualization, which illustrates to what degree the removal of the specific variable from the random forest model has on increasing the mean squared error. Higher values of %IncMSE indicate that the specific variable is “more important”.

Findings from Figure 11 align with the multivariate logistic regression results reported in Table 22. While we did not comment in detail on it in the context of our discussion of the logistic regression results, both exposure length and the fiscal year of PSA assessment impacted FTA rates. It makes intuitive sense that exposure time would be particularly important for predicting FTA rates, as more exposure time introduces more opportunity for failure to be realized, though of course this is a factor which is only observed post-assessment and thus would not be a suitable candidate for scale inclusion. Relatedly, different court-level policy choices which increased hearing prevalence (e.g., the increased use of status hearings during FY 2021 and FY 2022) and case completion time (e.g., the increased median case completion length in FY 2023) also increased FTA opportunity in specific years.

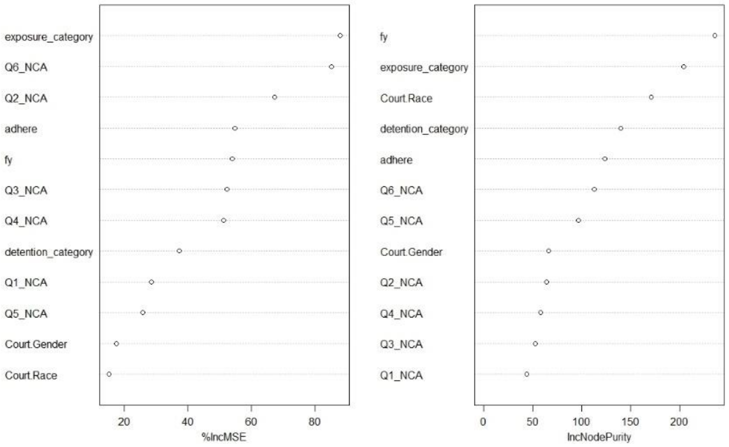
In terms of the factors specifically on the FTA scale, whether an individual had an FTA in the past two years (Q3_FTA) was the most predictive factor in terms of its effect on MSE relative to the other factors, which tended to have roughly similar effects on MSE reduction by excluding it from the prediction model.

Figure 11.
Variable Importance Plot from Random Forest Model Predicting FTA[19]



Findings from Figure 12 similarly align with the regression results reported in Table 22. The two most important factors within the scale in terms of error reduction were having a prior sentence served to incarceration (Q6_NCA) and having a current charge pending (Q2_NCA). Of the factors within the scale, age at current arrest (Q1_NCA) and having a prior violent conviction (Q5_NCA) were the least important for the model's performance. The finding with respect to age at current arrest is consistent with the finding of a negligible impact of age on NCA we reviewed in Table 4.

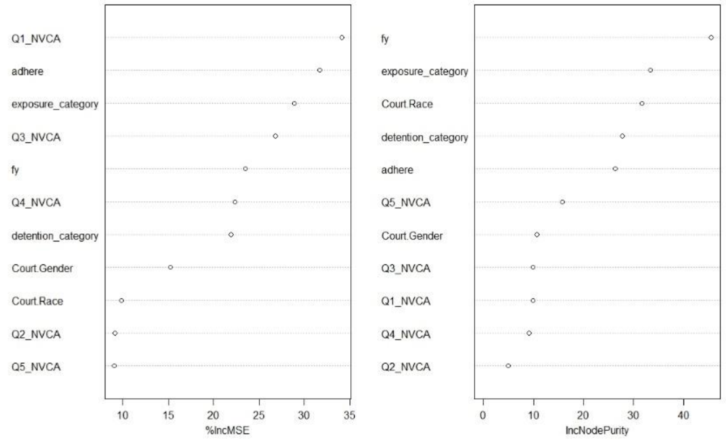
Figure 12.
Variable Importance Plot from Random Forest Model Predicting NCA[20]



[19] Q1_FTA = "Pending Charge at Time of Arrest". Q2_FTA = "Any Prior Conviction (Misdemeanor or Felony". Q3_FTA = "Prior FTA in Past Two Years". Q4_FTA = "Prior FTA Older Than Two Years".
[20] Q1_NCA = "Age at Current Arrest". Q2_NCA = "Pending Charge at Time of Arrest". Q3_NCA = "Prior Misdemeanor Conviction". Q4_NCA = "Prior Felony Conviction" . Q5_NCA = "Prior Violent Conviction". Q6_NCA = "Prior Sentence to Incarceration."

Findings from Figure 13 similarly align with the regression results reported in Table 12. The two most important factors in terms of error reduction were having a current violent offense (Q1_NVCA) and having a pending charge (Q3_NVCA). Of the factors within the scale, having a current violent offense and being < 20 Years Old (Q2_NVCA) and having a prior violent conviction (Q5_NVCA) were the least important for the model's performance. The finding with respect to age at current arrest is also consistent with the finding of a negligible impact of this risk factor we reviewed in Table 4.

Figure 13.
Variable Importance Plot from Random Forest Model Predicting NVCA[21]



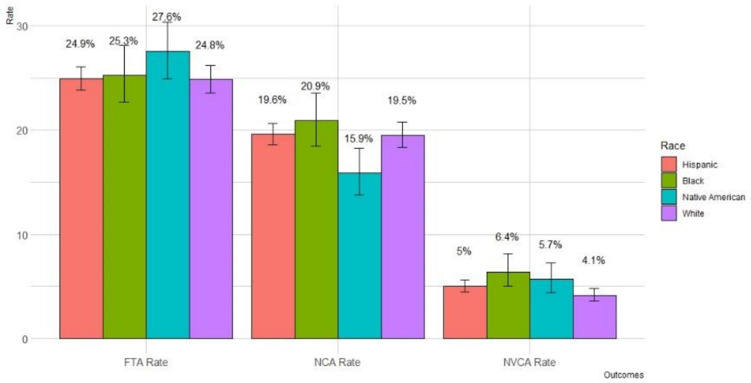
Outcome Measures by Race and Gender

It is important to explore how predictive validity varies by race and gender. Before evaluating predictive validity across groups, it is helpful to first get a sense of what failure base rates look like across different racial and gender groups. However, it is important to note that differences in failure base rates across different groups are not in and of themselves indicators of prediction bias.

In Figure 14, we show variation in the observed FTA, NCA, and NVCA base rates across race-ethnicity groups with 99% confidence intervals. Results suggest that while FTA rates were not significantly different across racial groups, there were some statistically significant differences across racial groups for NCA and NVCA, a pattern which deviates from the results of our 2021 validation report. For example, Native American individuals in the outcome sample had significantly lower NCA rates (NCA Rate: 15.9%) compared to Blacks (NCA Rate: 20.9%), Hispanics (NCA Rate: 19.6%), and Whites (NCA Rate: 19.5%). Similarly, White individuals in the outcome sample had statistically significantly lower NVCA rates (4.1%) than Blacks (NVCA Rate: 6.4%) and Hispanics (NVCA Rate: 5.0%). However, it is an open question as to whether the magnitude of these differences rises to the level of substantive significance (i.e., while statistically significant, is a 3.6% difference in NCA rates between Native Americans and Whites practically significant?)

