ADVANCING RESEARCH IN CORRECTIONS

School of Criminology and Criminal Justice | University of Nebraska – Omaha

6001 Dodge Street, Omaha, NE 68182 · 402.554.2610 · arcorrectionslab.org

Arizona Localized OST Violence and Recidivism Assessment

Development and Validation of Localized General and Violent Reconviction Scores for the Arizona Offender Screening Tool

Prepared for the Arizona Supreme Court, Administrative Office of the Courts, Adult Probation Services Division

Prepared by: John Ursino, Ph.D. and Zachary Hamilton, Ph.D.

This project was supported by Grant No. 15PBJA-23-GK-02436-DGCT awarded by the Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice. Points of view or opinions in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice

Table of Contents

Executive Summary iii
Primary Findings iv
Purpose and Background 1
The Offender Screening Tool and the Field Reassessment Offender Screening Tool 2
The Importance of Localization and Revalidation 3
Outcome-Specific Scores 5
Project Structure 6
Validation of the Existing OST and FROST 8
Validation Data and Methods 8
Findings 9
Predictive Accuracy 9
Performance Across Subgroups 10
Calibration 10
Implications for Tool Development 12
Localizing the OST 13
Study Population 13
Information Available for Development 14
Reconviction Outcomes 15
Assessment Items Considered 16
Criminal Behavior, Current Offense, and Criminal History 18
Needs and Circumstances 18
Age and Sex 18
Predictive Performance 19
Performance of the Localized OST 19
Fairness Assessment 21
How Fairness Was Evaluated 22
Findings by Sex 23
Findings by Race 25
Comparison of Updated and Existing OST Score 27

Sources of Improved Predictive Performance 28
Discussion: Implications for Arizona 30
Using the Scores and Next Steps 31
Implementation and Ongoing Evaluation 33
Recommended next steps 34
Limitations 34
Conclusion 36
References 38
Technical Appendix 41

Executive Summary

Treatment courts were developed to address substance use and mental health needs that contribute to individuals' involvement with the justice system. Although many early programs excluded individuals with violent offense histories, current treatment court standards emphasize individualized consideration of these applicants when legally permissible (All Rise, 2026). This shift has increased the importance of accurately assessing violent recidivism risk to inform eligibility decisions, supervision practices, and treatment planning. However, many correctional risk assessments were developed to predict general reoffending rather than violence specifically, potentially limiting their usefulness for these decisions.

Arizona currently uses the Offender Screening Tool (OST) to assess individuals on adult probation and the Field Reassessment Offender Screening Tool (FROST) to reassess risk during supervision. This project examines how effectively these instruments predict general and violent reconviction and whether their accuracy can be improved using Arizona's own assessment and outcome data. The goal is to strengthen the information available to probation agencies and treatment courts while building on assessment practices already established across the state.

The work was completed under the Adult Treatment Court Training Initiative, funded by the Bureau of Justice Assistance and awarded to All Rise. The broader initiative seeks to improve the assessment and treatment of individuals with histories of violent offending who may be appropriate candidates for treatment courts. Arizona's participation focuses on evaluating and refining its existing probation assessments to better support decisions involving this population.

Primary Findings

The project first validated Arizona's existing assessments using 178,675 OST and 95,563 FROST assessments (Ursino & Hamilton, 2026). The primary findings from that evaluation are summarized below.

- *The OST demonstrated weak accuracy for any and violent reconviction and moderate accuracy for felony reconviction, while the FROST demonstrated negligible-to-weak accuracy across outcomes.*
- *Both tools predicted general reconviction with reasonable calibration but overpredicted violent reconviction by more than 30 percentage points on average.*

The findings indicated that the OST's existing scoring structure provides limited information about violence risk and would substantially overstate the likelihood of violent reconviction if applied in treatment court screening. Since the OST items themselves capture relevant risk factors, they were retained and developed into localized General and Violent Reconviction tools, with scores weighted using Arizona subjects and using prior assessments and supplemented with administrative data. Performance was evaluated and the primary findings are summarized.

- *The localized General and Violent Reconviction scores both demonstrated strong predictive accuracy, compared with weak accuracy identified for the existing OST total score.*
- *Both localized scores were well calibrated, with predicted and observed rates differing modestly for both scales.*
- *Notably, the two scores draw on different information. The Violent Reconviction score is driven by domestic violence status, prior violence-related convictions, and relationship problems, while the General Reconviction score is driven by current felony offense, prior*

convictions, and recent drug use. Notably, several items that contribute heavily to one score may contribute little to the other.

- *Predictive accuracy was broadly similar across sex and racial subgroups. The Violent Reconviction score demonstrated strong predictive performance and the General Reconviction score moderate-to-strong performance for men, women, White individuals, and Non-White individuals.*

Together, the findings indicate that Arizona can obtain substantially improved prediction of violence risk from the tool developed using local subjects and data. Because the two scores draw on different risk factors, they allow treatment court teams to distinguish participants whose risk is primarily for general reoffending from those whose risk is specific to violence, rather than treating any history of violence as disqualifying. Further, the general recidivism tool provides an improved prediction for all individuals and may provide a potential replacement for the OST and FROST.

Purpose and Background

Treatment courts integrate judicial supervision with substance use and mental health treatment to address factors contributing to criminal behavior. Many courts have historically excluded individuals with current or prior violent offenses, in part because federal treatment court funding restricted participation by individuals defined as violent offenders (34 U.S.C. § 10611). Exclusions based on charge type alone, however, do not distinguish between individuals with a single violent conviction from individuals who indicate a meaningful likelihood of future violence. Recognizing this limitation, practice standards for treatment courts encourage admission decisions based on validated assessment rather than categorical exclusion (National Association of Drug Court Professionals, 2018).

Actuarial risk assessment provides one way to make these distinctions. Rather than relying solely on the nature of the current charge or on individual discretion, an actuarial assessment examines how case characteristics relate to subsequent outcomes among previously assessed individuals and combines those characteristics into a structured estimate of relative risk. Under the Risk-Need-Responsivity (RNR) framework, the intensity of supervision and services should generally match an individual's assessed risk, while intensive intervention with lower-risk individuals may produce limited benefit, or increase, criminal behavior (Andrews et al., 1990; Bonta & Andrews, 2007; Lowenkamp et al., 2006). As such, an assessment that separates general and violent risk allows treatment court teams to identify participants who can be served safely, those who may require additional supervision and services, and those for whom participation may not be appropriate.

The value of risk assessment lies in providing an informed basis for correction, treatment, and case-management decisions. By examining patterns of reoffending among individuals with

similar characteristics, these instruments estimate differences in the likelihood of future offending (Bucklen et al., 2021). Such estimates provide a consistent empirical foundation for professional judgment, allowing treatment court teams to consider assessed risk, alongside other relevant information when making eligibility, supervision, and treatment decisions. Their usefulness, however, depends on how accurately the underlying instruments assess recidivism risk among the populations they serve. Recognizing the importance of alignment between assessment structure and local populations, Arizona developed its own risk assessment instruments using information specific to its adult probation population.

The Offender Screening Tool and the Field Reassessment Offender Screening Tool

Arizona's use of the Offender Screening Tool (OST) provides an established foundation for assessing risk and needs among individuals on adult probation. The OST emerged from the Maricopa County Adult Probation Department's 1996 strategic planning process, which emphasized evidence-based practice and more consistent risk assessment. After reviewing existing risk and needs instruments, the department chose to develop its own tool, drawing on the assessment research available at the time, and emphasizing static and dynamic factors with established relationships to recidivism (Ferguson, 2002). The OST was fully implemented in Maricopa County by 1998 and adopted by the remaining Arizona counties by 2005, establishing a statewide framework for probation classification and case planning (Ferguson, 2002).

The Field Reassessment Offender Screening Tool (FROST) was introduced in 2005 to complement the OST by capturing changes in risk and need during supervision. Although the instruments are similar, the FROST focuses on individuals' behavior during the preceding year, allowing probation officers to update assessments as circumstances change. Both instruments contain 44 items across ten domains: Physical Health/Medical, Vocational/Financial, Education,

Family and Social Relationships, Residence and Neighborhood, Alcohol, Drug Abuse, Mental Health, Attitude, and Criminal Behavior. The Physical Health/Medical domain informs responsivity but is excluded from the total risk score (Ferguson, 2002). The OST and FROST therefore offer an opportunity to improve recidivism prediction using information routinely collected during routine supervision, without requiring additional assessments.

