Validation and Predictive Bias Testing of the Public Safety Assessment for Montgomery County, Pennsylvania

Christopher T. Lowenkamp
University of Missouri – Kansas City, Department of Criminal Justice & Criminology
University of Cincinnati, The Center for Justice and Communities

PRETRIAL RISK ASSESSMENT aims to balance public safety with fairness while reducing unnecessary detention (Mamalian, 2011; Desmarais et al., 2021). Decisions made during the pretrial phase significantly influence case outcomes, incarceration rates, and future recidivism (McCoy, 2007; St. Louis, 2023). The Public Safety Assessment (PSA), developed by the Laura and John Arnold Foundation, is an actuarial tool designed to predict three key outcomes: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA), using nine administrative data factors, which eliminates the need for resource-intensive defendant interviews (VanNostrand & Lowenkamp, 2013).

The PSA's development leveraged a large dataset of 746,525 cases from several U.S. jurisdictions, identifying nine robust predictors of pretrial failure, including prior convictions and current charge severity (VanNostrand & Lowenkamp, 2013). Since its introduction, the PSA has been implemented in numerous jurisdictions and subjected to independent validation. Table 1 (next page) summarizes the predictive validity findings from 12 validation studies encompassing 370,222 individuals released pretrial across diverse jurisdictions including Kentucky, Texas, Illinois, Ohio, California, Washington, Georgia, New Mexico, and Montana. Across these studies, the PSA demonstrates moderate predictive accuracy for court appearance and public safety outcomes. The weighted mean AUC for Failure to Appear (FTA) was .65 (95 percent CI [.63, .67]), and for New Criminal Activity (NCA) was .66 (95 percent CI [.65, .68]). Using Rice and Harris's (2005) effect size classifications, these values represent moderate predictive effects. The prediction of New Violent Criminal Activity (NVCA) shows differential performance depending on operationalization: the binary NVCA Flag yielded a weighted mean AUC of .55 (95 percent CI [.53, .57]), representing a small effect, whereas the 6-point NVCA Scale produced a weighted mean AUC of .66 (95 percent CI [.65, .67]), a moderate effect.Unweighted means were comparable. As Table 1 shows, individual study AUCs range in values reflecting the variability in predictive performance across jurisdictions and populations that underscores the importance of local validation.

Local validation is essential to ensure a tool's applicability to specific populations, as predictive performance may vary by jurisdiction due to differences in demographics, legal practices, and base rates of pretrial failure (Desmarais et al., 2021; Desmarais et al., 2022). This study validates the PSA in Montgomery County, PA, assessing its predictive accuracy and potential bias across demographic subgroups. By examining failure rates, AUC-ROC values, and logistic regression models, this research contributes to the evidence base for data-driven pretrial decision-making.

Method

Participants

Data for this validation study were obtained from Montgomery County, Pennsylvania's Pretrial Services, encompassing individuals processed through the pretrial system between 2021-2024. The total sample comprised 4,522 individuals, of whom 1,352 were detained pretrial and 3,170 were released. Among detained individuals, 1,084 (80.2 percent) did not receive a PSA. Among released individuals, 3,147 (99.3 percent) received PSA assessments and served as the analytical sample for predictive validity analyses.

Measures

Outcome Variables

Three binary outcome measures were examined. Failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA), were all coded as 0 or 1, with 0 indicating no failure and a 1 indicating that failure had occurred during the pretrial release period. These data were provided by the Montgomery County Pretrial Services.

PSA Risk Scores

The PSA risk scores were generated from the PSA assessment completed by Montgomery County Pretrial Services staff. All three scores are scaled to a six-point scale. The NVCA scale is further collapsed into a binary flag coded as 0 or 1, with a value of 1 indicating increased risk for NVCA. The PSA NVCA scale score was not available for analyses. As such, all analyses were performed using the binary increased violence risk indicator.