[21] Q1_NVCA = "Current Violent Offense". Q2_NVCA = "Current Violent Offense and < 20 Years Old". Q3_NVCA = "Pending Charge". Q4_NVCA = "Any Prior Conviction (Misdemeanor or Felony)". Q5_NVCA = "Prior Violent Conviction

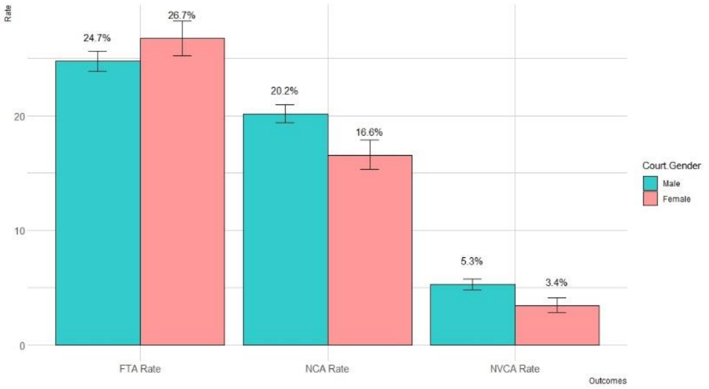
Figure 14.
Outcome Measures by Race-Ethnicity



We also used a chi-square (χ2) test of independence to evaluate the relationship between race and outcomes of interest[22]. Results indicated that while there was not a statistically significant association between race and FTA [χ2 (3, N=20,811) =6.1, p=0.11], there were statistically significant associations between race and NCA [χ2 (3, N=20,811) =16.8, p < 0.00] and NVCA [χ2 (3, N=20,811) =19.5, p < 0.00].

In Figure 15, we show variation in the FTA, NCA, and NVCA rates by gender with 99% confidence intervals. Figure 15 shows that while FTA rates were not significantly different between males and females, males had significantly higher NCA and NVCA rates (NCA Rate: 20.2%; NVCA Rate: 5.3%) than females (NCA Rate: 16.6%; NVCA Rate: 3.4%).

Figure 15.
Outcome Measures by Gender



We conducted a chi-square (χ2) test of independence to determine if there was a relationship between gender and outcomes of interest[23]. Results indicated that there were statistically significant associations

[22] Excluding the “Other Race” category reduced the sample size to 20,835.
[23] Excluding the “Other Race” category reduced the sample size to 22,292.

between gender and FTA [χ2 (2, N=22,247) =8.6, p=0.003], NCA [χ2 (2, N=22,247) =35.2, p <0.00] and NVCA [χ2 (2, N=22,247) =30.1, p <0.00].

Predictive Validity by Race and Gender

To evaluate predictive validity by race, we calculated AUC scores for the PSA by race. We present AUC scores with 99% confidence intervals for FTA, NCA, and NVCA in Table 23.

For FTA and NCA outcomes, the 99% confidence intervals overlap across all racial groups, implying no significant differences in predictive validity of the PSA by race, suggesting that the tool is well-calibrated. Per Desmarais and Singh's (2013) interpretative guidelines, the PSA would either be classified as having "good" or "fair" predictive performance for all failure outcomes, as point estimates of the AUC range from a low of 0.626 to a high of 0.684.

However, for the NVCA flag, the 99% confidence interval for Black defendants (0.471 - 0.589) is the only confidence interval which crosses the 0.50 threshold (i.e., an estimate as good as chance), and the point estimate of the AUC (0.530) is within the threshold of "Poor" classifier, suggesting that while the NVCA flag's predictive validity is not significantly different than other racial groups, the NVCA flag may be less predictively valid for the prediction of NVCA for Black individuals compared to other racial groups.

Table 23.
AUC ROC Scores and 99% CIs by Race

Outcome - Scale	**Hispanic (n = 10,170)**	**White (n = 7,169)**	**Native American (n = 1,782)**	**Black (n = 1,690)**
FTA – PSA Score	0.665 [0.649 – 0.680]	0.663 [0.645 – 0.682]	0.666 [0.629 – 0.703]	0.672 [0.634 – 0.711]
FTA – FTA Scale	0.661 [0.645 – 0.677]	0.659 [0.641 – 0.678]	0.662 [0.626 – 0.698]	0.665 [0.627 – 0.703]
NCA – PSA Score	0.626 [0.608 – 0.643]	0.644 [0.623 – 0.665]	0.684 [0.642 – 0.726]	0.649 [0.607 – 0.690]
NCA – NCA Scale	0.621 [0.604 – 0.639]	0.641 [0.621 – 0.661]	0.665 [0.624 – 0.707]	0.633 [0.593 – 0.674]
NVCA – PSA Score	0.572 [0.539 – 0.604]	0.593 [0.550 – 0.636]	0.628 [0.558 – 0.699]	0.549 [0.476 – 0.622]
NVCA – NVCA Scale	0.617 [0.585 – 0.648]	0.618 [0.573 – 0.663]	0.669 [0.599 – 0.738]	0.578 [0.509 – 0.647]
NVCA – NVCA Flag	0.556 [0.530 – 0.582]	0.585 [0.550 – 0.619]	0.626 [0.561 – 0.692]	0.530 [0.471 – 0.589]

We also computed AUCs for the PSA by gender. We present AUC scores for FTA, NCA, and NVCA for males and females in Table 24. AUC performance was generally "good" or "fair" for the prediction of FTA and NCA and was "fair" for the NVCA flag. There were not meaningful statistically significant differences in the AUCs of male and female defendants, suggestive of predictive parity despite significantly different base rates of NCA and NVCA between the two groups.

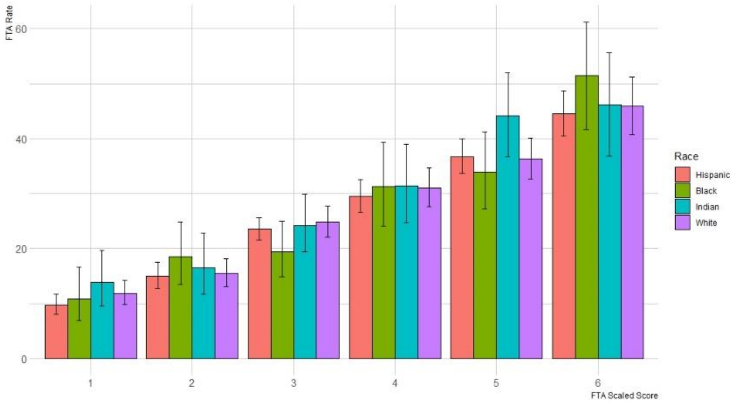
Table 24.
AUC ROC Scores and 99% CIs by Gender

Outcome - Scale	**Males** **(n = 16,648)**	**Females** **(n = 5,599)**
FTA – PSA Score	0.669 [0.657 – 0.681]	0.663 [0.641 – 0.684]
FTA – FTA Scale	0.663 [0.651 – 0.676]	0.662 [0.642 – 0.683]
NCA – PSA Score	0.637 [0.623 – 0.650]	0.651 [0.626 – 0.675]
NCA – NCA Scale	0.629 [0.616 – 0.642]	0.642 [0.618 – 0.666]
NVCA – PSA Score	0.587 [0.562 – 0.611]	0.566 [0.514 – 0.619]
NVCA – NVCA Scale	0.614 [0.590 – 0.639]	0.629 [0.576 – 0.682]
NVCA – NVCA Flag	0.568 [0.547 – 0.589]	0.570 [0.530 – 0.610]

