When "Good" Isn't: Officer-Level Variability in Risk Assessment Instruments

Christopher T. Lowenkamp
University of Missouri - Kansas City & University of Cincinnati Center for Justice & Communities
Anthony W. Flores
California State University - Bakersfield
Alexander M. Holsinger
University of Missouri - Kansas City

RISK ASSESSMENT HAS become ubiquitous in criminal justice decision-making. These instruments inform pretrial detention decisions, guide sentencing recommendations, determine supervision intensity for probation and parole, allocate treatment resources, and influence responses to technical violations (Andrews & Bonta, 2010; Monahan & Skeem, 2016). Given their widespread implementation and the consequential nature of these decisions, understanding the actual predictive validity of risk assessment instruments has become increasingly important for both research and practice.

Contemporary research on risk assessment predictive validity reveals consistent patterns. Fazel et al.'s (2022) systematic review of 27 independent validation studies found that commonly used risk assessment tools achieve Area Under the Curve (AUC) values ranging from .57 to .75, with most clustering between .60 and .65. Viljoen et al.'s (2025) meta-analysis provides additional context through direct comparison studies, finding that structured risk assessment tools achieve average AUCs of .689-.721 across different types of offending, while unstructured judgment achieves AUCs of .618-.625. While these findings demonstrate that structured tools statistically outperform unstructured judgment, the magnitude of improvement remains modest, particularly when considering the substantial resources required for training, implementation, and quality assurance of structured assessments.

Concurrent with these empirical findings, the field has witnessed an evolution in how predictive validity is characterized and evaluated. Rice and Harris (2005) originally provided descriptive benchmarks based on Cohen's (1988) effect size framework, classifying AUCs of .556, .639, and .714 as representing "small," "medium," and "large" effects respectively. These descriptive terms were subsequently reframed by Desmarais and Singh (2013) using evaluative labels, with AUCs of .55-.63 characterized as "fair," .64-.70 as "good," and above .71 as "excellent."

This seemingly minor shift from descriptive to evaluative terminology has potentially influenced how the field interprets and accepts validation results. Once AUCs around .60 were relabeled from "small effects" to "fair" or even "good" performance, it became easier to justify their acceptability. This linguistic drift has now culminated in the creation of circular "industry standards" that use common practice rather than scientific benchmarks to define adequacy. A report by D'Amato et al. (2023) exemplifies this circularity, explicitly defining an "industry standard" as AUC values between .55 and .70, based solely on the reasoning that "the majority of validation studies examining criminogenic risk and need assessments have AUC values that fall between .55 and .70" (p. 32). Similarly, the Pennsylvania ORAS validation (Vick et al., 2024) demonstrates the tension this has created. While acknowledging that AUC values for both instruments fell in the "fair" range (.57 to .63), the authors nonetheless characterized their overall findings as demonstrating "strong predictive validity," relying on logistic regression odds ratios and recidivism rate separation to support that characterization.

This acceptance of modest predictive validity as the professional standard raises critical questions. When aggregate validation statistics hover around .60, are we observing consistent mediocre performance, or might these aggregates mask important variation at the officer or some other smaller aggregate level? More fundamentally, if most officers achieve similar modest validity, what accounts for the consistent failure to achieve stronger prediction?

These questions have remained largely unexamined because validation studies traditionally report only aggregate statistics, treating implementation as uniform across assessors. This assumption of uniformity, that a validated instrument will perform similarly regardless of who administers it, underlies the entire enterprise of standardized risk assessment. Yet if this assumption is wrong, if substantial variation exists at the officer level, then aggregate validation statistics may obscure a more complex reality: some officers achieve strong predictive validity while others do not. Such variation would suggest

that the problem lies not in the instruments themselves but in their implementation, fundamentally changing how we understand and address the field's predictive validity ceiling. Moreover, if officer-level variation is substantial and unexplained by obvious factors like experience or caseload characteristics, it would call into question whether "evidence-based" risk assessment is truly evidence-based in practice. The drift from Rice and Harris's "small effects" to contemporary labels of "fair," "good," or "industry standard" has prevented us from even asking these critical questions. Why would we question performance the field has already deemed acceptable?

The present study examines officer-level predictive validity to address these questions, hypothesizing that aggregate validation statistics might obscure substantial between-officer variability. If such variation exists, it would suggest that the problem lies not in the instruments themselves but in their implementation, fundamentally changing how we understand the field's predictive validity ceiling. Such a finding would reframe the central question from "Why can't instruments exceed AUC = .70?" to "Why do some officers achieve strong predictive validity while others fail to do so using the same validated tools?"

Method

The present study employed a retrospective analysis of officer-level predictive validity across multiple risk assessment instruments and jurisdictions. Unlike traditional validation studies that aggregate results across all assessors, this investigation calculated separate predictive validity metrics for each individual officer, allowing for examination of the distribution and range of performance within and across instruments. This approach reveals whether aggregate statistics mask important variation in implementation quality and whether some officers consistently achieve better predictive validity than others using the same instruments. The analysis employed two measures of predictive validity for each officer: Area Under the Curve (AUC) to assess discrimination between recidivists and non-recidivists, and correlation coefficients to measure the linear relationship between risk scores and recidivism outcomes.

Data Sources and Sample

Data were obtained from two jurisdictions implementing post-conviction risk assessment instruments during routine supervision operations. The first jurisdiction assessed individuals during calendar years 2017 and 2018, with recidivism measured as reconviction within a fixed three-year follow-up period. The second jurisdiction conducted assessments between 2018 and 2022, with recidivism operationalized as revocation filed for either new convictions or technical violations during the supervision period.

While follow-up periods varied between jurisdictions, fixed three-year periods in the first jurisdiction and variable supervision periods in the second, within each jurisdiction, all officers used the same outcome measures and timing protocols.[1] Assessments were conducted by probation and parole officers as part of standard intake and case management procedures.