Demographic Variables

Demographic variables included age, sex, race, and ethnicity. Sex was recoded as 0 for female and 1 for male. Race was recoded into three binary variables for analyses: Black was coded as 0 for non-Black defendants and 1 for Black

defendants; White was coded as 0 for non-White defendants and 1 for White defendants, while Other was coded as 0 for Black and White defendants and 1 for those defendants that were not identified as Black or White. Ethnicity was recoded as 0 for non-Hispanic and 1 for Hispanic defendants.

Procedure

Descriptive Analysis

Comprehensive descriptive statistics were calculated for PSA scale scores, and outcomes across the overall sample and demographic subgroups. Cross-tabulations examined outcome rates by PSA scale scores to assess the instruments' ability to differentiate risk levels.

Predictive Validity Analysis

To assess predictive validity, area under the receiver operating characteristic curve (AUC-ROC or ROC) was calculated for each scale and outcome as well as for each subgroup (e.g., differing categories of sex, race, and ethnicity). The AUC-ROC represents the probability that a randomly selected individual who experiences an adverse outcome (e.g., pretrial failure) will receive a higher risk score than a randomly selected individual who does not experience that outcome (Hanley & McNeil, 1982). In practical terms, an AUC of 0.75 means there is a 75 percent chance the instrument will correctly rank a recidivist as higher risk than a non-recidivist. Stated more succinctly, the AUC-ROC measures how well a risk assessment tool distinguishes between individuals who will and will not reoffend. Values range from 0.50 (no better than guessing) to 1.0 (perfect prediction), with 0.71 or higher considered excellent performance (Rice & Harris, 2005).

ROC analysis was conducted for each PSA scaled score (ranging in value from 1 to 6 for the FTA and NCA scales and 0 to 1 for the NVCA flag). AUC values were calculated with 95 percent confidence intervals and interpreted using established benchmarks.

Failure rates were calculated for each scale score level to examine the PSA's ability to distinguish between risk groups. Chi-square tests assessed whether differences in failure rates across scale scores were statistically significant.

Bias Testing

Bias analysis followed the methodology used by Skeem and Lowenkamp (2016), testing for both slope bias (differential prediction) and intercept bias (differential calibration) across demographic subgroups. The subgroup comparisons made included male vs female; Black vs. White defendants; and Hispanic vs. non-Hispanic defendants. The statistical models used to test for slope and intercept bias are presented below:

Intercept Bias Testing:

1. Base Model: $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1(\text{PSA Score})$
2. Subgroup Model: $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1(\text{PSA Score}) + \beta_2(\text{Group})$

Slope Bias Testing:

1. Subgroup Model: $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1(\text{PSA Score}) + \beta_2(\text{Group})$
2. Interaction Model: $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1(\text{PSA Score}) + \beta_2(\text{Group}) + \beta_3(\text{PSA Score} \times \text{Group})$

Likelihood ratio chi-square tests compared model fit, with $\alpha = .01$ used to minimize Type I error given large sample sizes.

Data Analysis

All analyses were conducted using Stata 19.5. Analyses accounted for clustering by individual where appropriate by using clustered error at the individual defendant level. Missing data were handled through listwise deletion, given the administrative nature of the dataset and relatively low rates of missingness.

Results

Sample Characteristics

Table 2 presents the sample characteristics for the released cases included in this study. Males comprised 76.2 percent of cases, Black individuals represented 49.1 percent, and Hispanic individuals represented 6.3 percent.