FTA, NCA, and NVCA Scores and Outcomes by Race and Gender

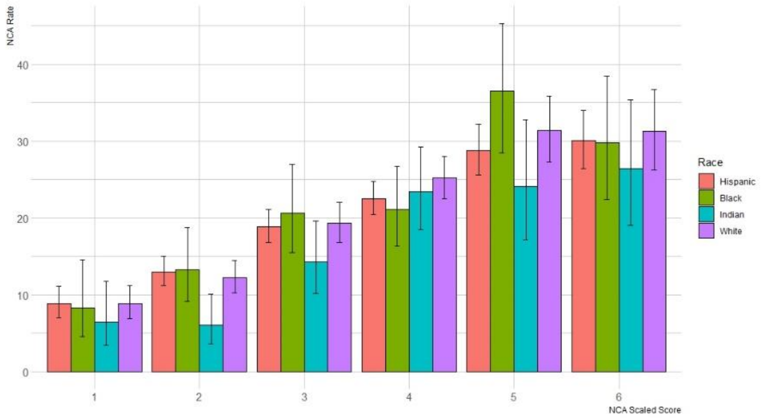
In Figure 16, we compared FTA rates across FTA scale scores by racial categories. Consistent with the findings of our validation study, while there was some variation across racial categories within scoring levels, this variation did not rise to the level of statistical significance. Across all racial groups, as scores on the FTA scale increased, so too did observed FTA rates. Thus, while there were significant inter-score differences in FTA rates within races, there were not intra-score differences in FTA rates across races.

Figure 16.
FTA Rates by Race and FTA Score



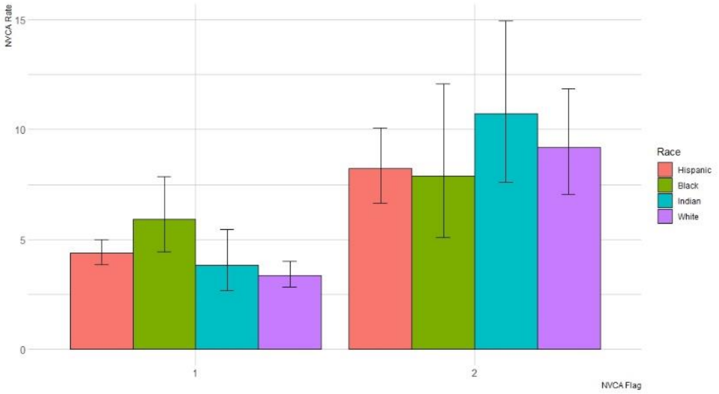
In Figure 17, we compared NCA rates across NCA scale scores by racial categories. Consistent with the findings of our validation study, while there was some variation across racial categories within scoring levels (e.g., Native Americans who scored 2 on the NCA Scale had significantly lower observed NCA rates than Whites and Hispanics), on balance, differences in NCA rates were not significantly different across groups. Across all racial groups, as scores on the NCA scale increased, so too did observed NCA rates.

Figure 17.
NCA Rates by Race and NCA Score



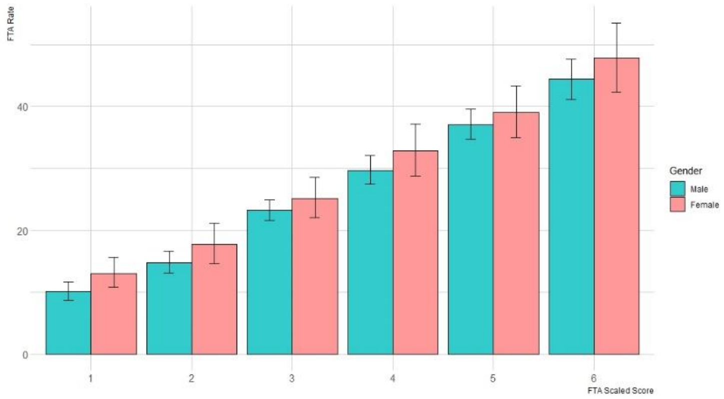
In Figure 18, we compared observed NVCA rates conditional on NVCA flags by racial categories. Across all racial groups, as scores on the NVCA scale increased, so too did observed NVCA rates. Across all groups except for Black defendants, the increase in observed NVCA rates between non-flagged and flagged cases were significantly different.

Figure 18.
NVCA Rates by Race and NVCA Flag



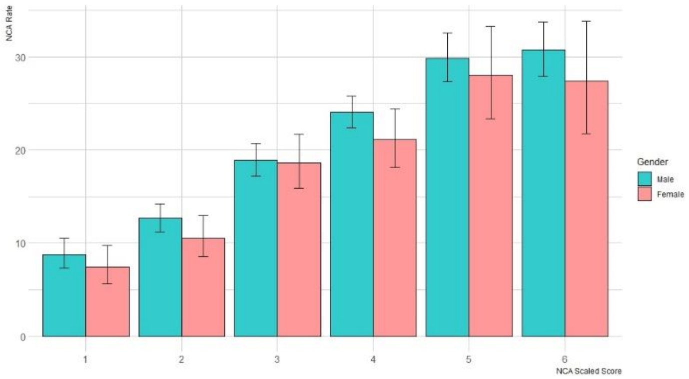
In Figure 19, we compared FTA rates across FTA scale scores by gender categories. As scores on the FTA scale increased, so too did observed FTA rates. Across all score categories, while females had higher rates of FTA than males, the differences in observed rates were not statistically significant.

Figure 19.
FTA Rates by Gender and FTA Score



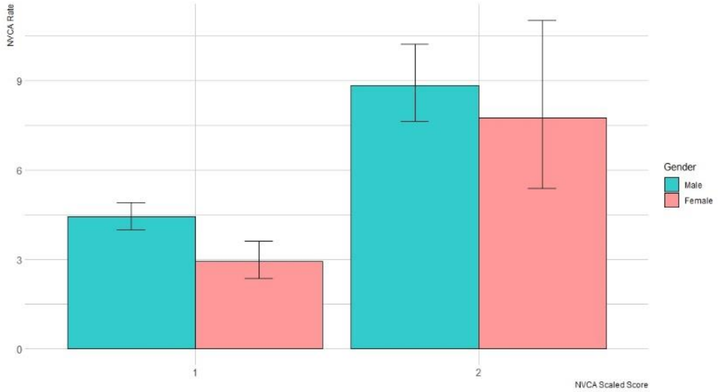
In Figure 20, we compared NCA rates across NCA scale scores by gender categories. As scores on the NCA scale increased, so too did observed NCA rates. Across all score categories, while males had higher rates of NCA than females, the differences in observed NCA rates within levels were not statistically significant, though there were aggregate differences in NCA rates when not conditioned by NCA scores (i.e., males had higher overall levels of NCA than females).

Figure 20.
NCA Rates by Gender and NCA Score



In Figure 21, we compared observed NVCA rates conditional on NVCA flags by gender. For both males and females, having an NVCA flag was associated with significant increase in observed NVCA rates relative to non-flagged cases. Interestingly, males without an NVCA flag had significantly higher rates of observed NVCA than females whereas NVCA rates were not significantly different between males and females with the NVCA flag.