The Importance of Localization and Revalidation

Although the OST and FROST were originally developed for Arizona, their scoring structures may no longer be appropriate for the state's current probation population. The OST was developed nearly three decades ago using the population characteristics and correctional practices of that period. Since then, changes in sentencing policies, supervision practices, and the composition of probation populations may have altered the relationships between individual risk factors and subsequent offending. Research on risk assessment has increasingly emphasized that predictive accuracy is not an enduring property of an instrument but depends partly on the population, setting, and outcomes against which it is evaluated (Desmarais et al., 2022; Hamilton et al., 2022). Instruments developed using local data can therefore become less accurate as the conditions under which they operate change. Periodic revalidation is necessary when tools remain unchanged or adjusted as the population served changes over time (Bucklen et al., 2021).

Concerns about predictive accuracy are particularly important when instruments are used to inform decisions about outcomes they were not specifically designed to support, as tools do not predict different forms of reoffending with the same level of accuracy. For example, in an earlier validation study the OST and FROST was found to predict technical violations with moderate accuracy but performed worse for new criminal offenses (Latessa et al., 2019). Differences in predictive accuracy are especially important when assessing less common

outcomes, such as violent reoffending. An instrument may correctly identify which individuals are at greater risk of violent reoffending while still overestimating how likely they are to commit a violent offense (Steyerberg, 2019). Accordingly, an instrument's ability to predict general recidivism does not establish that it can accurately estimate violent risk or reliably support decisions that depend on identifying individuals likely to commit violent offenses.

Limitations in outcome-specific prediction have direct implications for how the OST and FROST are used in Arizona, particularly when assessment results inform treatment court eligibility and other decisions involving public safety. Both instruments were designed primarily to assess general risk and criminogenic needs, rather than to produce separate estimates of violent reconviction. Under their existing scoring structures, individual items are assigned values on the same limited scale, generally ranging from one to three points, and summed to produce an overall risk score. Unfortunately, this approach does not account for how strongly individual items predict recidivism. For example, an item strongly associated with subsequent offending may contribute no more to the total score than one with a relatively weak relationship to that outcome.

The predictive importance of individual items may also differ depending on the outcome being assessed, where an item may be particularly informative for predicting general recidivism while being less useful for predicting violent offenses, and vice versa. Consequently, the existing scoring structures may not fully capture the information contained in the OST and FROST items, particularly when used to inform decisions about violence risk. For treatment courts, accurately estimating violent risk is especially important, as overestimation may unnecessarily restrict access to treatment-oriented alternatives, while underestimation may raise legitimate public safety concerns.

Local revalidation and refinement of the existing scoring structures offer a practical means of addressing these limitations while preserving Arizona's existing assessment infrastructure. Research suggests that the predictive contributions of individual risk factors can vary across jurisdictions, populations, and supervision contexts, supporting the use of local outcome data to evaluate and refine established instruments (Desmarais et al., 2022; Hamilton et al., 2022). Rather than assuming that each OST or FROST item should retain its original contribution to the total score, an empirically derived scoring structure can assign weights based on observed relationships with recidivism among individuals supervised in Arizona. The same information can also be used to develop outcome-specific scoring models, allowing existing assessment items to generate distinct estimates of general and violent recidivism risk. This approach builds on the state's longstanding investment in structured assessment while improving the extent to which resulting scores reflect current probation populations and the particular outcomes relevant to supervision and treatment court decisions.

Outcome-Specific Scores

The development of outcome-specific risk scores reflects a broader recognition that recidivism is not a single, uniform outcome. Individuals who are more likely to commit a new offense are not necessarily those most likely to commit a violent offense, and the factors associated with each outcome may differ in their relative importance. A single general risk score may therefore obscure meaningful differences in the types of reoffending individuals are most likely to engage in. Developing separate scores allows the same assessment information to be weighted according to its relationship with each outcome, producing more targeted estimates of risk without requiring additional information collection. This distinction is particularly relevant

when assessment results inform decisions about supervision intensity, treatment placement, or eligibility for programs where the likelihood of violent offending is an important consideration.

The current project applies these principles by developing separate General Reconviction and Violent Reconviction scores, each with its own empirically derived item weights. These scores are intended to answer related but distinct questions, rather than represent different levels of severity along a single risk continuum. The project also recognizes that jurisdictions differ in the assessment information available to agencies and their capacity to develop locally tailored instruments. Accordingly, the work includes both the refinement of Arizona's existing OST instrument and the development of pooled assessment models intended for use in settings where local model development is not feasible. The following section describes how these components were organized within the broader project.

Project Structure

This report is one product of a multistate effort to develop violence risk tools for treatment court populations. The original design proposed a single unified violence assessment developed from pooled data across participating states. Because the participating states use different assessment systems, with Arizona using the OST and the remaining states using the Level of Service Inventory-Revised (LSI-R) or the Ohio Risk Assessment System (ORAS; a derivative of the OST), there was not enough overlap in available items to responsibly build one common instrument. The project was therefore revised to produce:

1. *A state-specific localized assessment, built from each participating state's own instrument and outcome data; and*

2. *Two pooled violence assessments, one for jurisdictions using the LSI-R and one for jurisdictions using the Ohio Risk Assessment System (ORAS), for courts that cannot develop a jurisdiction-specific tool.*

The present report documents the Arizona state-specific assessment.[1] The next section presents the validation of the existing OST and FROST. The remainder of this report then describes the population and outcomes used to develop the assessment, the items considered for inclusion, the methods used to construct and validate the General and Violent Reconviction models, performance across demographic groups, and the translation of the statistical models into a scoring system suitable for agency use.

[1] Arizona data were not used to develop the pooled assessments, which will be described in a separate report.

Validation of the Existing OST and FROST

Before developing the localized assessment, the research team validated the OST and FROST as currently scored in Arizona (Ursino & Hamilton, 2026). The validation assessed how accurately each tool predicted reconviction within three years of assessment, whether accuracy differed by sex and race, and whether predicted risk aligned with observed reconviction rates. Its findings establish the baseline against which the localized assessment should be judged. The analyses also identify the aspects of existing performance most relevant to outcome-specific redevelopment.

Validation Data and Methods

The validation used Arizona's adult probation assessment records, including 178,675 OST assessments and 95,563 FROST assessments. The records included 136,154 assessments of men and 42,791 assessments of women. Predictive performance was evaluated for three outcomes: any new reconviction, felony reconviction, and violent reconviction, occurring within three years of assessment. For each outcome, performance was examined along two complementary dimensions: discrimination and calibration.

Discrimination measures how well an instrument distinguishes individuals who are subsequently reconvicted from those who are not. An instrument with stronger discrimination generally assigns higher risk scores to individuals who are reconvicted. Discrimination was measured using the Area Under the Receiver Operating Characteristic Curve (AUC), where values below 0.56 indicate negligible accuracy, 0.56 to 0.63 weak accuracy, 0.64 to 0.70 moderate accuracy, and 0.71 or greater strong accuracy (Rice & Harris, 2005). *Calibration* examines whether estimated probabilities of reconviction correspond to actual outcomes. For example, among individuals assigned a 20% probability of reconviction, approximately 20%

should be reconvicted if the estimates are accurate. Calibration was measured using the Expected Calibration Error (ECE), which summarizes the average difference between predicted and observed reconviction rates across score levels. An ECE of 0.03 indicates that predicted risk differed from observed reconviction rates by approximately 3 percentage points on average. The distinction between discrimination and calibration is important for interpreting the validation results. An instrument may successfully identify individuals at relatively higher risk while still overestimating or underestimating their actual likelihood of reconviction.

Findings

The following section describes findings from study analyses. As described, we examined overall predictive performance of the OST and FROST, but also whether that performance varies across demographic subgroups and whether estimated risk corresponds to observed outcomes. Results are presented first for discrimination, followed by subgroup differences and calibration.