Given the sensitive nature of these analyses and findings, to protect jurisdiction confidentiality, specific instruments and locations are not identified. The instruments are designated as Instrument A (with Scale 1 and Scale 2) and Instrument B. Both Instrument A and Instrument B represent actuarial risk assessment tools with established scoring protocols and documented aggregate predictive validity in the range typically reported in meta-analyses.

The analytic sample was restricted to officers who completed 20 or more assessments to ensure stable estimates of individual predictive validity. This threshold balanced statistical reliability with sample maximization. The final sample comprised 245 officer-instrument combinations. Some officers appear two times in this count if they used different instruments during the study period, as the unit of analysis is the officer-instrument combination rather than unique officers.

Measures

Two measures of predictive validity were calculated for each officer. The Area Under the Curve (AUC) measured each officer's ability to discriminate between individuals who did and did not recidivate, with values ranging from 0 to 1 where .50 indicates chance-level prediction. The point-biserial correlation coefficient (r) measured the relationship between risk scores and the binary recidivism outcome.

To examine whether caseload composition explained officer-level variability, several characteristics were calculated for each officer's assessed cases: percentage of males, percentage of Black individuals, average age at assessment, and average risk score. Officer experience with the assessment was operationalized as the number of assessments conducted.

Analytic Strategy

Analyses proceeded in three stages. First, descriptive statistics were calculated for officer-level AUC and correlation values by instrument. Second, regression models examined whether caseload characteristics and experience predicted officer-level predictive validity. Third, instrument controls were added to determine whether between-instrument differences explained officer variability. Sensitivity analyses examined varying thresholds for officer inclusion and weighted regression models.[2]

Results

Officer-Level Predictive Validity

Table 1 presents the distribution of officer-instrument combinations across the instruments. Instrument A accounted for the majority of combinations (71.43 percent), with Scale 1 and Scale 2 representing 40.41 percent and 31.02 percent respectively. Instrument B comprised 28.57 percent of the sample.

These 245 officer-instrument combinations represent standardized assessments expected to produce consistent predictive

[1] In the jurisdiction with variable follow-up periods, among officers meeting our inclusion criterion of 20+ assessments, 96 percent of officers using Instrument A and 100 percent of officers using Instrument B had assessed cases with follow-up periods exceeding 210 days (seven months). Flores, Holsinger, Lowenkamp, and Cohen (2017) demonstrated that AUC values reach stability at the 7-month mark, suggesting these variations are unlikely to substantially affect predictive validity estimates.

[2] Sensitivity analyses included: (a) varying the inclusion threshold from 10 to 30 assessments yielded similar patterns of unexplained variance, with R^2 ranging from .031 to .074 for AUC models across thresholds; (b) restricting to officers with 100+ assessments (n = 67, 16,608 total assessments) showed persistent variation with AUC ranging from .413 to .853 (M = .645, SD = .081), with 96.1 percent of AUC variance still unexplained by caseload characteristics and experience; (c) weighted regression using number of assessments as weights yielded similar results to unweighted models.

TABLE 1.
Distribution of Officers by Risk Assessment Instrument (N = 245)

Instrument	Scale	n	%
Instrument A	Scale 1	99	40.41
Instrument A	Scale 2	76	31.02
Instrument B	—	70	28.57
Total		245	100.00

validity. However, examining the actual performance of individual officers reveals a different picture. Table 2 reveals substantial variability in officer-level predictive validity both within and across instruments. Of the 245 officer-instrument combinations meeting inclusion criteria, 230 (93.9 percent) had sufficient outcome variation to calculate predictive validity metrics. The 15 combinations (6.1 percent) with incalculable AUCs represent officers whose caseloads resulted in uniform outcomes, either all successes or all failures, during the study period.

Among the 230 analyzable cases, AUC values ranged from .150 to 1.00 (M = .651, SD = .121), while correlation coefficients ranged from r = -.221 to .716 (M = .204, SD = .175). The most extreme variation occurred within Instrument A Scale 1, where AUCs ranged from .150 to .904, a span of .754. Instrument B showed the widest range in correlation coefficients (r = -.221 to .676). These negative correlations indicate that some officers' assessments were inversely related to outcomes; that is, higher risk scores predicted lower recidivism.

The presence of perfect discrimination (AUC = 1.00) alongside below-chance discrimination (AUC = .150) within the same instrument possibly suggests fundamental differences in how officers apply standardized tools.

The substantial variation persists even among high-volume officers. Restricting the analysis to officers who completed 100 or more assessments (n = 67, representing 16,608 total assessments), AUC values ranged from .413 to .853 (M = .645, SD = .081). Correlation coefficients ranged from r = -.081 to .438 (M = .179, SD = .115). Among these high-volume officers, the number of assessments completed ranged from 103 to 1,770 assessments (M = 247.9, SD = 320.9), with the upper quartile completing 202 or more assessments each. Even with sample sizes that should produce highly stable AUC estimates, officer performance varied from below chance to well above Rice and Harris's large effect benchmark (.714), eliminating statistical instability as an explanation for the observed variation.

Figure 1 presents the distribution of officer-level AUC values across the 230 officer-instrument combinations with calculable predictive validity. The histogram reveals the full distribution of officer-level performance masked by aggregate validation statistics. While the distribution centers near .65, consistent with typical validation studies' findings, the range extends from .15 to 1.00, encompassing both worse-than-chance performance and perfect discrimination.

The vertical reference lines indicate critical benchmarks: chance performance (.50),

TABLE 2.
Officer-Level Predictive Validity by Risk Assessment Instrument

			AUC				r			
Instrument	Scale	n	M (SD)	Min	Max	Range	M (SD)	Min	Max	Range
Instrument A	Scale 1	85	.620 (.125)	.150	.904	.754	.096 (.106)	-.131	.377	.508
Instrument A	Scale 2	75	.653 (.109)	.410	1.000	.590	.221 (.150)	-.089	.716	.805
Instrument B	—	70	.688 (.119)	.307	.899	.592	.316 (.192)	-.221	.676	.897
Overall		230	.651 (.121)	.150	1.000	.850	.204 (.175)	-.221	.716	.937

Note: Analysis includes 230 of 245 officer-instrument combinations (93.9 percent) with sufficient outcome variation for AUC calculation.