The distribution of PSA scores among released cases shows the expected pattern of lower-risk individuals being more likely to be released. For the FTA scale, 73.9 percent of released individuals scored between 1 and 3, while only 3.4 percent scored 6. Similar patterns emerged for the NCA scale, where 58.4 percent scored between 1 and 3, while only 5.5 percent scored 6. For the violence

TABLE 1
Area Under the Curve (AUC) Values Across PSA Validation Studies

Study	N	FTA	NCA	NVCA Flag	NVCA Scale
Kentucky (DeMichele et al., 2020)	164,597	.65	.65	—	.66
Harris County, TX (Greiner et al., 2020)	61,603	.60	.66	.55	—
McLean County, IL (Greiner et al., 2020)	6,345	.70	.67	.61	—
Lucas County, OH (Lowenkamp et al., 2020)	19,247	.62	.68	—	.63
Los Angeles, CA (Hess & Turner, 2021)	56,465	.73	.72	.53	.67
San Francisco, CA (Skog & Lacoe, 2021)	9,881	.63	.62	.57	.66
Kane County, IL (Greiner et al., 2021)	13,094	.67	.63	.61	—
Bernalillo County, NM (Ferguson et al., 2021)	10,289	.64	.64	.57	—
Pierce County, WA (DeMichele et al., 2023)	4,111	.56	.61	—	.66
Fulton County, GA (DeMichele et al., 2023)	20,209	.62	.65	—	.65
Thurston County, WA (DeMichele et al., 2023)	3,066	.67	.68	—	.64
Missoula County, MT (Pugmire, 2023)	1,315	.69	.68	—	—
Weighted M		.65	.66	.55	.66
95% CI		[.63, .67]	[.65, .68]	[.53, .57]	[.65, .67]
Unweighted M		.65	.66	.57	.65
95% CI		[.62, .67]	[.64, .67]	[.55, .60]	[.64, .66]

Note. N = sample size; M = mean; CI = confidence interval; FTA = Failure to Appear; NCA = New Criminal Activity; NVCA = New Violent Criminal Activity. NVCA Flag refers to the binary (0/1) indicator; NVCA Scale refers to the 6-point weighted scale. Dashes indicate the metric was not reported. Weighted $k = 12$, $N = 370{,}222$ for FTA and NCA; $k = 6$, $N = 157{,}677$ for NVCA Flag; $k = 7$, $N = 277{,}576$ for NVCA Scale.

flag, 85.0 percent of released individuals had no flag present. These distributions suggest that release decisions in Montgomery County were influenced by PSA risk scores, indicating the PSA was likely being used as intended to inform release recommendations.

Cross-Tabulation Results

Table 3 presents the failure rates by PSA scale scores for the three pretrial outcomes: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA). The table demonstrates clear risk gradients across all three assessment dimensions, confirming the PSA's ability to differentiate between individuals at varying risk levels.

For the FTA scale, failure rates show a consistent and substantial progression from the lowest to highest risk categories. Individuals scoring 1 on the FTA scale had a 10.5 percent failure rate among 572 cases, while those scoring 2 had a 19.6 percent failure rate among 917 cases. The failure rates continue to increase through the middle categories, with score 3 showing 24.6 percent failure among 857 cases and score 4 showing 29.0 percent failure among 418 cases. The highest risk categories demonstrate particularly elevated failure rates, with score 5 showing 49.6 percent failure among 276 cases and score 6 showing 51.4 percent failure among 107 cases. This represents nearly a five-fold increase in failure rates from the lowest to highest risk categories, indicating strong discrimination between risk levels.

The NCA scale similarly demonstrates effective risk stratification, though with one notable anomaly in the progression. Individuals scoring 1 had an 8.3 percent failure rate among 421 cases, while those scoring 2 had a 16.8 percent failure rate among 476 cases. However, individuals scoring 3 showed a 17.0 percent failure rate among 953 cases, which is nearly identical to the score 2 group despite representing a theoretically higher risk category. This minimal difference between scores 2 and 3 violates the expected monotonic progression where each successive score should correspond to meaningfully higher risk. The scale recovers its discriminative ability in the higher categories, with score 4 showing 29.4 percent failure among 756 cases, score 5 showing 35.7 percent failure among 367 cases, and score 6 showing 49.4 percent failure among 174 cases. The overall progression from 8.3 percent to 49.4 percent represents a six-fold increase in failure rates.