Predictive Accuracy

Table 1 displays AUCs for both tools across the three outcomes. The OST demonstrated weak accuracy for any reconviction (AUC = 0.63) and violent reconviction (AUC = 0.62) and reached the threshold of moderate accuracy for felony reconviction (AUC = 0.64). The FROST demonstrated negligible accuracy for any reconviction (AUC = 0.55) and weak accuracy for felony (AUC = 0.56) and violent reconviction (AUC = 0.59). Thus, neither tool meaningfully distinguished individuals who were later reconvicted of a violent offense from those who were not. Out of the 30 tests conducted, predictive discrimination was identified to be a negligible effect 7 times, a weak effect 11 times, and a moderate effect 12 times.

Table 1. AUCs for the OST and FROST Across Outcomes

Group	OST Any	OST Felony	OST Violent	FROST Any	FROST Felony	FROST Violent
Total	0.63	0.64	0.62	0.55	0.56	0.59
Male	0.62	0.64	0.62	0.55	0.57	0.59
Female	0.64	0.65	0.65	0.54	0.55	0.61
White	0.64	0.64	0.61	0.51	0.54	0.61
Non-White	0.64	0.65	0.56	0.58	0.62	0.51

Note: AUC values below 0.56 indicate negligible accuracy, 0.56 to 0.63 weak accuracy, 0.64 to 0.70 moderate accuracy, and 0.71 or greater strong accuracy.

Performance Across Subgroups

Differences by sex were small. The OST performed slightly better for women, reaching moderate accuracy for all three outcomes (AUCs range: 0.64 to 0.65), while it reached moderate accuracy for men only for felony reconviction (AUC = 0.64). The FROST demonstrated negligible, or weak, accuracy for both men and women across outcomes. The OST demonstrated moderate accuracy for White and Non-White individuals for any and felony reconviction (AUCs of 0.64 to 0.65). For violent reconviction, its accuracy was weak for White individuals (AUC = 0.61) and borderline negligible-to-weak for Non-White individuals (AUC = 0.56). The largest gap appeared for the FROST's prediction of violent reconviction, where accuracy was weak for White individuals (AUC = 0.61) and little better than chance for Non-White individuals (AUC = 0.51).

Calibration

Table 2 displays ECEs for both tools. Calibration errors were smallest for any reconviction, with predicted and observed rates differing by roughly 6 percentage points for the OST and 8 for the FROST. Errors were larger for felony reconviction (ECEs of 0.10 for the OST and 0.08 for the FROST). Both tools substantially overpredicted violent reconviction, with predicted risk exceeding observed rates by about 33 percentage points for the OST and 31 for the

FROST. This level of overprediction leads to supervision and treatment decisions based on inflated estimates of violent risk.

Table 2. ECEs for the OST and FROST Across Outcomes

Outcome	OST	FROST
Any Reconviction	0.06	0.08
Felony Reconviction	0.10	0.08
Violent Reconviction	0.33	0.31

Figure 1 illustrates the calibration of the OST for general reconviction. The diagonal dotted line represents perfect calibration, where predicted probabilities match observed reconviction rates. For example, a predicted probability of 0.50 is associated with a return rate of 50%. The calibration trend generally follows this line across the lower and middle score ranges but falls below it at higher scores, indicating that the OST overestimates the likelihood of reconviction among individuals with the highest predicted risk.

Figure 1. OST Calibration for Any Reconviction

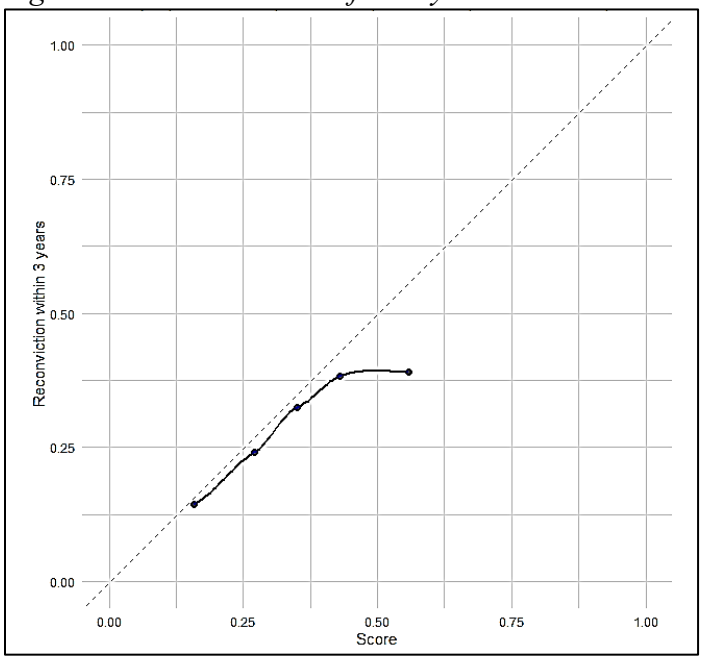
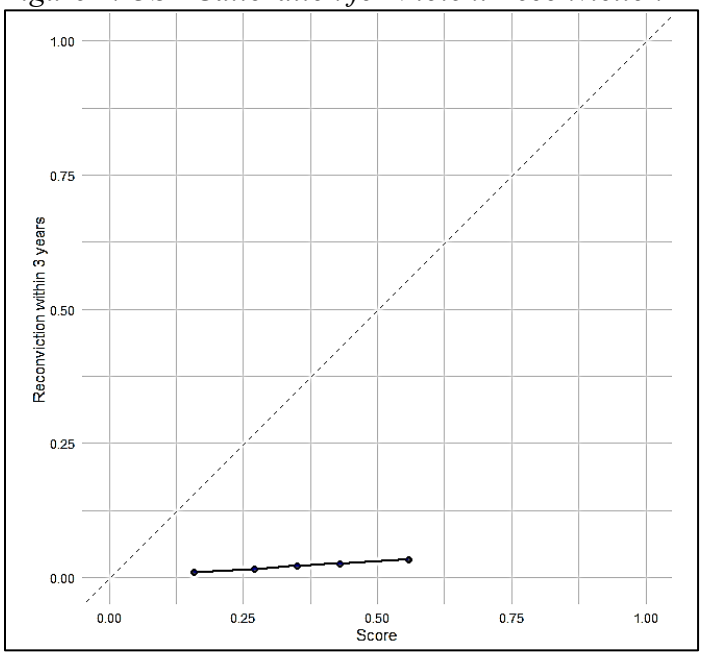


Figure 2 presents the corresponding results for violent reconviction, revealing substantially greater miscalibration. Unlike the pattern for any reconviction, the calibration trend

remains well below the diagonal across the entire score range. Predicted probabilities reach approximately 55%, while observed violent reconviction rates remain below 5%. This substantial gap indicates that the OST's existing risk estimates considerably overstate the likelihood of violent reconviction. Consequently, individuals may be assigned to higher risk categories than their observed likelihood of violent reconviction would warrant, affecting supervision intensity and treatment decisions.

Figure 2. OST Calibration for Violent Reconviction



Implications for Tool Development

The current validation found that, although the OST and FROST may provide information relevant to general risk classification and case planning, neither instrument demonstrated sufficient predictive performance to support decisions specifically concerning violent recidivism risk and often failed to produce consistent and efficient prediction of any and felony outcomes. The FROST's comparatively weak performance across outcomes also raised concerns about its utility for reassessing risk during supervision.

Further, the findings reinforce the importance of developing outcome-specific scoring models and informed the selection of assessment data for the localized models. A localized tool reweights assessment items and administrative information separately for general and violent reconviction, allowing each factor's contribution to reflect its relationship with the outcome being predicted rather than relying on the existing total score. The sections that follow describe the development of these models and evaluate whether they improve on the performance documented here.

Localizing the OST

Because the initial validation established the need for outcome-specific scoring, the next phase used Arizona assessment and administrative records to develop separate General and Violent Reconviction models, then evaluated both models on data withheld from development. This cross-validation approach allows the development results to be assessed against a previously documented baseline rather than interpreted in isolation. The following sections describe the sample, predictors, and estimation procedures.

Study Population

The localized OST was developed using records for 178,945 individuals assessed with the OST in Arizona between June 2000 and March 2023. While records in the data extended to 2025, cases after March 2023 did not allow a sufficient follow-up period to identify reconvictions and were therefore excluded. Records were drawn from Arizona adult probation assessment data provided by the Adult Probation Services Division. The records allowed the research team to observe each individual's assessment information and then determine whether a new conviction occurred during the three years that followed.