TABLE 3.
Distribution of Officers by Rice & Harris (2005) Effect Size Benchmarks[3]

Effect Size Category	AUC Range	n	% of Valid	% of Total
Worse than chance	< .50	18	7.83	7.35
Negligible	.50-.555	24	10.43	9.8
Small	.556-.639	64	27.83	26.12
Medium	.639-.714	61	26.52	24.9
Large	> .714	63	27.39	25.71
Subtotal (Valid)		230	100	93.88
Missing/Incalculable	—	15	—	6.12
Total		245		100

[3] Distribution of performance varied significantly by instrument ($\chi^2 = 20.23$, p = .009). Among officers with calculable AUCs, the percentage performing worse than chance ranged from 5.3 percent (Instrument A Scale 2) to 9.1 percent (Instrument A Scale 1), while those achieving large effects ranged from 15.2 percent (Instrument A Scale 1) to 45.7 percent (Instrument B). Notably, Instrument A Scale 1 showed the highest rate of incalculable AUCs (14.1 percent vs. 1.3 percent for Scale 2 and 0 percent for Instrument B), suggesting more officers with uniform outcomes. Despite Instrument B showing the highest proportion of large effects, all three instruments exhibited substantial within-instrument variability, with 14.3 percent of Instrument B users still performing below Rice & Harris's small effect threshold. The key finding is not which instrument performs marginally better, but that all three show concerning levels of implementation failure regardless of their design or complexity.

Rice and Harris's (2005) small effect threshold (.556), medium effect (.639), and large effect (.714). A substantial proportion of officers (18.26 percent) fall below the .556 threshold, indicating performance that barely exceeds or actually falls below chance prediction. The concentration of officers between .50 and .70 represents those performing at levels the field considers "acceptable" but which provide minimal improvement over random classification or classification by unaided practitioners (see Viljoen, et al., 2025).

Classification by Rice and Harris (2005) Benchmarks

Table 3 distributes officer performance according to Rice and Harris's (2005) original effect size benchmarks, supplemented with categories for performance below the small effect threshold. Eighteen officer-instrument combinations (7.35 percent) achieved AUCs below .50, performing worse than random chance. An additional 24 (9.80 percent) performed between .50 and .556, barely exceeding chance levels. Together, these 42 officer-assessment combinations, representing 17.14 percent of the sample (N = 245 including those with incalculable AUCs), conducted assessments that Rice and Harris would not have considered to demonstrate even a "small" effect.

The distribution reveals a bimodal split: nearly equal proportions of officers achieve "Large" effect sizes (27.39 percent) and fall into the combined worse-than-chance and negligible categories (18.26 percent), suggesting not a normal distribution around mediocrity but rather a split between competent assessors and those whose assessments fail to exceed chance levels of discrimination.

More broadly, 46.09 percent of officers (106 of 230) perform below Rice & Harris's .639 medium effect threshold; that is, nearly half cannot achieve what was considered merely moderate performance in 2005. The modal category is "Small" effect (27.83 percent), clustering just above the .556 threshold that barely exceeds chance. This concentration at marginal performance levels indicates that the most common outcome for officers using standardized instruments is minimal predictive validity. This poor performance becomes even more concerning when examining the convergence between different validity metrics.

Figure 2 presents the relationship between two commonly reported measures of predictive validity: Area Under the Curve (AUC) and the point-biserial correlation (r) for each officer-instrument combination. This scatter plot reveals not only the extent of officer-level variability but also the methodological complexity inherent in evaluating risk assessment performance.

The strong positive relationship between AUC and correlation values (r = .846) indicates these metrics share approximately 72 percent of their variance. However, this correlation drops to .738 for officers with more than 100 assessments (n=67), suggesting that some of the observed relationship, and particularly the anomalous patterns, may reflect statistical instability in smaller samples.

The AUC metric remains relatively unaffected by base rates while correlation coefficients are sensitive to them. With very low base rates (e.g., 10 percent recidivism),

FIGURE 1.
Distribution of Officer-Level Predictive Validity (n = 230)