The violence flag, which is a binary indicator rather than a scaled score, shows meaningful discrimination between groups. Among the 2,695 individuals without a violence flag, only 3.0 percent experienced a new violent criminal activity during the pretrial period. In contrast, among the 448 individuals with a violence flag present, 10.0 percent experienced such an outcome. This represents more than a three-fold increase in the likelihood of violent criminal activity when the flag is present, demonstrating the flag's utility in identifying individuals at elevated risk for violence relative to the lower-risk group.

These failure rate progressions validate the PSA's fundamental ability to rank-order individuals by risk level. The clear separation between risk categories for FTA and NCA outcomes, along with the meaningful discrimination provided by the violence flag, supports the instrument's use for pretrial decision-making. The one area of concern is the lack of discrimination between NCA scores 2 and 3, which suggests a potential opportunity for scale refinement. This could involve either collapsing these two categories or recalibrating the scoring algorithm to ensure meaningful differences between all adjacent risk levels. Overall, the failure rate patterns confirm that the PSA successfully identifies individuals at varying levels of pretrial risk in Montgomery County, with higher scores consistently associated with higher rates of adverse outcomes.

TABLE 2.
Sample Characteristics

	Released Cases (*N* = 3,170)	
Measure	*N*	%
Detained	—	—
Male	2,416	76.2
Black	1,555	49.1
Hispanic	201	6.3
FTA Scale		
1	572	18.0
2	917	28.9
3	857	27.0
4	418	13.2
5	276	8.7
6	107	3.4
NCA Scale		
1	421	13.3
2	476	15.0
3	953	30.1
4	756	23.9
5	367	11.6
6	174	5.5
Violence Flag		
No Flag	2,695	85.0
Flag Present	448	14.1

TABLE 3.
Failure Rates by PSA Scale Scores

	FTA		NCA		Violence Flag	
Scale Score	N	% Failure	N	% Failure	N	% Failure
1	572	10.5	421	8.3	2,695	3.0
2	917	19.6	476	16.8	448	10.0
3	857	24.6	953	17.0	—	—
4	418	29.0	756	29.4	—	—
5	276	49.6	367	35.7	—	—
6	107	51.4	174	49.4	—	—

Base Rates by Demographic Groups

Base rate analysis provides important context for understanding potential disparities before examining instrument bias. Table 4 presents the failure rates by demographic subgroups among the 3,170 individuals in the analytical sample.

As shown in Table 4, failure rates were relatively consistent across racial groups, with differences of approximately 2 percentage points or less for all three outcomes. Chi-square tests indicated no statistically significant differences between White and Black individuals for any outcome (all $p > 0.05$). Sex comparisons revealed that males had significantly higher rates of new criminal activity (23.55 percent vs. 19.50 percent, $\chi^2 = 6.84$, $p = 0.009$) and new violent criminal activity (4.22 percent vs. 3.18 percent, $\chi^2 = 4.92$, $p = 0.027$) compared to females, while failure to appear rates were similar. Hispanic individuals showed lower failure rates across all outcomes compared to non-Hispanic individuals, though these comparisons should be interpreted cautiously given the small Hispanic sample size (n = 201), which limits statistical

power and precision of estimates.

These base rate findings indicate relatively equal failure rates across racial groups, modest sex differences favoring females for criminal activity outcomes, and lower failure rates among Hispanic defendants. These baseline differences are important to consider when interpreting subsequent bias analyses, as they represent genuine population differences rather than instrument-induced disparities.

Predictive Validity Results

Table 5 presents the AUC values across demographic subgroups, demonstrating the PSA's predictive accuracy for different populations. Overall, the PSA achieved moderate predictive validity with AUC values of 0.65 (95 percent CI: 0.62-0.68) for the FTA Scale, 0.66 (95 percent CI: 0.64-0.69) for the NCA Scale, and 0.61 (95 percent CI: 0.55-0.67) for the NVCA Flag. These values fall within the "fair" to "good" range according to established criteria and are consistent with PSA performance in other jurisdictions.