To develop and evaluate the localized OST, the study population was randomly divided into an 80% training sample and a 20% test sample, stratified by reconviction outcome to ensure similar proportions of individuals who were reconvicted in each sample. The training sample was used to identify the assessment items associated with reconviction and estimate their respective scoring weights. These weights were then applied to the independent test sample to evaluate how accurately the completed assessment predicted reconviction among individuals who were not involved in its development. The test sample was not used to select items or estimate weights, providing a more realistic estimate of predictive performance than evaluating the assessment using the same individuals on which it was developed. Table 3 summarizes the study population, with the training and test samples displayed separately to document the comparability of their characteristics.

Table 3. Characteristics of the Arizona Study Population

Characteristic	Training Sample	Test Sample	Total
Individuals (N)	143,156	35,789	178,945
Male (%)	76.1	76.0	76.1
White (%)	49.0	49.2	49.1
Non-White (%)	51.0	50.8	50.9
General Reconviction within 3 years (%)	33.4	33.4	33.4
Violent Reconviction within 3 years (%)	10.9	11.1	11.1

Information Available for Development

The project did not begin with a new list of risk factors. Instead, it evaluated information Arizona already collects through the OST and its supervision records. Candidate measures came from three existing sources:

- *OST assessment items covering criminal behavior, education, vocational history, family and social relationships, residence, alcohol use, drug use, mental health, physical health, and attitudes.*

- *Current offense and supervision status information, including whether the current offense is a felony or undesignated offense and whether the individual carries domestic violence, DUI, or sex offender status.*

- *Criminal history and demographic information, including counts of prior felony, misdemeanor, and violent convictions, prior prison sentences, age at assessment, highest grade completed, and sex.*

Most OST items record whether a risk factor is present, absent, or, for some items, present as a strength. For model development and scoring, each OST item was treated as indicating whether the risk factor was present, with protective responses grouped with absence of the risk factor. The technical appendix describes protective responses as recoded to missing before imputation; this procedure must be reconciled with operational zero-point scoring. Cases missing an outcome or item responses were excluded.

Reconviction Outcomes

An assessment is understood in relation to the outcome it was developed to predict. Two outcomes were selected for this project, both measured over the three years following assessment. A fixed follow-up period allows the two outcomes to be evaluated on a consistent time horizon.

General Reconviction was defined as a criminal reconviction within three years of assessment. Among individuals with complete three-year follow-up, 33.4% experienced General Reconviction. This outcome captures the broader likelihood of renewed criminal justice involvement serious enough to result in a conviction.

Violent Reconviction was defined as a new conviction for a violent offense within three years. Among individuals with complete three-year follow-up, 11.1% experienced Violent Reconviction. Violent Reconviction was modeled separately because violence represents a

particular public-safety concern for treatment court admission and supervision decisions and occurs far less often than general reconviction.

Assessment Items Considered

Fifty-six candidate measures were evaluated for each outcome, including 44 OST items and 12 measures of current offenses, criminal history, and demographic characteristics. Separate models were developed for General and Violent Reconviction to identify which items contributed to predicting each outcome and how much weight they should receive. Model development accounted for overlap among assessment items. For example, individuals with multiple prior convictions may also have prior felony convictions or a history of increasingly serious offenses. Although each measure may be associated with reconviction, they often capture similar information. We used ridge regression to account for these relationships, select the most informative items, and estimate their contributions to each outcome. All items were coded so that higher values indicated greater risk, and items receiving negative weights were excluded from the final models. Additional details on model development and estimation procedures are provided in Appendices A1 and A2. Table 4 summarizes the candidate measures by domain and the number retained in each model.

Table 4. Candidate and Retained Items by Domain

Domain	Candidate Items	Retained, General	Retained, Violent	Content
Criminal Behavior	9	5	7	Age at first arrest, number and seriousness of prior convictions, post-conviction designation, prior felony and violence-related convictions
Education	3	0	1	Educational attainment and reading ability
Vocational	5	3	3	Current employment, unemployment history, job loss, financial situation
Family / Social	8	6	8	Upbringing, family criminal history, family relations, childhood exposure to violence, romantic and peer relationships
Residence	2	2	2	Residential arrangement and stability
Alcohol	3	1	2	Alcohol-related arrests and the effect of alcohol use on the individual's life
Drug	3	3	1	Recent illegal drug use and the individual's perception of that use
Mental Health	2	2	2	History of mental illness and current functioning
Health	2	2	1	Health problems that interfere with supervision or employment
Attitude	7	2	3	Attitudes supportive of crime, attitudes toward authority, motivation to change
Current Offense and Status	5	3	1	Felony or undesignated current offense; DUI, domestic violence, and sex offender status
Criminal History Counts	4	3	2	Prior felony, misdemeanor, and violent convictions; prior prison sentences
Demographics	3	3	3	Age at assessment, highest grade completed, sex
Total	**56**	**35**	**36**	

The complete item wording and response options are provided in the accompanying scoring workbook. Retention counts are useful for describing model composition, but they should not be read as a ranking of the causal importance of domains. Items with similar content may cancel out each other's predictive weight, and a domain with many retained items may

contribute less to the final score than a domain with only one or two highly weighted items. Likewise, an item assigned no points in one model is not necessarily unrelated to that outcome in isolation; but may offer little additional information once other items are included.

Criminal Behavior, Current Offense, and Criminal History

Criminal history is consistently among the strongest predictors of future offending (Gendreau et al., 1996). The localized OST draws on this information in several forms: the OST Criminal Behavior items, which record the timing, extent, and seriousness of prior offending; current offense and status indicators from supervision records; and counts of prior convictions and prison sentences. These measures overlap, and the modeling approach described in the next section outlines the predictive value each provides.

Needs and Circumstances

The Education, Vocational, Family / Social, Residence, Alcohol, Drug, Mental Health, Health, and Attitude domains describe an individual's current circumstances and needs. Unlike prior convictions, many of these items can change through intervention or changing life circumstances. Their inclusion means that an individual's scores can change at reassessment when circumstances such as employment, substance use, relationships, or residential stability change. Their predictive contributions should not be taken as evidence that addressing any particular need will necessarily reduce reconviction.

Age and Sex

Age is among the most consistently observed predictors of offending, with risk generally declining across adulthood. Sex was included as a candidate predictor so that the model could account for baseline differences in the rates of reconviction between men and women.

Predictive Performance

Once the General and Violent Reconviction models were finalized using the training sample, they were applied to test sample. Performance was evaluated using discrimination and calibration, with results compared against the existing OST to determine whether the updated models improved predictive accuracy. Discrimination was measured using the AUC statistic to assess how well each model distinguished between individuals who were and were not reconvicted (Rice & Harris, 2005). Calibration was evaluated by comparing predicted probabilities with observed reconviction rates across ten groups of approximately equal size and summarized using Expected Calibration Error (ECE), with lower values indicating better agreement. Calibration plots were also examined to identify patterns of overprediction and underprediction, and bootstrap confidence intervals were calculated for ECE estimates.

Performance of the Localized OST

Table 5 summarizes discrimination and calibration for the localized versions of Arizona's tool, using the test sample. General Reconviction achieved an AUC of 0.71 (95% CI 0.70–0.71), and Violent Reconviction achieved an AUC of 0.73 (95% CI 0.72–0.73), both within the "strong" predictive accuracy range. The corresponding ECEs were approximately 0.02 and 0.01, indicating 2% and 1% expected error, or "excellent" calibration.

Table 5. Test Sample Performance

Outcome	AUC (95% CI)	ECE
General Reconviction	0.71 (0.70–0.71)	0.02
Violent Reconviction	0.73 [0.72-0.73I]	0.01

Although the models demonstrated strong discrimination, their ability to distinguish individuals by risk does not necessarily mean that predicted probabilities accurately reflect the likelihood of reconviction. Accurate probability estimates are particularly important for violent

reconviction since violent offenses are relatively uncommon, meaning that individuals identified as high-risk relative to others may still have a relatively low overall probability of being reconvicted for a violent offense. The calibration analyses therefore examine how closely predicted probabilities derived from risk scores correspond to observed reconviction rates.

Figure 3 presents the calibration results for General Reconviction, comparing average predicted probabilities with observed reconviction rates across ten risk groups. Predictions closely tracked observed outcomes throughout the distribution, with the largest differences remaining below 3.5 percentage points. In the high-end of the range, where error was largest, predicted risk averaged 51.4% and the observed rate was 48.1%. The consistency of the updated tools indicates that the model provides more accurate estimates of general reconviction risk across both lower- and higher-risk groups, as compared to the current OST and FROST tools.