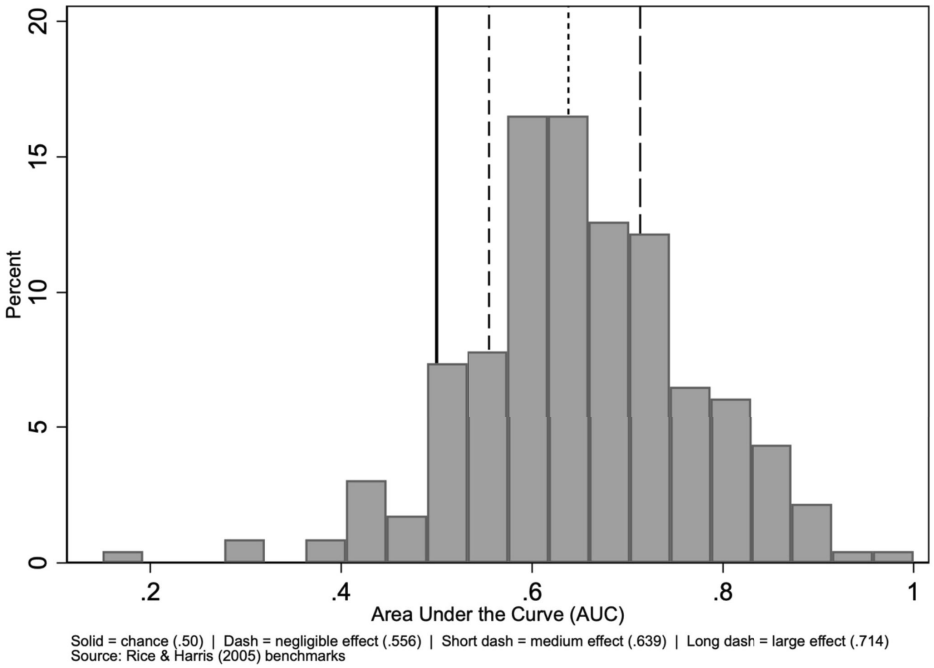
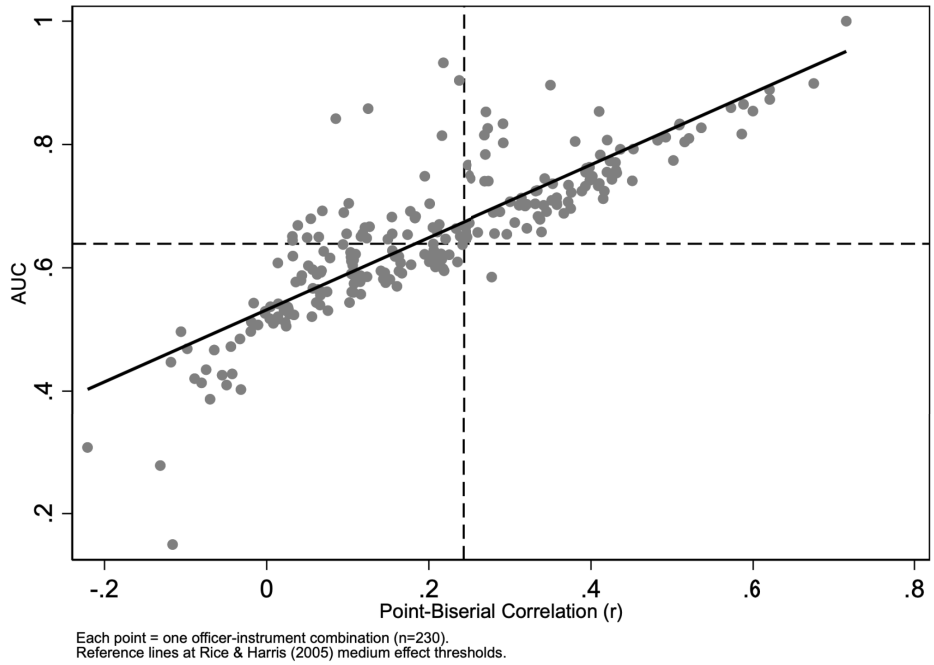


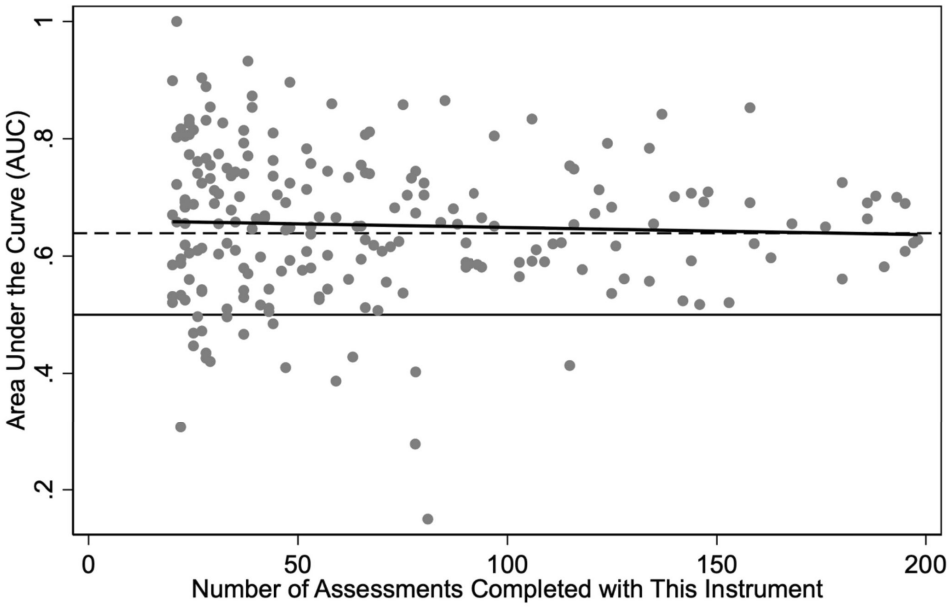
FIGURE 2.
Officer Performance: AUC vs. Correlation (each point represents one officer-instrument combination; n = 230)



the restricted variance in the outcome limits the maximum achievable correlation, while AUC can still reach high values if discrimination is good. This base rate dependency likely explains some of the scatter in Figure 2. Officers with similar discrimination ability (AUC) may show vastly different correlations depending on their caseload's base rate.

Reference lines at Rice and Harris's (2005) medium effect size benchmarks (AUC = .639, r = .243) divide the plot into four quadrants. A substantial proportion of officers fall in the lower left quadrant, failing to achieve medium effect sizes by either metric. The anomalous cases, officers with AUCs above .50 but negative correlations, may partly reflect the combination of small sample sizes (most have n = 20-30 assessments) and extreme base rates. For example, an officer with only 20 assessments and a 15 percent base rate would have just 3 expected recidivists. In a sample that small, statistical noise could easily produce a negative correlation or push AUC below chance.

FIGURE 3.
Officer Performance Versus Experience with Instrument



Regression: β = -0.0001, p = 0.493, R^2 = 0.002
Solid line = chance (.50), Dashed line = medium effect (.639)
Note: Truncated at 200 assessments for clarity

TABLE 4.
Regression Models Predicting Officer-Level Predictive Validity[4]

	Model 1: AUC		Model 2: Correlation	
Predictors	b	p	b	p
% Male	-0.0004	0.605	0.0001	0.944
% Black	-0.0015	0.03	-0.0021	0.020
Average Age	0.0022	0.48	0.0049	0.217
Average Score	0.0015	0.226	0.0085	< 0.001
Experience (# of assessments)	9.12e-07	0.982	5.17e-6	0.923
Constant	0.608	< 0.001	-0.063	0.739
R^2	0.054		0.242	
Adjusted R^2	0.033		0.225	
N	230		230	

[4] Sensitivity analyses using weighted regression (weighted by number of assessments) with instrument controls produced R^2 = .092 for AUC and R^2 = .406 for correlations (adjusted R^2 = .067 for AUC and adjusted R^2 = .390 for correlations). While weighting improved the correlation model, even these weighted models leave 90.8 percent of AUC variance and 59 percent of correlation variance unexplained.