Racial comparisons revealed differential performance for the FTA Scale. White individuals showed significantly higher AUC values (.69) compared to Black individuals (.62), with non-overlapping confidence intervals indicating a statistically significant difference (full confidence intervals are available from the author). This pattern of differential prediction by race is consistent with findings in Kentucky (DeMichele et al., 2020) and Los Angeles (Hess & Turner, 2021). The AUC values fall within the small (Black) and moderate (White) effect size ranges per Rice and Harris (2005) classifications, indicating the FTA Scale retains some predictive utility for both groups. The NCA Scale showed more comparable performance across racial groups, with AUC values of .67 for White individuals and .65 for Black individuals, and overlapping confidence intervals. For the NVCA Flag, the pattern reversed slightly, with Black individuals showing an AUC of 0.622 compared to 0.598 for White individuals.

Sex comparisons demonstrated remarkably similar predictive validity across all three outcomes. The FTA Scale performed nearly identically for males (0.653) and females (0.664). The NCA Scale showed comparable performance with AUC values of 0.664 for males and 0.643 for females. The NVCA Flag showed slightly better performance for males (0.620) compared to females (0.574), though both values indicate moderate predictive utility.

Ethnicity comparisons revealed more variable results, though interpretation is limited by the small Hispanic sample size (n = 201). Hispanic individuals showed AUC values of .66 for FTA and .69 for NCA, both comparable to or slightly better than the overall sample performance. However, the NVCA Flag showed notably poor performance for Hispanic individuals, with an AUC of .45, which falls below the threshold for acceptable prediction. This anomalous result likely reflects the extremely low base rate of violent offending in this small subsample (only 1.99 percent failure rate), making reliable prediction difficult. Although AUC-ROC is base rate invariant, stable estimation requires adequate numbers of events; with only four NVCA events among Hispanic defendants, the resulting AUC estimate is unreliable regardless of overall sample size. Non-Hispanic individuals, comprising the majority of the sample, showed consistent performance across all three outcomes (.65 for FTA, .66 for NCA, .62 for NVCA) that closely mirrored overall results.

The consistency of AUC values across most demographic comparisons, with the majority falling between 0.60 and 0.69, suggests that the PSA maintains relatively stable predictive validity across different populations. The one notable exception, the poor NVCA performance for Hispanic individuals, may be an artifact of small sample size, the use of the binary violence risk indicator, and low base rates rather than a systematic bias issue. These findings support the PSA's use as a general pretrial assessment tool while highlighting the importance of adequate sample sizes for reliable validation across all demographic subgroups.

Bias Testing Results

Table 6 presents the bias analysis results examining both slope bias (differential prediction) and intercept bias (differential calibration) across demographic subgroups. Using the Skeem and Lowenkamp (2016) methodology with a significance threshold of $\alpha = 0.01$ (χ^2 critical value = 6.635), none of the bias tests reached statistical significance, providing evidence that the PSA performs equitably across demographic groups.[1]

For the FTA outcome, slope bias tests yielded chi-square values below the critical threshold for all comparisons: Black versus White ($\chi^2 = 5.84$), Hispanic versus Non-Hispanic ($\chi^2 = 0.08$), and Female versus

[1] The $\alpha = .01$ threshold controls for familywise error across multiple bias tests (3 outcomes × 3 demographic comparisons × 2 bias types = 18 tests). This threshold is not conservative given sample sizes exceeding 2,000 in most comparisons; rather, it appropriately adjusts for the increased risk of false positives when conducting multiple tests. Additionally, multivariate models that include PSA scores as predictors have greater statistical power than simple group comparisons because controlling for a strong predictor reduces residual variance. AUC confidence intervals use the conventional $\alpha = .05$ because they describe precision of individual estimates rather than test multiple hypotheses.