Figure 21.
NVCA Rates by Gender and NVCA Flag



Predictive Fairness by Race and Gender: Odds Ratios

Odds Ratios by Race

In Table 25, we report odds ratios for all failure outcomes from subgroup logistic regression evaluating the odds of failure for each unit increase in the outcome-specific score by racial category[24]. Results suggest statistically similar odds of FTA across all racial groups. Substantively, an OR of 1.47 for FTA indicates that there was a 47% increase in the odds of failing to appear at court for each one-point increase in the score on the FTA scale for Hispanic defendants. Similarly, results suggest statistically similar odds of NCA by NCA scores across all racial groups. For example, an OR of 1.34 for NCA for Hispanics indicates that there was a 34% increase in the odds of engaging in new criminal activity for each one-point increase in the score on the NCA scale. However, we observe the most variance across racial groups in failure odds with the NVCA flag as well as more uncertainty in the estimates of the odds ratios. For example, Hispanics had a 1.96 OR for NVCA indicating that going from not receiving the NVCA flag to receiving the flag was associated with a 96% increase in the odds of new violent criminal activity. Notably, as observed in prior analyses, specifically for the NVCA flag, the 99% confidence interval for the odds ratio of Black defendants crossed 1.0, implying that there was not a statistically significant difference in the odds of engaging in NVCA for Black defendants conditional on receipt of the NVCA flag (i.e., that the NVCA flag does not allow one to accurately forecast NVCA for Black defendants).

[24] In Appendix B, we report supplementary results of a multivariate logistic model which predicted pretrial failure outcomes but which, in addition to the explicitly modelled interaction between race and scale scores, controlled for adherence, detention length, exposure time, gender, and fiscal year. The only point of difference in results between the bivariate subgroup regressions and the more in-depth multivariate model is that the latter finds evidence of a statistically significant interaction between the NVCA flag and gender. The interaction indicates that NVCA_Scaled_Score has a stronger relationship with the likelihood of pretrial failure for females than for males. Specifically, if the NVCA_Scaled_Score increases by one unit, the odds of pretrial failure increase more for females than they do for males by a factor of 1.86. However, this is consistent with the fact that while there were not significantly different base rates between males with and without the NVCA flag, there were significantly different base rates between females with and without the NVCA flag.

Table 25.

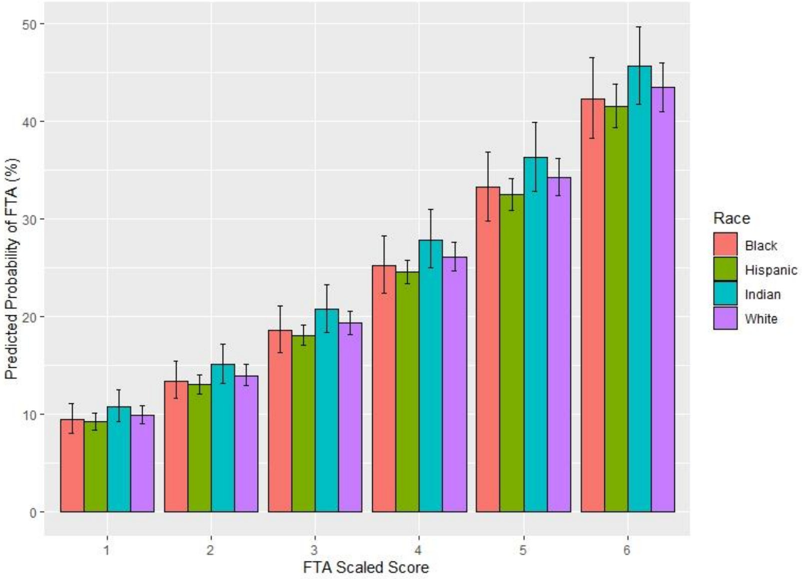
Odds Ratios and 99% Confidence Intervals by Race

Race	FTA OR [99% CI]	NCA OR [99% CI]	NVCA OR [99% CI]
Hispanic	1.47 [1.41 – 1.53]	1.34 [1.28 – 1.40]	1.96 [1.49 – 2.54]
White	1.44 [1.38 – 1.51]	1.40 [1.33 – 1.48]	2.90 [2.05 – 4.04]
Native American	1.45 [1.33 – 1.60]	1.46 [1.31 – 1.64]	3.01 [1.77– 5.15]
Black	1.48 [1.34 – 1.63]	1.36 [1.23 – 1.52]	1.37 [0.76 – 2.37]

In Figure 22, we plot the predicted probability of FTA conditional on race from the logistic models with 99% confidence intervals across different FTA scale scores. When comparing the predicted probabilities from the logistic model with the observed FTA rates by race we saw in Figure 10, we note that there were not significant differences across figures in the observed or estimated FTA rates.

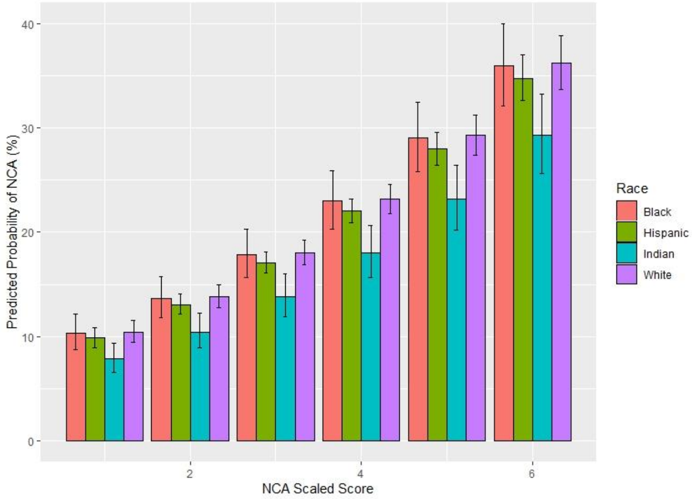
Figure 22.

FTA Rates and Logistic Fit by Race



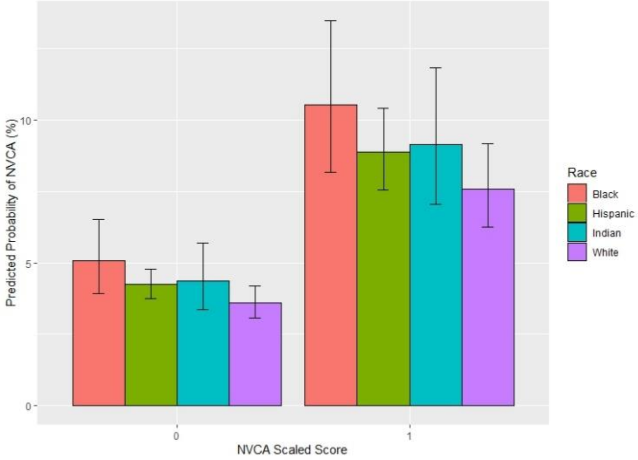
In Figure 23, we plot the predicted probability of engaging in NCA conditional on race from the logistic models with 99% confidence intervals across different NCA scale scores. When comparing the predicted probabilities from the logistic model with the observed NCA rates by race we saw in Figure 11, we note that the logistic model generates some different estimates of the NCA rate relative to the observed rates. While both the observed and expected data suggest that within scores, there were not statistically significant differences across races in NCA rates, on balance, the logistic model tends to overestimate NCA rates across groups (e.g., whereas Whites who scored at Level 6 on the NCA scale engaged in NCA at a rate of 31%, the logistic model predicted they engaged in NCA at a rate closer to 43%). The difference between the observed and fitted data may reflect the influence of unobserved variables in the logistic regression which intersect with race to impact NCA rates (e.g., omitted variable bias).