Predictors of Officer-Level Prediction

Table 4 presents regression models examining whether officer caseload characteristics and experience with the assessment predict individual predictive validity. Model 1 examines predictors of AUC values, while Model 2 examines predictors of correlation coefficients. Both models explained minimal variance in officer-level predictive validity. For AUC values (Model 1), only 5.4 percent of the variance was explained (R^2 = .054). Most notably, officer experience showed virtually no relationship with predictive validity (b = 9.12 × 10^{-7}, p = .982), indicating that officers conducting 200 assessments achieved no better predictive validity than those conducting 20.

The correlation model (Model 2) explained more variance than the AUC model (24.2 percent vs. 5.4 percent), though over three-quarters of the variance in correlation values still remained unexplained. Average risk score emerged as a significant predictor of correlations (b = .0085, p < .001) but not AUC (b = .0015, p = .226), suggesting that officers working with higher risk caseloads achieved stronger linear relationships between scores and outcomes, possibly due to greater variance in outcomes. However, this relationship did not translate into better discrimination ability. The percentage of Black individuals in an officer's caseload showed a small negative relationship with both AUC (b = -.0015, p = .030) and correlation values (b = -.0021, p = .020), though the effect sizes are minimal.

What these models fail to explain is equally important. Even with caseload characteristics and experience with the assessment included, 94.6 percent of variance in AUC values and 75.8 percent of variance in correlations remained unexplained. The failure of experience with the assessment to predict performance is of particular interest, suggesting that implementation problems are not resolved through practice or repetition. This finding supplements Flores et al. (2006), who found that training and use of assessment results in case planning predicted accuracy. The present results suggest that without such structured support, simply conducting more assessments does not improve predictive validity. Figure 3 plots officer AUC against the

number of assessments completed with the instrument. The near-zero slope (β = -.0001, R^2 = .002) confirms that volume of experience with the assessment does not predict predictive validity.

This null relationship has implications for quality assurance. If experience does not improve performance, then officers who begin with poor assessment practices will continue performing poorly indefinitely. Traditional assumptions about training effects, supervision benefits, or gradual skill development through practice find no support in these data. The field's reliance on initial training followed by minimal ongoing oversight appears inadequate when predictive validity shows no improvement across thousands of cumulative assessments.

Instrument Controls

Table 5 adds instrument controls to examine whether between-instrument differences explain officer-level variability. While significant differences existed between instruments, with Instrument B performing best, followed by Instrument A Scale 2, then Scale 1, the models still failed to explain the vast majority of variation. Even with instrument controls, 91 percent of variance in AUC values and 71.2 percent of variance in correlations remained unexplained. These findings eliminate the most obvious explanations for officer-level variation. Neither caseload composition, experience with the assessment, or instrument type adequately accounts for why some officers achieve AUCs above .90 while others perform worse than chance.

Several notable patterns emerge with instrument controls. First, the substantial differences between instruments (with Instrument B showing markedly better performance than Instrument A Scale 1) might suggest that some tools produce better results than others. However, as shown in Table 2, even Instrument B included officers with AUCs as low as .307 and correlations of -.221, worse than random guessing. The between-instrument differences in average performance mask the more critical within-instrument variation.

Second, average risk score becomes significantly negative for both AUC (b = -.0101, p = .010) and correlation (b = -.0113, p = .025) when controlling for instrument, suggesting that within instruments, officers handling higher risk caseloads actually achieve worse discrimination. This counterintuitive finding likely reflects restriction of range effects: Homogeneous high-risk caseloads compress score distributions and create ceiling effects, reducing the instrument's discriminative capacity. This psychometric artifact could explain why officers managing higher risk populations achieve lower AUCs despite showing stronger correlations (due to greater outcome variance).

Finally, the demographic effects observed in Table 4 disappear when controlling for instrument, indicating these were likely artifacts of different instruments being used in different jurisdictions with varying demographic compositions rather than true caseload effects. The persistence of sizeable unexplained variance even after controlling for the most plausible explanatory factors, instrument type, caseload characteristics, and experience with the assessment, underscores that the problem in risk assessment is likely rooted in implementation quality, not instrument design or caseload composition.

Discussion

Our findings reveal two interconnected issues that have been obscured by aggregate validation statistics and lowered standards. The first is empirical: considerable unexplained variation in officer-level performance, with some officers achieving near-perfect discrimination while others perform worse than chance. The second is conceptual: the field's gradual acceptance of mediocre performance as adequate, facilitated by drift from descriptive to evaluative terminology.

When 8 percent of officers perform no better than random chance, and when neither experience with the assessment nor caseload characteristics explain this variation, fundamental questions arise about the nature of risk assessment implementation. The presence of negative correlations indicates that some officers systematically assign higher risk scores to those who are less likely to recidivate, a pattern that, if relied upon for decision-making, would direct resources away from those who need them most.

The implications extend beyond individual assessment failures. Officers performing worse than chance may continue using these instruments for years, producing consequential decisions based on assessments with no predictive validity. This is particularly concerning given evidence that inappropriate interventions based on inaccurate risk classifications can increase recidivism (Lowenkamp & Latessa, 2004). The field currently has no systematic mechanism to detect or correct these failures at the officer level.

The findings from Viljoen et al. (2025) provide useful context for interpreting our results. Their meta-analysis found that unstructured judgment typically achieves AUCs between .618 and .625 in predicting recidivism, a range similar to the aggregate performance of many structured risk assessments. If structured assessments, which require substantial resources for training, implementation, and quality assurance, achieve predictive validity only marginally better than unstructured judgment, the cost-benefit ratio warrants examination. This consideration becomes more pressing at the officer level, with 46 percent of officers failing to achieve even a medium effect size (AUC < .639), a threshold only marginally above the upper limit of unstructured practitioner judgment documented by Viljoen et al. (2025).