TABLE 4.
Base Rates for FTA, NCA, and NVCA by Demographic Group

Subgroup	N	FTA (%)	NCA (%)	NVCA (%)
Race				
White	1,548	25.32	23.00	4.07
Black	1,555	23.22	22.44	3.99
Sex				
Male	2,416	24.30	23.55*	4.22*
Female	754	23.61	19.50	3.18
Ethnicity				
Hispanic	201	18.41	13.43	1.99
Non-Hispanic	2,969	24.52	23.21	4.11

TABLE 5.
AUC Values by Race and Sex

Outcome	Black	White	Hispanic	Non-Hispanic	Male	Female
FTA	0.616	0.689	0.656	0.654	0.653	0.664
NCA	0.648	0.672	0.685	0.658	0.664	0.643
NVCA	0.622	0.598	0.447	0.618	0.620	0.574

Male ($\chi^2 = 0.06$). The odds ratios for the interaction terms ranged from 0.855 to 1.018, all close to 1.0, indicating that the relationship between FTA risk scores and actual failure to appear is consistent across demographic groups. Intercept bias tests were similarly non-significant, with chi-square values ranging from 0.62 to 1.38.

The NCA outcome showed comparable results. Slope bias chi-square values remained well below significance for Black ($\chi^2 = 0.64$), Hispanic ($\chi^2 = 0.18$), and Female ($\chi^2 = 1.30$) comparisons. Intercept bias tests also failed to reach significance across all comparisons, though the Hispanic intercept comparison ($\chi^2 = 6.44$) was the highest value observed across all tests. This value did not exceed the critical threshold of 6.635 and should be interpreted cautiously given the small Hispanic sample size (n = 201).

For the NVCA outcome, all bias tests remained non-significant despite the relatively rare nature of violent offending. Chi-square values for both slope and intercept comparisons were uniformly low across all demographic groups. The Hispanic comparison (slope $\chi^2 = 2.86$, intercept $\chi^2 = 2.12$) should be interpreted cautiously given that only four Hispanic defendants experienced violent offending, rendering both the AUC estimate and bias tests unreliable.

These comprehensive bias testing results provide robust evidence that the PSA does not exhibit systematic bias across racial, ethnic, or sex groups in Montgomery County. The consistency of non-significant findings across all three outcomes and all demographic comparisons strengthens confidence that the instrument performs similarly across groups. This pattern of results supports the continued use of the PSA as an unbiased pretrial assessment tool while emphasizing the importance of ongoing monitoring to ensure unbiased performance is maintained over time.

Discussion

Key Findings

This validation study provides robust evidence supporting the continued use of the Public Safety Assessment in Montgomery County, Pennsylvania. The PSA demonstrates moderate predictive validity, effectively differentiates between individuals at different risk levels, and shows no evidence of systematic bias across racial, ethnic, or sex subgroups. The comprehensive bias testing, which revealed no significant slope or intercept bias across any demographic comparisons, is particularly encouraging given ongoing concerns about algorithmic fairness in criminal justice risk assessment.

Predictive Validity

The AUC values of .65 for FTA, .66 for NCA, and .61 for NVCA observed in Montgomery County align closely with the weighted means from 12 prior PSA validation studies summarized in Table 1 (N = 370,222). Montgomery County's FTA and NCA results exactly match prior weighted means, while the NVCA Flag exceeds the prior weighted mean of .55. Using Rice and Harris (2005) effect size classifications, the FTA and NCA values represent moderate predictive effects. These results also compare favorably to Desmarais et al.'s (2021) systematic review of six pretrial instruments, which found pooled AUC values of .631 for FTA and .626 for NCA using categorical risk levels. The clear progression in failure rates across scale scores, from 10.5 percent to 51.4 percent for FTA, 8.3 percent to 49.4 percent for NCA, and 3.0 percent to 10.0 percent for the violence flag, confirms the instrument's ability to meaningfully differentiate risk levels.