Figure 23.
NCA Rates and Logistic Fit by Race



In Figure 24, we plot the predicted probability of engaging in NVCA conditional on race from the logistic models with 99% confidence intervals by the NVCA flag. When comparing the predicted probabilities from the logistic model with the observed NVCA rates by race we saw in Figure 12, we note that the logistic model generates some different estimates of the NVCA rate relative to the observed rates. While both the observed and expected data suggest that within scores, there were not statistically significant differences across races in NVCA rates, on balance, the logistic model tends to overestimate NVCA rates for Blacks, Hispanics, and Whites and underestimate the NVCA rate of Native Americans. Again, the difference between the observed and fitted data may reflect the influence of unobserved variables in the logistic regression which intersect with race to impact NVCA rates (e.g., omitted variable bias).

Figure 24.
NVCA Rates and Logistic Fit by Race



Odds Ratios by Gender

In Table 26, we report odds ratios for all failure outcomes derived from subgroup logistic regression evaluating the odds of failure for each unit increase in the outcome-specific scale score by gender[25].

[25] In Appendix B, we report supplementary results of a multivariate logistic model which predicted pretrial failure outcomes but which, in addition to the explicitly modelled interaction between gender and scale scores, controlled for adherence, detention

Results suggest statistically similar odds increases of FTA, NCA, and NVCA as scores on the scales increase between males and females. For example, recall that we can interpret an NCA OR of 1.39 for females as indicating that there is a 39% increase in the odds of engaging in new criminal activity for each one-point increase in the score on the NCA scale.

Table 26.

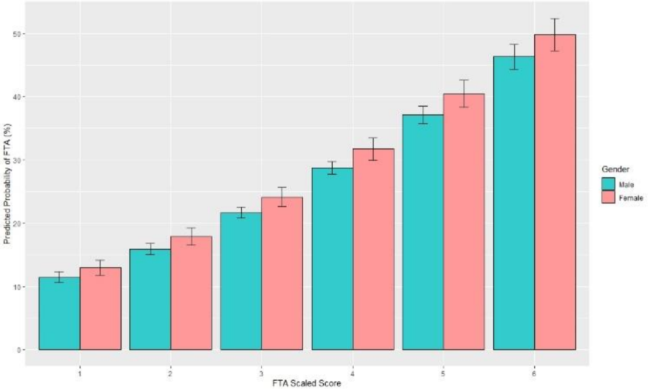
Odds Ratios and 99% Confidence Intervals by Gender

Gender	FTA OR [99% CI]	NCA OR [99% CI]	NVCA OR [99% CI]
Male	1.47 [1.43 – 1.52]	1.36 [1.32 – 1.41]	2.09 [1.72 – 2.53]
Female	1.44 [1.37 – 1.51]	1.39 [1.31 – 1.48]	2.79 [1.76 – 4.31]

In Figure 25, we plot the predicted probability of engaging in FTA conditional on gender from the logistic models with 99% confidence intervals across different FTA scale scores. When comparing the predicted probabilities from the logistic model with the observed FTA rates by gender we saw in Figure 10, we note that the logistic model tends to underestimate actual FTA rates specifically for FTA Scaled Scores 1 - 4. However, both fitted and observed data suggests that within each score category, there were not statistically significant differences by gender in FTA rates. Moreover, results suggest a positive relationship between FTA scores and estimated FTA rates.

Figure 25.

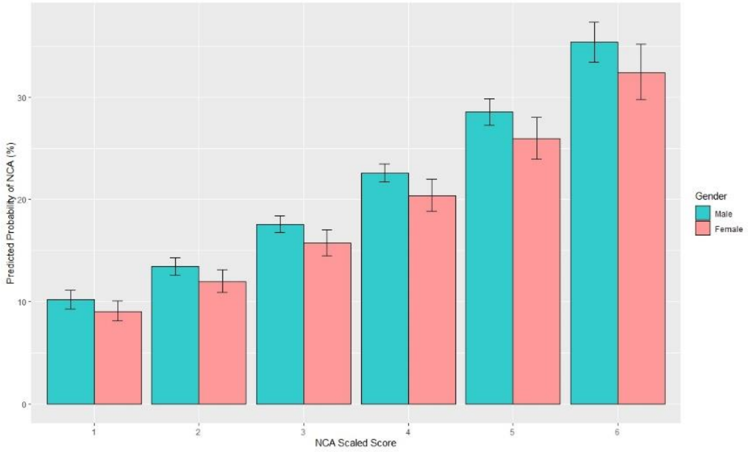
FTA Rates and Logistic Fit by Gender



Similarly, in Figure 26, we plot the predicted probability of engaging in NCA conditional on gender from the logistic models with 99% confidence intervals across different NCA scale scores. When comparing the predicted probabilities from the logistic model with the observed FTA rates by gender we saw in Figure 10 and compare against the FTA findings, we see that the logistic model tends to do a more accurate job estimating NCA rates by gender, though there are some marginal differences between expected probabilities and observed rates. However, both fitted and observed data suggests that within each score category, there were not statistically significant differences by gender in NCA rates. Moreover, results suggest a positive relationship between the NCA score and estimated NCA rates.

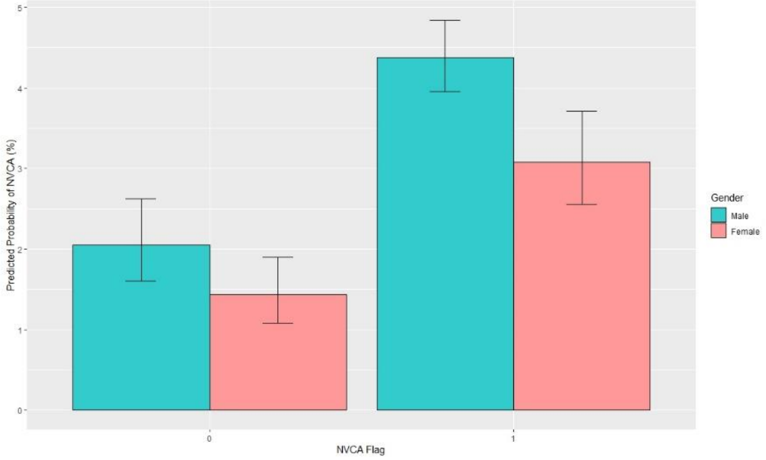
length, race, and year. The only point of difference in results between the bivariate subgroup regressions and the more in-depth multivariate model is that the latter finds evidence of a statistically significant interaction between the NVCA flag and gender. The interaction indicates that NVCA_Scaled_Score has a stronger relationship with the likelihood of pretrial failure for females than for males. Specifically, if the NVCA_Scaled_Score increases by one unit, the odds of pretrial failure increase more for females than they do for males by a factor of 1.86. However, this is consistent with the fact that while there were not significantly different base rates between males with and without the NVCA flag, there were significantly different base rates between females with and without the NVCA flag.

Figure 26.
NCA Rates and Logistic Fit by Gender



In Figure 27, we plot the predicted probability of engaging in NVCA conditional on gender from the logistic models with 99% confidence intervals across different NVCA flag. When comparing the predicted probabilities from the logistic model with the observed NVCA rates by gender we saw in Figure 15, we see that the logistic model tends to do a relatively accurate job estimating NVCA rates by gender. Both fitted and observed data suggests that for those with the NVCA flag, males had higher rates of NVCA than females, even though there was not a statistically significant difference by gender in NVCA for those who did not receive the NVCA flag. Moreover, results suggest a positive relationship between receipt of the NVCA flag and estimated NVCA rates.