The inability of our models to explain officer-level variation eliminates the most

TABLE 5.

Regression Models with Instrument Controls

	AUC (n=230)		Correlation (n=230)	
Predictors	**b**	**p**	**b**	**p**
% Male	-.0004	.534	-.0001	.944
% Black	-.0004	.653	-.0002	.842
Average Score	-.0101	.010	-.0113	.025
Experience	8.32e-6	.838	1.81e-5	.729
Instrument				
A Scale 2 (ref: A Scale 1)	.153	.003	.261	.000
B (ref: A Scale 1)	.265	.003	.443	.000
Constant	.702	.000	.149	.044
R^2	.090		.288	
Adjusted R^2	.065		.269	

obvious explanations. It is not simply that some officers assess more difficult caseloads. Demographic composition and risk levels explained less than 5 percent of variance in performance. It is not that officers improve with experience; those conducting their 200th assessment performed no differently than those conducting their 20th. Even instrument differences, while statistically significant, left over 90 percent of variance in AUC values unexplained.

This unexplained variance points toward factors our data could not capture: assessment fidelity, clinical override practices, or fundamental differences in how officers approach the assessment process. Some officers may treat assessments as mechanical exercises, while others engage more thoughtfully with the instruments. Some may systematically override actuarial scores based on clinical judgment, judgment that might not improve upon, and could undermine, actuarial prediction.

The substantial officer-level variation documented in this study raises important questions about implementation quality in risk assessment practice. Previous research has documented widespread implementation challenges in correctional settings (Taxman et al., 2007), and studies have consistently shown that training quality (Teo et al., 2012) and implementation fidelity (Vincent et al., 2012) are critical factors often overlooked in validation studies. Our findings suggest that these implementation factors may have far greater impact on real-world performance than previously recognized.

Our findings reveal a paradox in risk assessment: Instruments can be simultaneously validated and invalid. The same tool achieving acceptable aggregate validity in research studies can produce assessments ranging from near-perfect to worse-than-chance in practice. This paradox exists because validation studies treat implementation as a constant when it is actually a variable and possibly a consequential one. An instrument is only as valid as its implementation, yet the field has focused almost exclusively on instrument development while neglecting implementation quality.

The conceptual problem, the acceptance of mediocre performance as adequate, may have enabled the empirical problem to persist undetected. When validation studies report only aggregate statistics, when AUCs of .60 are labeled "good," and when "industry standards" are defined by common practice rather than decision-making requirements, there is little impetus to examine implementation quality. The evolution from Rice and Harris's descriptive benchmarks to contemporary evaluative labels has not merely changed terminology; it has lowered expectations and obscured failure. The D'Amato et al. (2023) report exemplifies this problem: Despite claiming an "industry standard" of .55-.70, their own validation found no IRAS tool achieved an overall AUC above .60, with values clustering near the floor of their already-lowered standard.

Instrument or Implementation?

Our findings also suggest that substantial officer-level variation exists, and the problem may lie in implementation rather than the instruments themselves. Officers achieving AUCs above .80 using these same tools as officers achieving AUCs below .60 (or lower) demonstrate their potential capability when properly implemented. The problem appears to be the absence of quality assurance systems ensuring consistent implementation.

The field may have inadvertently concealed this problem by redefining modest predictive validity as acceptable. When aggregate AUCs of .60 became "fair" or met "industry standard," we stopped asking why instruments weren't performing better. This reframing may have prevented us from investigating whether the predictive validity ceiling reflected instrument limitations or implementation failures. Our findings suggest it may be the latter, aggregate mediocrity perhaps represents implementation failure, not uniform modest performance.

Viljoen et al.'s (2025) finding that unstructured judgment achieves AUCs of .618-.625 provides useful context. When implemented poorly, structured assessment may offer little advantage over clinical judgment while consuming substantially more resources. However, 27 percent of officers achieve large effects (AUC > .714), demonstrating performance substantially exceeding unstructured judgment. The question is not whether structured assessment can outperform clinical judgment, data suggest it can, but whether we have systems ensuring it does so consistently.

Abandoning structured risk assessment would be a dangerous misinterpretation of these observational findings. While 46 percent of officers fail to achieve medium effect sizes, 54 percent exceed that threshold. The solution may lie in understanding what enables effectiveness in high-performing officers and systematically implementing those practices. The existence of officers achieving AUCs above .80 suggests the instruments can work; the existence of officers performing worse than chance suggests our oversight systems have not detected these failures.

Implications for Practice

These findings suggest changes in how risk assessments are implemented and monitored are required. The distribution of officer performance raises critical questions about minimum standards for risk assessment practice. Using Rice and Harris's (2005) original benchmarks, 46 percent of officers perform below the threshold for a medium effect (AUC < .639), suggesting nearly half would benefit from additional training, supervision, or support. More broadly, the field should reconsider whether the cited "industry standard" of .55-.70 represents adequate performance, particularly given that many officers within this range demonstrate predictive validity that barely exceeds chance levels.

Agencies should move beyond aggregate validation to continuous monitoring of individual officer performance. Our data reveal 23 officers with negative correlations who have collectively conducted 1,099 assessments, representing over a thousand decisions potentially informed by invalid assessments. Given these findings, agencies should implement continuous performance monitoring and targeted remediation. Officers performing below acceptable thresholds should be identified for immediate intervention.

We envision a tiered training model where officer-level AUC metrics drive targeted interventions (see Appendix for a more detailed discussion). Officers producing AUCs at or below small effect size levels (≤ .556) would receive immediate comprehensive retraining covering the entire assessment process and administration protocols. Those performing in the small effect range (.556-.639) would receive targeted training focused on specific skill deficits in scoring or data collection, while officers in the moderate range (.639-.714) would participate in skill enhancement programs. Officers already achieving large effect sizes but at the lower end (close to .714) would receive advanced training designed to push their performance to the limits of what has been achieved in actuarial prediction in correctional contexts. While we don't detail specific curricula for each track, officers producing assessments with less than small effect size prediction would clearly benefit from foundational retraining in assessment fundamentals, proper administration

techniques, and ongoing monitoring. The content for higher-tier tracks would require additional research comparing the practices and approaches of high-performing officers (AUC ≥ .80) with those of lower-performing officers to identify the specific competencies and techniques that distinguish officers achieving larger effect sizes from those achieving smaller effect sizes.