Bias Analysis

The absence of systematic bias across demographic groups represents a critical validation finding. The comprehensive testing revealed no significant slope bias (differential prediction) or problematic intercept bias (differential calibration) for racial, ethnic, or sex comparisons when using appropriate statistical controls for multiple comparisons. While some chi-square values approached significance, none exceeded the critical threshold, and the overall pattern strongly supports equitable performance.

Scale Optimization Considerations

Analysis reveals specific opportunities to improve PSA performance in Montgomery County. First, the NCA scale shows virtually no discrimination between scores 2 and 3, with failure rates of 16.8 percent and 17.0 percent respectively. This violates the principle of monotonic risk progression. Collapsing these categories into a 5-point scale would improve face validity and provide more stable risk estimates through larger category sample sizes while maintaining predictive accuracy.

Second, individual item-level data were not available for this study but have begun to be collected. Future analyses should examine item-level performance to identify which factors drive prediction and whether local reweighting or cutoff adjustments could improve overall accuracy. This includes examining the 6-point NVCA Scale to determine whether modifications to scoring or cutoffs could enhance violence prediction beyond what the binary flag currently achieves. Similar analyses of the FTA and NCA scales could identify opportunities to increase accuracy while ensuring any existing bias is corrected and none is introduced.

Limitations

Several limitations warrant consideration when interpreting these findings. First, the relatively small Hispanic sample (n = 201) limits the precision of estimates for this subgroup, particularly for the rare NVCA outcome where only four individuals experienced violent

TABLE 6.
Bias Analysis Results - Slope and Intercept Comparisons

Outcome	Subgroup	Slope Bias χ^2	Intercept Bias χ^2	OR Interaction	OR Comp Group
FTA	Black	5.84	0.62	0.855	0.934
	Hispanic	0.08	1.38	0.963	0.800
	Female	0.06	1.00	1.018	0.903
NCA	Black	0.64	0.88	0.947	0.921
	Hispanic	0.18	6.44	1.070	0.594
	Female	1.30	1.38	0.912	0.882
NVCA	Black	0.01	0.18	1.009	0.927
	Hispanic	2.86	2.12	1.000	0.508
	Female	0.08	0.14	1.181	0.918

Note. OR = Odds Ratio. Critical value for significance at $\alpha = .01$ is $\chi^2 = 6.635$ (df = 1). None of the slope or intercept bias tests reached statistical significance.

offending. This small sample size makes it difficult to draw definitive conclusions about PSA performance for this population.

Second, the follow-up period varied across cases based on the length of time from pretrial release to case disposition. Individuals with longer pretrial periods had greater opportunity for outcome events to occur, which may introduce variability in outcome detection based on case processing time rather than actual risk differences.

Third, the study examined only pretrial outcomes and did not assess longer-term post-disposition recidivism. While this aligns with the PSA's intended purpose as a pretrial tool, it limits understanding of whether risk predictions remain valid beyond the immediate pretrial period.

Finally, individual item-level data for PSA risk factors were not available for analysis, limiting opportunities for local calibration or refinement. As noted above, these data have begun to be collected.

Implications for Policy and Practice

The validation results support continued PSA implementation in Montgomery County with several evidence-based enhancements. Systematic quality assurance procedures should ensure consistent data collection and scoring practices. Regular monitoring protocols should track PSA performance across demographic groups to identify any emerging disparities. Staff training should emphasize appropriate use of PSA scores as one factor among many in pretrial decision-making, not as deterministic predictors.

Conclusion

This validation study demonstrates that the Public Safety Assessment performs adequately in Montgomery County, Pennsylvania, meeting established standards for predictive validity while showing no evidence of systematic bias across demographic groups. The PSA effectively differentiates between risk levels, provides meaningful improvements over chance prediction, and operates equitably across racial, ethnic, and sex subgroups. These findings support continued implementation with appropriate quality assurance, monitoring, and training procedures.