Figure 3. Observed and Predicted General Reconviction by Predicted-Probability Decile

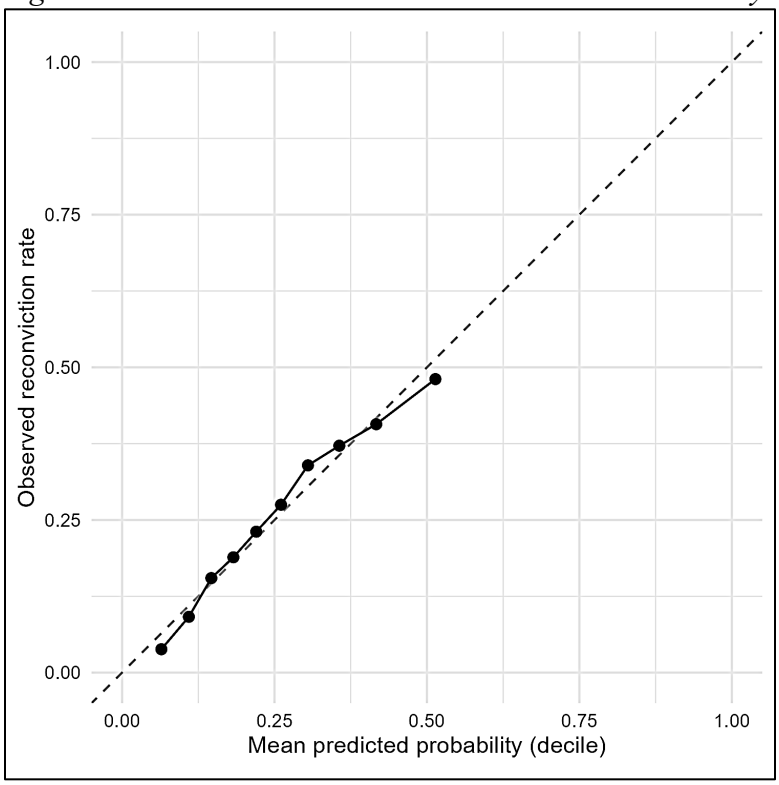
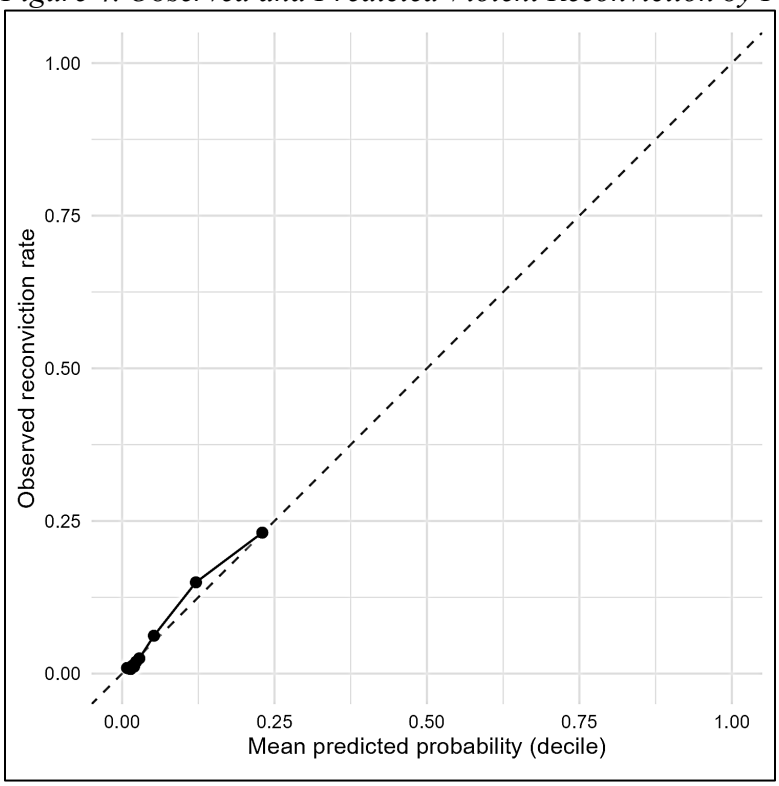


Figure 4 presents the corresponding results for Violent Reconviction. Despite this outcome's lower base rate, predicted probabilities closely aligned with observed rates across most risk groups. The largest departure occurred near the middle of the range, where the model predicted a 12.1% reconviction rate compared with an observed rate of 14.9%, a less than 3% difference. Notably, predicted and observed rates were nearly identical at the highest-risk end of the scale (24.0% predicted and 24.1% observed). Taken together, the updated model provides more accurate estimates of violent reconviction, as compared to the current OST and FROST tools.

Figure 4. Observed and Predicted Violent Reconviction by Predicted-Probability Decile



Fairness Assessment

An assessment can perform well for the overall population while performing differently for particular groups. For that reason, the localized OST was also evaluated separately by sex and by race. The purpose of this analysis was not simply to determine whether groups received

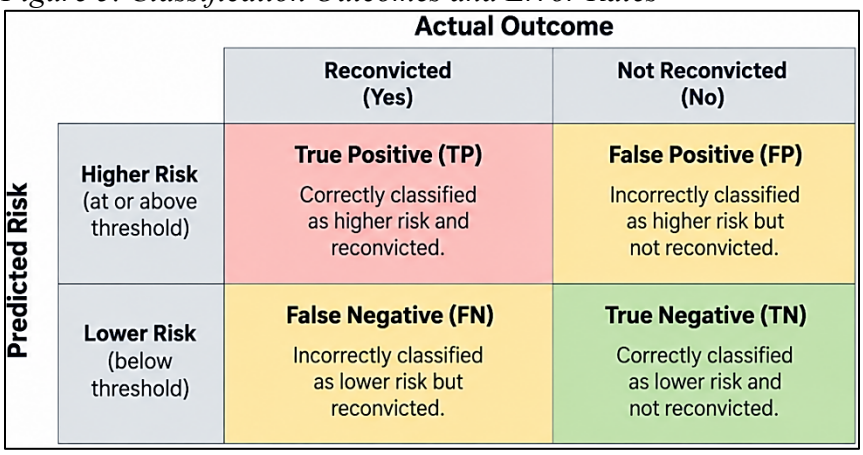
different average scores. Groups can differ in their observed rates of reconviction, and score distributions may differ even when an assessment predicts equally well. The analysis instead examined whether the assessment functioned similarly when predicting reconviction within each group.

As stated previously, race was not used to calculate either score and was retained only to evaluate the completed models. Sex was included as a predictor, as described earlier, so that each model could account for differences in reconviction rates between men and women. These choices should be considered when interpreting both the subgroup comparisons and the operational implications of the scores. The exclusion of race does not rule out group differences arising through other predictors.

How Fairness Was Evaluated

Predictive performance was examined across demographic groups using AUCs and comparisons of predicted and observed reconviction rates. Classification errors were also examined by applying the same risk threshold across groups for each outcome. Figure 5 illustrates the four possible classification outcomes. True positives and true negatives represent correct classifications, while false positives and false negatives represent incorrect classifications. The false-positive rate (FPR) measures the percentage of individuals who were not reconvicted but were classified as higher risk. The false-negative rate (FNR) measures the percentage of individuals who were reconvicted but were classified as lower risk. Comparing these rates across demographic groups helps identify whether certain groups are more likely to experience either type of classification error.

Figure 5. Classification Outcomes and Error Rates



To compare classification errors across groups, we used reference thresholds of 26% predicted probability for general reconviction and 4.3% for violent reconviction. These thresholds were used solely for comparison and are not recommended for classification or eligibility decisions. Appendix A4 presents alternative thresholds and their associated error rates. For the racial subgroup analyses, individuals were classified as White or Non-White based on available administrative records. The Non-White category combined multiple racial and ethnic groups because the number of reconviction events within individual groups, particularly violent reconvictions, was insufficient to produce reliable estimates. Although this aggregation improved statistical stability, it does not imply that the groups have similar outcomes and limits the ability to identify differences among them.