Figure 27.
NVCA Rates and Logistic Fit by Gender



Adherence

As we noted in our discussion about the limitations of only looking at the two-way relationship between PSA COR recommendations and different failure outcomes, while a bivariate relationship can provide a first-take look at the relationship between two variables, it does not account for other factors that might explain the relationship (i.e., the fact that judges can choose to impose supervision or release conditions that are either more or less restrictive than those recommended by the PSA DMF which can, in consequence, impact failure rates). Accordingly, it is important to evaluate the scope of adherence to the

DMF COR recommendations and see whether the choice to adhere to or deviate from the DMF COR recommendations correlated with failure rates.

In Table 27, we showed how adherence varied as a function of different collapsed PSA COR recommendation categories. Across the subset of cases considered (i.e., excluding cases where there were no conditions of release assigned), the overall adherence rate was 79.2% (n = 15,567). In 8.8% of cases, judges imposed conditions of release which were less restrictive than the PSA recommended (n = 1,737), and in 12.0% of cases, judges imposed conditions of release which were more restrictive than the PSA recommended (n = 2,349). Judges tended to impose more restrictive conditions in cases where the DMF matrix recommendation was "Detention/Max" categories (i.e., 19.4% of all cases were deviated from in the direction of more restrictive conditions). We conducted a chi-square test which revealed a significant association between adherence and FTA [X2 (2, N=14,383) =110.59, p=<.001], although the size of this effect was relatively small (Cramer's V= 0.087)[26].

Table 27.
Adherence or Deviation by Collapsed PSA COR Recommendation Category

PSA Category	Total	% of All Adherence Cases	Less Restrictive	Adherence	More Restrictive
ROR	5,077	25.8%	0.0%	82.7%	17.3%
ROR – PML 1	2,558	13.0%	29.7%	67.0%	3.2%
ROR – PML 2	3,159	16.1%	13.4%	81.0%	5.6%
ROR – PML 3	4,295	21.9%	7.7%	83.5%	8.8%
ROR – PML 4	896	4.6%	5.8%	80.6%	13.6%
Detain/Max	3,668	18.7%	4.6%	76.0%	19.4%
Total	19,653	100%	8.8%	79.2%	12.0%

Study Limitations

There are some limitations to the analysis we present in the revalidation report worth considering. First, given the nature of the data and the breadth of some reported analyses (i.e., two-way, bivariate analyses) (e.g., our analysis of the raw correlations between tool factors and failure outcomes), we were not able to make any *causal* claims. Relatedly, while the use of multivariate regression can reduce concerns about potential confounding, these models are still limited with respect to the variables we included which could theoretically impact the relationship between the PSA and failure outcomes (e.g., there may be factors beyond those included in the models which influence the relationship between PSA scores and failure outcomes). Moreover, the specific methodology we used does not allow us to answer causal questions about the impact of different factors on failure outcomes (e.g., we cannot say definitively that when judges deviate from PSA COR recommendations, this *causes* a reduction in failure likelihoods, even though the two factors are correlated). Thus, we caution readers against extracting strong causal conclusions, as findings are correlative and thus suggestive.

Second, given the uniqueness of Bernalillo County and the uniqueness of the use of the PSA within Bernalillo County, the results are not generalizable to other municipalities within New Mexico or

[26] The logistic regression results we report in Table 22 suggest that adherence, after statistically adjusting for other potential factors, is related to some types of failure in unique ways. Specifically, relative to cases where judges adhered to PSA COR recommendations, cases when judges deviated (i.e., specifically, when judges imposed more restrictive COR than what the PSA recommended) tended to be associated with significantly lower odds of FTA (OR: 0.69). Cases where judges imposed less restrictive COR than what the PSA recommended were associated with significantly lower odds of NCA (OR: 0.85). We invite readers to revisit the discussion of Table 15 for a more detailed analysis of the effect of adherence on failure outcomes.

elsewhere. The overall results (i.e., AUC metrics) align with those from other jurisdictions (DeMichele et al., 2020). Bernalillo County is unique in several ways relative to other jurisdictions in which the PSA has been previously validated (e.g., the scope of ethnic heterogeneity; scope of polysubstance use; geographic scale; the use of PSA exclusively for felony cases and not misdemeanors).

Recommendations

Results of the revalidation suggest a series of recommendations as well as pathways toward future analyses worth considering.

1. As noted in recent snapshot reports of Bernalillo County's MDC (Ferguson & Goldberg, 2023), the MDC jail population has aged over time (e.g., the median age of the MDC population in 2017 was 32; the median age of the MDC population in 2023 was 34). That the median age of defendants has increased over time and that the relative proportion of defendants who scored within the younger thresholds for age-embedded factors on the NCA and NVCA scales has decreased over time (i.e., <22 years old for NCA; having a current violent offense and being < 20 years old for NVCA) suggests there may be increasing uncertainty surrounding the relationship between these factors and failure over time (i.e., as the sample size of the number of younger defendants has increasingly decreased over time, there may be greater statistical uncertainty surrounding the relationship between these factors with failure rates when each subsequent year is analyzed in isolation). Moreover, replicating a finding of the 2021 validation report, we presented evidence that the age factor on the NCA scale and the age factor on the NVCA scale were not significantly correlated with failure outcomes at either the bivariate or multivariate level of analysis. For these reasons, we advise that the AOC analyze in more detail the relationship between defendant age and failure rates. Following such an analysis, if practicable, we advise either revising the scoring on the PSA tool for the existing factors which embed age (i.e., reducing the points assigned to individuals who scored at the younger age factors on the NCA and NVCA factors), changing the thresholds for age on these factors to better reflect the correlation of age and failure locally (to the extent that such a correlation is apparent in the data), or deleting the factors from the tool given their lack of predictive power.
2. We found that while there were not statistically significant differences across racial categories in the AUC scores for the NVCA flag, the AUC score for the NVCA flag for Black defendants crossed the 0.50 threshold, signaling that the NVCA flag performed no better than random chance at predicting NVCA failure for Black defendants. It is important for follow up studies to explore why this may be the case. One possibility is that unobserved factors may influence the relationship between race and failure rates (i.e., the regression which generated the AUC estimates only explored the bivariate relationship between PSA scores and failure outcomes within racial groups). Another possibility is that the relationship between race and failure is interactive (i.e., the analysis which generated the AUC does not account for heterogeneity *within* the racial category – say, by gender – which could also impact the overall predictive quality of the flag within race). Probing potential explanations for why the NVCA flag had poor predictive performance among Black defendants would be an interesting next step for research.
3. To our awareness, not much research has explored the impact of Covid-19 and related policy changes on pretrial failure rates. One thing we observed in the data was that over time (i.e., particularly following the 2021 NM Supreme Court ruling which mandated status case reviews), FTA rates significantly increased between the pre-Covid and during-Covid periods alongside this change as this policy change, in tandem with lengthier case processing times, directly increased failure opportunity (e.g., 19.3% FTA for the pre-Covid period to 33.2% FTA for the post-Covid period). It may be helpful for the AOC to engage in more regular monitoring of failure rates over time (i.e., annual reporting) so that they can more closely-in-time understand the potential impact of such policy changes on failure outcomes. We also would advise conducting more analyses of the effects of different Covid-era policies on failure outcomes to isolate the effects of different

types of policy changes on failure outcomes (i.e., partial out the extent to which the increase in observed failure rates reflected an increased proportion of virtual versus in-person hearings versus increased opportunities to failure because of increased status case volume versus natural increases or statistical noise).