The study contributes to the growing evidence base supporting actuarial risk assessment in pretrial decision-making while highlighting the critical importance of local validation. As jurisdictions nationwide grapple with pretrial reform, this research demonstrates that properly validated risk assessment tools can support both public safety and fairness.

References

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. https://doi.org/10.1111/1745-9133.12481

DeMichele, M., Tueller, S., Burtch, E., Dawes, D., & Lattimore, P. (2023). *Validation and predictive bias testing of the Public Safety Assessment for Pierce County, Washington.* RTI International. https://advancingpretrial.org/psa/research/

DeMichele, M., Tueller, S., Inkpen, C., Dawes, D., & Lattimore, P. (2023). *Validation and predictive bias testing of the Public Safety Assessment for Fulton County, Georgia*. RTI International.

DeMichele, M., Tueller, S., Janda, K., Dawes, D., & Lattimore, P. (2023). *Validation and predictive bias testing of the Public Safety Assessment for Thurston County, Washington*. RTI International.

Desmarais, S. L., Zottola, S. A., Duhart Clarke, S. E., & Lowder, E. M. (2021). Predictive validity of pretrial risk assessments: A systematic review of the literature. *Criminal Justice and Behavior*, 48(4), 398–420.

Desmarais, S. L., D'Amora, D. A., & Tavárez, L. P. (2022). *Advancing fairness and transparency: National guidelines for post-conviction risk and needs assessment*. U.S. Department of Justice, Bureau of Justice Assistance, and Council of State Governments Justice Center.

Ferguson, E., De La Cerda, H., & Guerin, P. (2021). *Bernalillo County Public Safety Assessment review*. University of New Mexico, Institute for Social Research. https://isr.unm.edu/reports/2021/bernalillo-county-public-safety-assessment-validation-study.pdf

Greiner, D. J., Stubenberg, M., & Halen, R. (2020). *Validation of the PSA in Harris County, TX.* Access to Justice Lab, Harvard Law School. https://a2jlab.org/public-safety-assessment-validation-studies/

Greiner, D. J., Stubenberg, M., & Halen, R. (2021). *Validation of the PSA in Kane County, IL.* Access to Justice Lab, Harvard Law School. https://a2jlab.org/public-safety-assessment-validation-studies/

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.

Hess, J., & Turner, S. (2021). *Validation of the PSA in Los Angeles County.* Center for Evidence-Based Corrections, University of California, Irvine.

Lowenkamp, C. T., & VanNostrand, M. (2013). *Exploring the impact of supervision on pretrial outcomes*. Laura and John Arnold Foundation.

Lowenkamp, C. T., DeMichele, M., & Warren, L. K. (2020). Replication and extension of the Lucas County PSA project. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727443

Mamalian, C. A. (2011). *State of the science of pretrial risk assessment*. Bureau of Justice Assistance, U.S. Department of Justice.

McCoy, C. (2007). Caleb was right: Pretrial decisions determine mostly everything. *Berkeley Journal of Criminal Law*, *12*(2), 135-150.

Pugmire, T. (2023). *Missoula County Public Safety Assessment validation analysis.* Missoula County Community Justice Department. University of Montana, Department of Sociology. https://www.missoulacounty.gov/media/tgldqleo/public-safety-assess-summary-formal-report_2023.pdf

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*(5), 615-620.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*(4), 680-712.

Skog, A., & Lacoe, J. (2021). *Validation of the PSA in San Francisco.* California Policy Lab. https://www.capolicylab.org/wp-content/uploads/2021/08/Validation-of-the-PSA-in-San-Francisco.pdf

St. Louis, S. (2023). The pretrial detention penalty: A systematic review and meta-analysis of pretrial detention and case outcomes. *Justice Quarterly*, *41*(3), 347-370. https://doi.org/10.1080/07418825.2023.2193624

VanNostrand, M., & Lowenkamp, C. T. (2013). *Assessing pretrial risk without a defendant interview*. Laura and John Arnold Foundation.