Findings by Sex

Predictive accuracy was similar for men and women, although both models performed slightly better for women. The General Reconviction score demonstrated borderline moderate-to-strong accuracy for men (AUC = 0.70) and strong accuracy for women (AUC = 0.73), while the Violent Reconviction score demonstrated strong accuracy for both groups (AUC = 0.71 for men and 0.75 for women). Despite these relatively small differences in predictive accuracy,

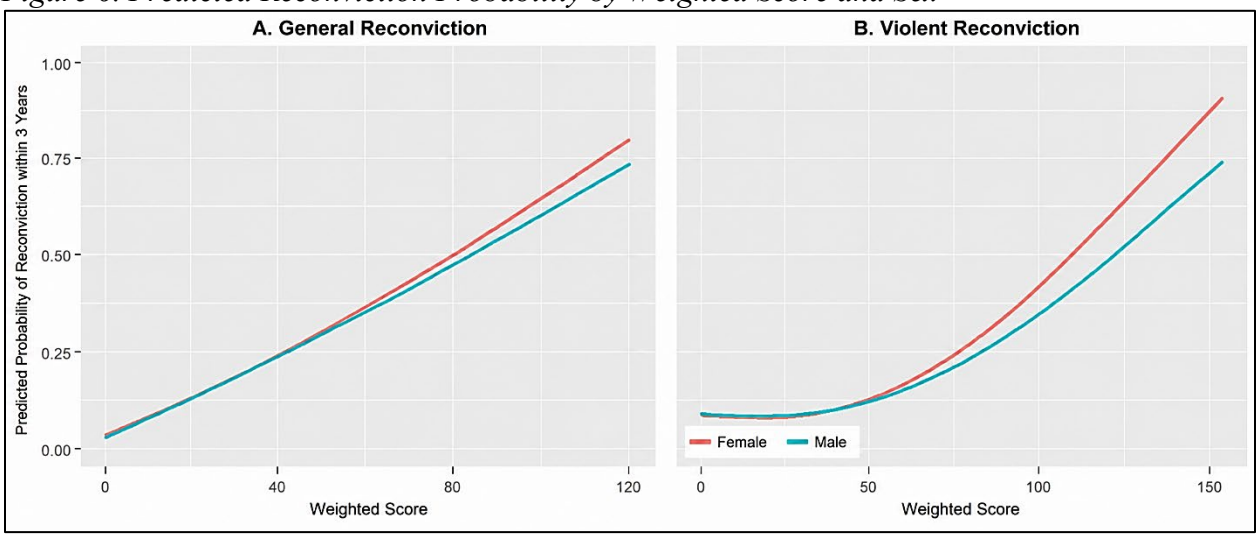
classification error rates at the derived thresholds varied more noticeably. For General Reconviction, men had a modestly higher false-positive rate than women (37.5% versus 32.6%) but a slightly lower false-negative rate (32.6% versus 34.2%). Differences were more pronounced for Violent Reconviction, where the false-positive rate was 22.2% for men compared to 13.8% for women, in part due to the rarity of cases where women violently reoffend.

Table 6. Performance by Sex, Test Sample

Outcome	Group	N	Reconviction (%)	AUC	FPR	FNR
General	Men	27,184	36.1	0.70	37.5	32.6
General	Women	8,605	24.7	0.73	32.6	34.2
Violent	Men	27,184	14.4	0.71	22.2	20.2
Violent	Women	8,605	9.8	0.75	13.8	22.7

Figure 6 illustrates how predicted reconviction probabilities change across weighted risk scores for men and women. For General Reconviction (Panel A), the curves follow nearly identical patterns across most scores, with a modest separation emerging at the upper end of the risk scale. The difference reaches approximately six percentage points only at the highest scores shown. Violent Reconviction (Panel B) follows a similar pattern, with predicted probabilities increasing more sharply at higher scores and women receiving somewhat higher estimates at equivalent scores. Overall, both panels demonstrate considerable consistency in how weighted scores translate into predicted probabilities across sexes, with differences concentrated at the upper ends of the score ranges. The divergence between men and women at these scores may reflect the smaller number of women at the extremes of the score distributions, due to then committing violent offenses less often and having lower base rates for general reoffending.

Figure 6. Predicted Reconviction Probability by Weighted Score and Sex



Taken together, the findings indicate that both models perform similarly across sexes, with slightly stronger predictive accuracy for women. The close alignment of predicted probability curves further suggests that weighted scores correspond to comparable estimates of reconviction risk for men and women. Overall, the results provide encouraging evidence of comparable predictive performance for men and women, while identifying classification errors as an area warranting continued attention when selecting thresholds for translating tool scores into operational decisions.

Findings by Race

Differences in predictive accuracy by race were modest. The General Reconviction score demonstrated strong accuracy for White individuals (AUC = 0.71) and borderline moderate-to-strong accuracy for Non-White individuals (AUC = 0.70), a less-than-substantial difference. Similarly, the Violent Reconviction score demonstrated strong accuracy for both White (AUC = 0.73) and Non-White (AUC = 0.72) individuals.

However, classification error rates at the reference thresholds differed across groups. For General Reconviction, White individuals possessed a higher false-positive rate than Non-White

individuals (32.9% versus 23.2%), while false-negative rates were similar (15.1% versus 16.7%). For Violent Reconviction, Non-White individuals had a higher false-positive rate (14.4% versus 9.3%) but a substantially lower false-negative rate (16.7% versus 35.4%).

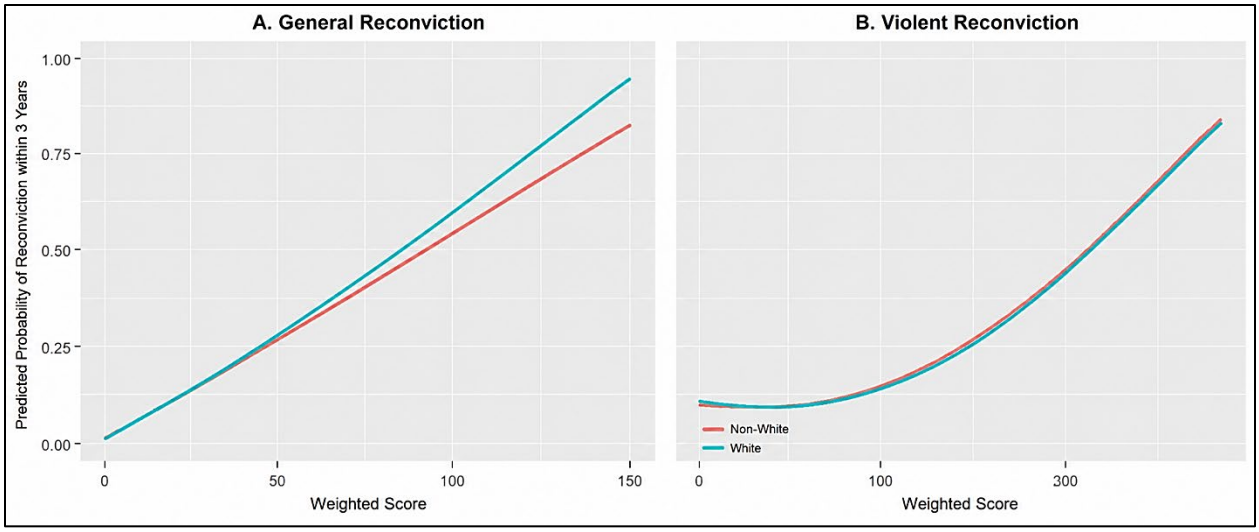
Table 7. Performance by Race, Test Sample

Outcome	Group	N	Reconviction (%)	AUC	FPR	FNR
General	White	17,592	31.8	0.71	32.9	15.1
General	Non-White	18,197	34.7	0.70	23.2	16.7
Violent	White	17,592	10.6	0.73	9.3	35.4
Violent	Non-White	18,197	12.5	0.72	14.4	16.7

Figure 7 illustrates the relationship between weighted scores and predicted reconviction probabilities across racial categories. For General Reconviction (Panel A), the trendlines are closely aligned through the lower and middle score ranges, with some separation emerging at higher scores (approximately 10%). White individuals receive somewhat higher predicted probabilities at equivalent scores toward the upper end of the risk score distribution, although the overall patterns remain similar. For Violent Reconviction (Panel B), the curves overlap almost entirely across the full score range, indicating virtually identical predicted probabilities at equivalent scores. Together, the panels demonstrate broadly consistent relationships between weighted scores and estimated risk across racial categories, particularly for violent reconviction.