4. Figures 8 and 9 of the report find that while NCA rates increased with NCA score categories, the types of NCA defendants committed within the pretrial period tended not to be of the highest severity levels (i.e., there were low volumes of first-degree through third-degree felonies committed when defendants were released pretrial across all PSA NCA risk scores). Following Moore, Ferguson, and Guerin (2024), we suggest that future research into the relationship between the PSA risk scores and pretrial NCA or NVCA failure continue to disaggregate pretrial criminal failure by charge severity instead of defaulting to the use of binary indicators of recidivism, as the binary approach to the measurement of NCA fails to capture important nuance in the broader public safety narrative surrounding pretrial risk assessments (i.e., that most individuals, even those at highest risk of NCA, did not commit NCA and that of the minority who did commit an NCA while released, most NCAs were either of a similar charge severity or less severe charge severity than the source charge).

Conclusions

In the present report, we revalidated the PSA within Bernalillo County using a sample of 22,387 cases of 15,831 unique defendants spanning July 1, 2017, through June 30, 2023. Results of the revalidation suggested the PSA, as used through June 2023, was a predictively valid instrument for evaluating risk of pretrial failure in Bernalillo County with AUC performance in the range of 0.58 – 0.69, signaling “fair” to “good” levels of predictive validity. Generally, as scores on the FTA and NCA scales increased, observed failure rates increased. For FTA and NCA outcomes, we found no statistically significant differences in the PSA's predictive validity by race, suggesting reasonable calibration of the tool, with AUC estimates generally classified as "good" or "fair". FTA and NCA AUCs were generally "good" or "fair," and "fair" for the NVCA flag for both male and female defendants. Statistically, there were no significant differences in AUCs between male and female defendants, despite differing base rates of NCA and NVCA between genders. Our primary conclusions about the predictive validity and calibration of the PSA replicate those we reported in our 2021 validation study.

However, our results suggest there are ways in which the PSA could be optimized locally with respect to (1) how risk factors which have lower predictive power are weighted within the tool (e.g., age-related factors embedded within the NCA and NVCA scales have low predictive value) and (2) the limits of the predictive power of the NVCA flag specifically for Black defendants (i.e., the 99% confidence interval of the AUC for Black defendants crossed the 0.50 threshold, suggesting the use of the NVCA flag for Black defendants may be less informative than chance at predicting possibility of pretrial NVCA). In exploring variation by race, we also replicated a finding from our 2021 validation report that found for Native American defendants, the risk of NCA was significantly lower when conditioning by scores on the NCA scale (i.e., compared to defendants of other racial groups who scored the same as Native American defendants, on balance, Native American defendants had significantly lower rates of NCA for the same score). These results suggest potential pathways forward toward modifying the PSA tool locally which the AOC may consider (e.g., considering revising the age range used for PSA factors where age range is included and/or downweighing this factor; consider upweighting or rescoring factors which were consistently more predictive of failure per random forest results; consider presenting judges with the overall NVCA score instead of flag given variation by race in the predictive quality of the NVCA flag).

We also present evidence of the effect of the Covid-19 pandemic. For example, results in Appendix A suggest that following the onset of the Covid-19 pandemic, case processing times and case volume – particularly for status hearing cases - increased which, by definition, increased risk of failure by increasing overall exposure time and thus opportunity to fail. We observed this effect primarily in elevated FTA rates for the during and post-Covid 19 subpopulations when compared against the FTA rate observed during the pre-Covid 19 period. In the future, it may be worthwhile for researchers to explore how policy choices related to the Covid-19 pandemic (e.g., the choice to move to virtual hearings) contributed to changes in FTA rates, as there is currently a gap in the research literature on the effects of Covid-19 related policy-changes on different types of pretrial failure outcomes.

References

AdvancingPretrial.org. (2020). About the Public Safety Assessment (PSA). Retrieved from: https://advancingpretrial.org/psa/about/.

AdvancingPretrial.org. (2024). How it works. Advancing Pretrial Policy & Research. Retrieved July 25, 2024, from https://advancingpretrial.org/psa/factors/

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91-93.

Arnold Ventures. (2019). Public Safety Assessment FAQs (PSA 101). Retrieved from https://craftmediabucket.s3.amazonaws.com/uploads/Public-Safety-Assessment-101_190319_140124.pdf.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., & Comfort, M. (2020). Public safety assessment: Predictive utility and differential prediction by race in Kentucky. *Criminology & Public Policy*, 19(2), 409-431.

Desmarais, S., & Singh, J. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States.

Duwe, G., & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections*, 1(3), 155-176.

Ferguson, E., de la Cerda, H., Guerin, P., & Moore, C. (2021). *Bernalillo County public safety assessment validation.* Institute for Social Research, Center for Applied Research and Analysis, University of New Mexico. Retrieved from: http://isr.unm.edu/reports/2021/bernalillo-county-public-safety-assessment-validation-study.pdf.

Georgiou, G. (2019). Weights matter: Improving the predictive validity of risk assessments for criminal offenders. *Journal of Offender Rehabilitation*, *58*(2), 92-116.

Greiner, D. J., Stubenberg, M., & Halen, R. (2020). Validation of the PSA in Harris County, TX.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4), 906-917.

Monahan, J., Skeem, J., & Lowenkamp, C. (2017). Age, risk assessment, and sanctioning: Overestimating the old, underestimating the young. *Law and Human Behavior*, 41(2), 191–201

Moore, C., Ferguson, E., & Guerin, P. (2024). *How Much Risk, and Risk of What? A Closer Look at Pretrial Rearrest and Risk Assessment.* Manuscript submitted for publication.

National Academies of Sciences, Engineering, and Medicine (2022). The Limits of Recidivism: Measuring Success After Prison. Washington, DC: The National Academies Press. https://doi.org/10.17226/26459.

National Association of Pretrial Services Agencies. (2020). Standards on Pretrial Release: Revised 2020. Retrieved from: https://napsa.org/eweb/DynamicPage.aspx?Site=napsa&WebCode=standards.

Skog, A., & Lacoe, J. (2021). *Validation of the PSA in San Francisco*. eScholarship, University of California.

Appendix

Appendix A – Frequencies of Failure Rates and Case Completion by Court and Fiscal Year

Table 1.
Failure Rate by Court

Court	Failure Rate
Metro-Only (2)	21%
District-Only (3)	46%
Metro-District (1)	39%

Table 2.
Failure Rate by Fiscal Year

Fiscal Year	FTA Rate
FY 2018	7.3%
FY 2019	23.2%
FY 2020	24.6%
FY 2021	27.4%
FY 2022	26.4%
FY 2023	36.3%

Table 3.
Average and Median Time for Case Completion by Fiscal Year

Fiscal Year	Average	Median
FY 2018	63.3 days	51 days
FY 2019	129.9 days	53 days
FY 2020	121.7 days	54 days
FY 2021	139.7 days	59 days
FY 2022	172.4 days	56 days
FY 2023	201.3 days	72 days

Table 4.
Average and Median Time to Failure by Fiscal Year

Fiscal Year	Average	Median
FY 2018	45 days	36 days
FY 2019	51 days	31 days
FY 2020	58 days	48 days
FY 2021	71 days	53 days
FY 2022	77 days	51 days
FY 2023	64 days	35 days

Table 5.
Average and Median Time to Failure by Fiscal Year and Court

FY and Court	Average	Median
FY 2018		

Metro-Only (2)	24 days	11 days
District-Only (3)	88 days	50 days
Metro-District (1)	51 days	41 days
FY 2019		
Metro-Only (2)	30 days	15 days
District-Only (3)	89 days	34 days
Metro-District (1)	61 days	43 days
FY 2020		
Metro-Only (2)	48 days	49 days
District-Only (3)	68 days	36 days
Metro-District (1)	73 days	47 days
FY 2021		
Metro-Only (2)	66 days	52 days
District-Only (3)	73 days	60 days
Metro-District (1)	97 days	58 days
FY 2022		
Metro-Only (2)	70 days	51 days
District-Only (3)	96 days	66 days
Metro-District (1)	105 days	55 days
FY 2023		
Metro-Only (2)	44 days	35 days
District-Only (3)	64 days	12 days
Metro-District (1)	158 days	53 days

Table 6.