Given that experience with the assessment does not seem to improve performance, initial training and ongoing quality assurance become critical. The null relationship between the number of assessments conducted and performance suggests that poor practices established early persist indefinitely without intervention. Quality assurance would benefit from shifting from assuming improvement over time to active monitoring and correction of poor performance.

Implications for Research

Future validation studies, when possible, should expand reporting beyond aggregate statistics. Full distributions, ranges, and examinations of outliers provide more complete pictures of how risk assessment tools perform in practice. The wide variation in officer performance documented here suggests critical research questions about implementation factors that current validation approaches cannot address.

Research comparing the practices and approaches of high-performing officers (AUC ≥ .80) with those of lower performing officers could identify specific competencies and techniques that distinguish effective implementation. This might include examining assessment fidelity, decision-making processes, override patterns, and scoring consistency. Understanding what enables some officers to achieve near-perfect discrimination while others fail to achieve acceptable discrimination could inform targeted interventions.

The development of implementation fidelity measures that can be assessed independently of outcomes represents a critical research priority. Such measures could help identify quality implementation practices before predictive validity deteriorates. Currently, the field lacks tools to distinguish between officers likely to achieve strong versus poor predictive validity until after numerous assessments have been conducted, and potentially misguided decisions made.

Given that experience with the assessment does not seem to improve predictive validity, experimental studies testing different training, supervision, and quality assurance approaches are essential. Research should examine whether intensive initial training, regular booster sessions, peer review processes, or automated scoring systems can reduce the variance in officer performance we documented.

Implications for Policy

The substantial variation in officer-level predictive validity raises fundamental questions for policies mandating or encouraging risk assessment use. When the same validated instrument produces dramatically different results depending on who administers it, the reliability of individual assessments in high-stakes decisions becomes questionable. Courts and policymakers must consider not just whether an instrument has been validated in aggregate, but whether adequate systems exist to ensure consistent implementation quality.

Mandating risk assessment use without corresponding infrastructure for quality monitoring may not achieve intended policy goals. Our findings suggest that validated instruments are only as effective as their implementation, and ensuring implementation quality requires resources and oversight beyond initial tool adoption. Policy frameworks should account for assessor variability and establish mechanisms for ongoing quality assurance. Without such frameworks, risk assessment may not consistently provide the improvements over unstructured judgment that validation studies suggest are possible.

Limitations

Several limitations qualify the interpretation of our findings. We lacked comprehensive measures at multiple levels that could explain officer-level variation. At the caseload level, we controlled for demographic composition and average risk scores but could not examine offense type specialization; officers assessing populations such as sex offenders, where general instruments may perform differently, could show systematically different validity. At the officer level, the number of assessments completed serves as an imperfect proxy for experience. We lacked data on years in corrections, educational background, assessment-specific training and certification, and participation in ongoing professional development. At the agency level, we could not examine quality assurance processes, availability of coaching related to assessments, or recertification protocols. Agencies with robust oversight might show less officer-level variation than those with minimal infrastructure.

Methodological constraints also limit our conclusions. The variation in follow-up periods between jurisdictions, variable supervision periods for Instrument A versus fixed three-year periods for Instrument B, could influence predictive validity estimates. However, among officers meeting our inclusion criterion, approximately 95 percent using Instrument A Scale 1 and 100 percent using Instrument A Scale 2 had average follow-up periods exceeding seven months. Given that Flores et al. (2017) demonstrated AUC stability after seven months, these variations are unlikely to substantially affect our findings. The threshold of 20-plus assessments, while necessary for statistical stability, excluded newer officers who might show different patterns during initial implementation.

Finally, our examination focused exclusively on post-conviction risk assessment instruments. Pretrial assessments, which operate under different time constraints, with different information availability, and predict different outcomes (failure to appear and dangerousness) might show different patterns of officer-level variation. Similarly, specialized assessments (violence risk, sex offense risk, intimate partner violence) versus general recidivism tools might demonstrate different implementation challenges. The anonymization of instruments and jurisdictions, while necessary given the sensitive nature of findings, prevents instrument-specific recommendations or targeted remediation efforts.

Despite these limitations, the magnitude of unexplained variance, over 90 percent for AUC values even with available controls, suggests that unmeasured factors, while important, are unlikely to fully explain the dramatic variation from AUC = .15 to 1.00. The consistency of findings across three instruments from two jurisdictions, combined with the null relationship with experience, points toward fundamental implementation challenges that may transcend specific contexts or populations.

Conclusion

The acceptance of modest aggregate predictive validity as "industry standard" has obscured a fundamental problem with implementation in risk assessment practice. When officers using the same validated instrument can range from perfect discrimination to systematic mis-prediction, and when this variation cannot be explained by caseload characteristics,

the assessment instrument, or experience with the assessment, fundamental questions arise about the nature of risk assessment implementation.

The legal and ethical foundations of using risk assessment in criminal justice decision-making rest on the assumption that these tools improve accuracy. Our findings suggest this assumption may not hold when implementation quality goes unmonitored. Moving forward, the field needs continuous performance monitoring, evidence-based training interventions, and the willingness to acknowledge that a validated instrument in the hands of a poorly performing officer is not valid at all. Without such reforms, risk assessment will continue to fall short of its potential contribution to both public safety and individual liberty.

References

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Anderson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

D'Amato, C., Barnes, J. C., Khade, N., & Smith, M. S. (2023). Indiana Risk Assessment System and Youth Assessment System: A re-validation study: Final report. University of Cincinnati Corrections Institute.

Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Council of State Governments Justice Center.