Overall, the models demonstrated comparable ability to distinguish individuals by risk across racial categories, although differences in classification errors were more pronounced, particularly for violent reconviction. The differences in error rates highlight the importance of examining classification performance across groups rather than relying solely on overall predictive accuracy when evaluating the instruments.

Figure 7. Predicted Reconviction Probabilities by Weighted Score and Race/Ethnicity



Comparison of Updated and Existing OST Score

The existing OST total score was also evaluated using the same test sample and reconviction outcomes to determine whether the new models predict general and violent reconviction more accurately than the existing assessment.

The earlier OST validation (Tables 1 and 2) used all available assessments. For an apples-to-apples comparison, Table 8 evaluates the existing and updated OST using the same test sample. The updated models improved both predictive accuracy and calibration. For General Reconviction, AUC values increased from 0.63 to 0.71, while ECE decreased from 0.06 to 0.02. For Violent Reconviction, AUC increased from 0.62 to 0.73, while ECE decreased from 0.33 to 0.01. Notably, the updated models moved the predictive accuracy needle from a weak, to a strong model for both general and violent outcomes. Improvements in predictive accuracy are particularly important for violent reconviction, as continued use of the existing OST could lead to substantially overestimations of individuals' likelihood of violent reoffending, potentially resulting in unnecessarily restrictive supervision and treatment decisions. In contrast, the updated tools demonstrate that improved methods and localized weighting will provide substantial improvement to Arizona's prediction of recidivism.

Table 8. Comparison of Localized and Existing OST Scores

Outcome	Original AUC	Localized AUC	Original ECE	Localized ECE
General Reconviction	0.63	0.71	0.06	0.02
Violent Reconviction	0.62	0.73	0.33	0.01

Sources of Improved Predictive Performance

The final models show which items are the strongest predictors for both outcomes. Criminal history and current offense information contribute heavily to both outcomes, consistent with the broader recidivism literature (Gendreau et al., 1996), but the specific measures differ by outcome. These patterns provide context for understanding why separate outcome-specific models were retained.

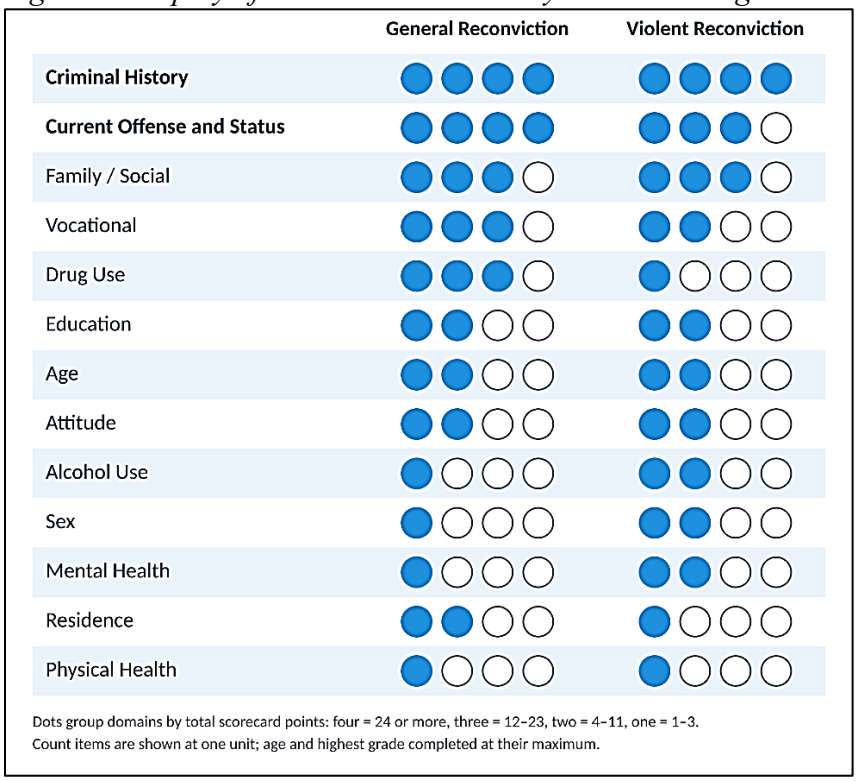
For General Reconviction, the largest contributions come from a current felony offense, a felony or undesignated post-conviction designation, three or more prior convictions, an increase in offense seriousness, and prior prison sentences. Needs-related items also contribute meaningfully, particularly recent drug use and the individual's perception that drug use interferes with life, repeated unemployment, and negative peer relationships. These contributions combine information about criminal history with measures of current circumstances. Their weights describe predictive associations after accounting for other retained items.

For Violent Reconviction, the largest contributions come from domestic violence status, a felony or undesignated post-conviction designation, prior violence-related convictions, and three or more prior convictions. Relationship and family information also contribute more to the violence model than to general reconviction. All eight Family/Social items carry weight in the violent model, led by a problematic or absent romantic relationship and/or negative family relations. Drug use items contribute far less to violence than to general reconviction prediction, while alcohol-related arrests contribute more.

Retaining separate models for general and violent reconviction creates a multiband assessment that provides distinct risk estimates for different outcomes. Prior research supports this approach, demonstrating the value of assessing multiple recidivism outcomes rather than relying on a single overall risk score (Duwe 2014; 2024; Hamilton et al., 2022; 2025).

The differing contributions of assessment domains highlight the value of developing separate scoring models for general and violent reconviction rather than relying on a single measure of risk. Figure 8 illustrates the relative contribution of each domain to the General and Violent Reconviction scores. Criminal history carries the greatest weight in both models, while current offense information contributes more to general reconviction. Vocational and drug-related factors also carry greater weight in the general model, whereas family and social factors contribute more prominently to violent reconviction.

Figure 8. Display of Which Domains Carry the Most Weight



A current felony offense contributes 16 points to General Reconviction and 0 to the Violent Reconviction model, while domestic violence status contributes 12 points to Violent Reconviction and none to General Reconviction. These examples further illustrate how the same information can have different predictive value depending on the outcome.

Although item weights help predict recidivism, they should not be interpreted as direct measures of treatment needs, which should continue to be identified through appropriate clinical and case-planning processes. The differences between the models also have an important practical implication. A single overall risk score can obscure variation in the type of outcome being predicted. Presenting the General and Violent Reconviction scores side by side allows teams to consider broad recidivism risk without assuming that it is interchangeable with violence-specific risk. A higher general score does not automatically imply an equivalently high violent score. Decisions should therefore refer to the specific outcome relevant to the question being considered.

Discussion: Implications for Arizona

The findings support a practical strategy for improving correctional assessment: using locally observed outcomes to reassess how existing information is combined. Arizona currently collects extensive information through the OST and related records. Outcome-specific weighting makes it possible to extract a more accurate predictive model from those same data with limited additional collection burden. The value of this approach lies partly in preserving existing assessment workflows while changing how their information is weighted.

The separate models also clarify a distinction that is easy to lose in practice. General reconviction and violent reconviction are related but not interchangeable. A tool optimized for general criminal justice involvement may be insufficiently informative for decisions specifically

concerned with violence. The observed differences in item contributions provide a rationale for maintaining two scores rather than treating violence risk as merely the upper end of one general risk continuum.

At the same time, the development results do not demonstrate that implementing the scorecard will improve treatment court outcomes. The models were evaluated retrospectively on historical probation cases. They have not yet established prospective accuracy among treatment court referrals, nor have they been evaluated as part of an admission or supervision decision process. The next stage is therefore not simply to select a numerical cutoff; it is to determine how the scores can improve the consistency and proportionality of decisions without displacing individualized clinical and judicial review.

Using the Scores and Next Steps

The localized OST produces separate General and Violent Reconviction scores that can inform treatment court screening, supervision, and service decisions. However, the thresholds used to translate these scores into risk classifications have not yet been established. Although the statistical development process determines how individual factors contribute to predicted risk, it does not determine where agencies should draw the boundaries between risk levels. Those decisions depend on how the scores will be used, the populations being assessed, and the consequences of incorrectly classifying someone as higher or lower risk. A threshold appropriate for identifying individuals who may benefit from additional services, for example, may not be appropriate for determining eligibility for a treatment court.