FTA Rate by Covid-Period

Period	**FTA Rate**	**Total Count**
Pre-Covid	19.7%	10,038
During-Covid	29.7%	11,566
Post-Covid	33.5%	827

Appendix B – Multivariate Interactive Logistic Models by Race[27]

	NCA_YN	NVCA_YN
	(2)	(3)
NCA_Scaled_Score2	1.62***	--
	(0.24)	--
NCA_Scaled_Score3	2.29***	--
	(0.33)	--
NCA_Scaled_Score4	3.33***	--
	(0.46)	--
NCA_Scaled_Score5	4.38***	--
	(0.64)	--
NCA_Scaled_Score6	4.63***	--
	(0.71)	--
NVCA_Scaled_Score	--	2.12***
	--	(0.27)
Race: Black	0.81	1.50***
	(0.30)	(0.24)
Race: Indian	0.61	1.15
	(0.23)	(0.20)
Race: White	1.13	0.85
	(0.20)	(0.09)
Adherence: Less Restrictive COR	0.80***	0.90
	(0.07)	(0.13)
Adherence: More Restrictive COR	1.07	0.96
	(0.08)	(0.12)
Detention Length: 1-2 Weeks	1.08	1.56***
	(0.11)	(0.26)
Detention Length: 2-3 Weeks	1.27*	1.58**
	(0.16)	(0.33)
Detention Length: 3-4 Weeks	1.36**	1.73***
	(0.17)	(0.36)
Detention Length: 4+ Weeks	1.09	1.43**
	(0.11)	(0.25)
Gender: Female	0.90*	0.76***
	(0.05)	(0.08)
Year Fixed Effects	Yes	Yes

[27] FTA results not presented as the logistic model did not converge.

NCA Score 2 X Race: Black	**1.07**	
	(0.48)	
NCA Score 3 X Race: Black	**1.83**	
	(0.75)	
NCA Score 4 X Race: Black	**1.23**	
	(0.49)	
NCA Score 5 X Race: Black	**2.02***	
	(0.84)	
NCA Score 6 X Race: Black	**1.21**	
	(0.52)	
NCA Score 2 X Race: Indian	**0.83**	
	(0.39)	
NCA Score 3 X Race: Indian	**1.08**	
	(0.47)	
NCA Score 4 X Race: Indian	**1.58**	
	(0.66)	
NCA Score 5 X Race: Indian	**1.12**	
	(0.51)	
NCA Score 6 X Race: Indian	**1.21**	
	(0.54)	
NCA Score 2 X Race: White	**0.80**	
	(0.17)	
NCA Score 3 X Race: White	**0.98**	
	(0.20)	
NCA Score 4 X Race: White	**0.96**	
	(0.19)	
NCA Score 5 X Race: White	**0.95**	
	(0.20)	
NCA Score 6 X Race: White	**0.99**	
	(0.23)	
NVCA Flag x Black	**--**	**0.61**
	--	**(0.19)**
NVCA Flag x Indian	**--**	**1.12**
	--	**(0.32)**
NVCA Flag x White	**--**	**1.24**
	--	**(0.26)**
Constant	0.11***	0.05***
	(0.02)	(0.01)

Observations	13,358	13,358
Log Likelihood	-6,315.73	-2,613.33
Akaike Inf. Crit.	12,703.45	5,266.66
Note:		

Appendix C – Multivariate Interactive Logistic Models by Gender

	Dependent variable:		
	FTA_YN	NCA_YN	NVCA_YN
	(1)	(2)	(3)
FTA_Scaled_Score	1.00		
	(2,485.82)		
NCA_Scaled_Score2		1.50***	
		(0.18)	
NCA_Scaled_Score3		2.33***	
		(0.27)	
NCA_Scaled_Score4		3.54***	
		(0.40)	
NCA_Scaled_Score5		4.45***	
		(0.54)	
NCA_Scaled_Score6		4.69***	
		(0.60)	
NVCA_Scaled_Score			1.93***
			(0.19)
Court.GenderFemale	1.00	0.92	0.66***
	(15,510.44)	(0.16)	(0.08)
adhere2	1.00	0.80***	0.90
	(10,674.90)	(0.07)	(0.13)
adhere3	1.00	1.07	0.95
	(10,254.31)	(0.08)	(0.12)
detention_categoryOne to Two Weeks	1.00	1.08	1.56***
	(15,572.00)	(0.11)	(0.26)
detention_categoryTwo to Three Weeks	1.00	1.27*	1.62**
	(20,188.89)	(0.16)	(0.34)
detention_categoryThree to Four Weeks-	1.00	1.37**	1.71**
	(20,298.69)	(0.17)	(0.36)
detention_categoryFour or More Weeks	1.00	1.10	1.44**
	(16,253.78)	(0.12)	(0.26)

Court.RaceBlack	1.00	1.14	1.30*
	(12,097.81)	(0.10)	(0.18)
Court.RaceIndian	1.00	0.73***	1.20
	(11,585.78)	(0.07)	(0.16)
Court.RaceWhite	1.00	1.06	0.89
	(6,844.44)	(0.05)	(0.08)
Year2018	1.00	0.93	0.81
	(10,683.73)	(0.07)	(0.11)
Year2019	1.00	0.86**	0.89
	(10,552.69)	(0.06)	(0.12)
Year2020	1.00	0.95	1.24
	(10,944.55)	(0.07)	(0.16)
Year2021	1.00	0.57***	0.79
	(12,130.79)	(0.05)	(0.12)
Year2022	1.00	0.38***	0.15*
	(33,185.68)	(0.13)	(0.15)
FTA_Scaled_Score:Court.GenderFemale	1.00		
	(4,389.80)		
NCA_Scaled_Score2:Court.GenderFemale		0.94	
		(0.21)	
NCA_Scaled_Score3:Court.GenderFemale		1.09	
		(0.23)	
NCA_Scaled_Score4:Court.GenderFemale		0.84	
		(0.17)	
NCA_Scaled_Score5:Court.GenderFemale		1.13	
		(0.25)	
NCA_Scaled_Score6:Court.GenderFemale		0.98	
		(0.24)	
NVCA_Scaled_Score:Court.GenderFemale			1.86***
			(0.42)
Constant	0.00	0.11***	0.05***
	(0.0000)	(0.01)	(0.01)
Observations	13,358	13,358	13,358
Log Likelihood	-0.0000	-6,321.69	-2,612.35
Akaike Inf. Crit.	36.00	12,695.38	5,260.69
Note:			*p**p***p<0.01