Fazel, S., Burghart, M., Fanshawe, T., Gil, S., Monahan, J., & Yu, R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *Journal of Criminal Justice, 81,* 1-9. https://doi.org/10.1016/j.jcrimjus.2022.101902

Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice, 34*(5), 523-529. https://doi.org/10.1016/j.jcrimjus.2006.09.007

Flores, A. W., Holsinger, A. M., Lowenkamp, C. T., & Cohen, T. H. (2017). Time-free effects in predicting recidivism using both fixed and variable follow-up periods: Do different methods produce different results. *Criminal Justice and Behavior*, *44*(1), 121-137. https://doi.org/10.1177/0093854816678649

Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections*, 3-8.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*, 489-513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

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. https://doi.org/10.1007/s10979-005-6832-7

Taxman, F. S., Cropsey, K. L., Young, D. W., & Wexler, H. (2007). Screening, assessment, and referral practices in adult correctional settings. *Criminal Justice and Behavior, 34*(9), 1216-1234. https://doi.org/10.1177/0093854807304431

Teo, A. R., Holley, S. R., Leary, M., & McNiel, D. E. (2012). The relationship between level of training and accuracy of violence risk assessment. *Psychiatric Services*, *63*(11), 1089-1094. https://doi.org/10.1176/appi.ps.201200019

Vick, B., Gartside, C., & Orth, R. (2024). Analysis and validation of ORAS assessment across four Pennsylvania counties 2017-2022. Research and Training Center, Indiana University of Pennsylvania. Pennsylvania Commission on Crime and Delinquency.

Vincent, G. M., Guy, L. S., Gershenson, B. G., & McCabe, P. (2012). Does risk assessment make a difference? Results of implementing the SAVRY in juvenile probation. *Behavioral Sciences & the Law*, *30*, 384-405.. https://doi.org/10.1002/bsl.2014

Viljoen, J. L., Goossens, I., Monjazeb, S., Cochrane, D. M., Vargen, L. M., Jonnson, M.R., Blanchard, A. J. E., Li, S. M. Y., & Jackson, J. R. (2025). Are risk assessment tools more accurate than unstructured judgments in predicting violent, any, and sexual offending? A meta-analysis of direct comparison studies. *Behavioral Sciences & the Law, 43*(1), 75-113. . https://doi.org/10.1002/bsl.2698

Appendix

Traditional fidelity monitoring approaches for risk assessment instruments rely on supervisor ratings based on process compliance, with quarterly or annual reviews measuring protocol adherence, completion timing, and professional demeanor. However, these approaches fail to measure what matters most: whether assessments actually predict outcomes. This appendix presents an evidence-based alternative that uses objective performance metrics to identify effective practices and improve assessment quality.

Current fidelity monitoring has fundamental flaws. First, supervisors rate officers on criteria that, while intuitively appealing, have no demonstrated relationship to predictive validity. Fidelity monitoring forms used for decades in the field have not been validated against actual predictive outcomes. Without clear metrics tied to predictive validity, subjective ratings become compliance exercises rather than quality improvement. Second, traditional monitoring fails to connect process to outcomes. An officer may follow every prescribed step yet produce assessments with no predictive value. As documented in this study, individual-level AUC values ranged from .15 to 1.00, with correlation coefficients ranging from -.221 to .716. Traditional fidelity monitoring cannot identify these performance differences because it focuses on process compliance rather than predictive validity. Third, infrequent monitoring cannot detect or correct ongoing problems. By the time annual reviews occur, officers producing assessments below acceptable thresholds may have misclassified hundreds of individuals.

This appendix presents a shift from process-focused to outcome-focused fidelity monitoring using the following evidence-based approach:

Phase 1: Establish Performance Baselines

Calculate individual officer-level AUC values for all assessors with sufficient case volume (minimum n = 30 assessments). Generate performance distributions to identify officers performing below standards. Using enhanced categories based on Rice and Harris's (2005) benchmarks, categorize performance as: smaller than small effect (< .556), small effect (.556-.639), medium effect (.639-.714), and large effect (> .714).

Phase 2a: Identify and Study Excellence

Identify officers performing substantially above system baselines (e.g., AUC > .80) for intensive study. Conduct structured observations of high-performing officers' assessment practices, interviewing them about their approaches to gathering information, scoring protocols, and decision-making processes. Document specific behaviors and practices that distinguish top performers.

Phase 2b: Study Poor Performance to Validate Differences

Interview and observe officers performing below .556 (small effect threshold). This

comparative approach validates that practices actually differ between high and low performers, identifies unexpected systemic barriers or misunderstandings, avoids survivorship bias, and allows for the development of targeted remediation by understanding specific failure points.

Phase 3: Develop Evidence-Based Training

Create training modules based on documented differences between high and low performers. Focus on specific behaviors that differentiate those achieving AUC > .80 from those below .556. Establish automatic retraining requirements for officers falling below .556. Design tiered interventions, for example: mandatory comprehensive retraining for those below .556, targeted skill development for those between .556 and system baselines, and advanced training for those performing well but with room for improvement.

Phase 4: Validate Through Outcomes

Implement training interventions to allow for comparison. Measure post-training AUC values to determine effectiveness. Scale only interventions that demonstrate measurable improvements in predictive validity. Generate evidence-based protocols derived from practices that empirically improve outcomes.

Phase 5: Continuous Monitoring and Improvement

Establish dashboards displaying officer-level AUC values. Create automatic alerts for performance below acceptable thresholds. Implement regular reviews to identify trends and emerging problems. Enable rapid identification and remediation of performance deficits.

The findings presented in this study demonstrate that officer-level predictive validity varies dramatically. This variation challenges current quality assurance practices. By focusing on what matters, accurate prediction of recidivism, rather than on compliance with process checklists, this approach could enhance the reliability and validity of risk assessment implementation. While this study examined officer-level variation, future implementations of performance-based monitoring should also consider supervisor-level effects, as systematic differences in team performance may reflect variations in management practices, training quality, or oversight approaches that could be identified and replicated. The goal is to provide practitioners with the tools and training needed to achieve the predictive validity these instruments were designed to deliver.