Establishing these thresholds therefore requires Arizona stakeholders to first identify the decisions each score is intended to support. The two scores provide different, but complementary information, and their practical value depends on how that information is incorporated into

existing policies and procedures. Several potential applications illustrate how the scores could be used:

- **Treatment court screening:** The Violent Reconviction score should inform admission decisions by identifying candidates who may warrant additional review based on their risk of violent reoffending. This would support more individualized decisions without relying exclusively on current charges, offense histories, or automatic exclusions based on a single score that combines general and violent reconviction risk.
- **Supervision classification:** The General Reconviction score should inform decisions about supervision intensity, consistent with the risk principle that higher-risk individuals generally benefit from more intensive interventions while intensive supervision of lower-risk individuals may be counterproductive. Classification thresholds would help Arizona agencies allocate supervision resources according to assessed risk.
- **Targeted intervention and services:** The Violent Reconviction score will help identify individuals who may benefit from additional assessment for violence-focused interventions, including domestic violence programming or cognitive-behavioral treatment. In this context, the score would serve as an initial indicator of risk rather than a substitute for evaluating individual treatment needs.

Considering the General and Violent Reconviction scores together allows agencies to distinguish between different types of risk when making supervision and treatment decisions. An individual with elevated general risk but relatively low violent risk may warrant more intensive supervision without necessarily requiring violence-focused intervention. Conversely, an individual with lower general risk but elevated violent risk may require a different combination of supervision, assessment, and services.

Although the project initially focused on developing a violence risk assessment for Arizona treatment court eligibility decisions, the resulting models have broader applications. Using information already collected statewide, the research team developed updated OST models predicting both general and violent reconviction. Given their improved predictive performance relative to the existing OST, we recommend that Arizona consider adopting the updated models statewide to improve risk estimation across the probation population, while also supporting treatment court eligibility decisions.

Regarding next steps, the courts and probation agencies should determine which populations and decisions the updated models will best be used and establish appropriate risk classification thresholds. Arizona stakeholders should identify the proportion of individuals to be assigned to each risk category and the corresponding policy implications, including supervision intensity, treatment referrals, and treatment court eligibility. These priorities should guide the selection of score cutoffs that align with the intended uses of each model.

Implementation and Ongoing Evaluation

Implementation should specify when scores are calculated, who verifies the underlying OST and administrative inputs, how missing or disputed information is handled, and how the results of the general and violent models are presented to decisionmakers (All Rise, 2026). Written guidance should also address the documentation of overrides and the use of scores in treatment planning. Clear procedures can help to promote consistent calculation and interpretation across settings.

After implementation, Arizona should monitor score distributions, missingness, item coding, use of overrides, and the proportion of referrals assigned to each risk category. As three-year outcomes become available, the updated assessment should be revalidated, as indicated by

industry standard guidelines (see Desmaris et al., 2022). Routine monitoring of risk scores and category proportions is also needed following implementation to ensure that the tool is functioning as intended

Recommended next steps

Implementing the updated OST will require Arizona to establish risk classification thresholds, integrate the scoring procedures into existing workflows, and prepare staff to use the results consistently. Therefore, we offer four recommendations. First, Arizona stakeholders should determine how the General and Violent Reconviction scores will be used and establish risk classification thresholds that reflect the desired distribution of individuals across risk categories and the associated implications for supervision, treatment, and eligibility decisions. Second, the updated scoring procedures should be integrated into Arizona's existing assessment workflow or case management system, allowing officers to calculate both scores without collecting additional information. Third, staff should receive training on how to calculate and interpret the scores, with particular emphasis on their distinct purposes and appropriate uses.

Fourth, following implementation, Arizona should routinely monitor risk classifications and compare predicted probabilities with observed reconviction rates. Once sufficient follow-up data have accumulated, the updated OST should be revalidated using newly assessed individuals to determine whether its predictive performance remains consistent over time. Together, these recommendations provide a practical framework for implementing the updated OST while supporting its continued accuracy and usefulness in correctional decision-making.

Limitations

While the localized OST was developed from Arizona's own population and evaluated on individuals who played no role in its development, it is not without limitations. The following

limitations identify where the findings should be interpreted cautiously. They also describe areas requiring additional evidence before widespread operational use.

Development population and intended use. The assessment was developed using the general adult probation population rather than treatment court participants specifically. Treatment court participants are selected for substance use or mental health needs and receive more intensive supervision and services, which may alter the relationships between assessment items and reconviction. Performance among treatment court participants should be examined once enough participants have been assessed and followed.

Conviction-based outcomes. Both outcomes rely on new convictions. Convictions capture only behavior that is detected, charged, and adjudicated, and plea agreements can result in conviction offenses that differ from the underlying behavior. Violent reconviction in particular may undercount violent behavior, and the assessment predicts officially recorded violence rather than all violence.

Race and ethnicity. Combining multiple racial and ethnic populations into a single Non-White category made stable evaluation possible, but it can conceal differences among the populations within that category. Performance across racial and ethnic groups should continue to be evaluated as additional cases and outcomes accumulate following implementation. Future monitoring should prioritize more detailed subgroup estimates when the number of outcome events supports reliable comparisons. Until then, the combined category should be interpreted cautiously.

Informing rather than determining decisions. The assessment estimates reconviction risk using the information available in Arizona's records and does not capture every consideration relevant to an individual case. It is intended to provide a consistent, empirically derived source of

information for treatment court teams and supervision staff, not to replace their judgment. A score should therefore support, rather than determine, the final decision.

These limitations underscore the importance of treating validation as an ongoing process rather than a one-time exercise. By building on Arizona's existing assessment infrastructure and using local outcome data to refine risk estimates, the project provides a stronger empirical foundation for treatment court screening and supervision decisions. Continued monitoring and collaboration with Arizona stakeholders can further improve the models' accuracy, fairness, and practical utility over time.

Conclusion

Arizona developed the OST nearly three decades ago to bring structured, evidence-informed assessment into adult probation practice. Its statewide adoption created a valuable foundation, but locally developed instruments still require periodic revalidation as populations, supervision practices, and intended uses evolve. The initial validation found that the OST generally outperformed the FROST, yet both instruments showed limited ability to predict violent reconviction and substantially overstated recidivism likelihood. The validation results demonstrate why localization improves a tool’s predictive performance and why assessment tools must be periodically reevaluated and refined to support agency.

The localized OST demonstrates the potential value of deriving outcome-specific risk scores from information Arizona already collects. Using Arizona probation data, both the General and Violent Reconviction models achieved strong predictive discrimination, each identifying different combinations of factors relevant to each outcome. The improvement in predictive accuracy for both outcomes support continued development of a two-score approach rather than reliance on a single, general risk measure for decisions involving violence.

Moving forward, Arizona stakeholders should define how the assessments will be used in practice, establish decision thresholds and their implications for case management, and continue collecting assessment and outcome data to support future validation and refinement. With appropriate implementation and continued validation, the localized scores could provide a more consistent empirical basis for screening and supervision decisions while preserving the role of professional judgment and individualized treatment planning. Ultimately, the value of the updated OST lies not simply in more accurate risk prediction, but in helping Arizona make better-informed supervision and treatment decisions.

References

All Rise. (2026). Adult treatment court best practice standards (2nd ed., updated January 2026). https://allrise.org/publications/standards/

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19–52.

Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation*. Public Safety Canada.

Bucklen, K. B., Duwe, G., & Taxman, F. S. (2021). Guidelines for post-sentencing risk assessment. *National Institute of Justice*.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845.

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, Office of Justice Programs, Bureau of Justice Assistance.

Duwe, G. (2014). The development, validity, and reliability of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR). *Criminal Justice Policy Review*, *25*(5), 579-613.

Duwe, G. (2024). Evaluating bias, shrinkage and the home-field advantage: Results from a revalidation of the MnSTARR 2.0. *Corrections*, *9*(1), 20–42.

Ferguson, J. L. (2002). Putting the "what works" research into practice: An organizational perspective. Criminal Justice and Behavior, 29(4), 472–492.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*(4), 575–608.

Hamilton, Z., Kigerl, A., & Kowalski, M. (2022). Prediction is local: The benefits of risk assessment optimization. *Justice Quarterly, 39*(4), 722–744.

Hamilton, Z., Kigerl, A., Allen, B., Ursino, J., & Krushas, A. (2025). Never going to let you down: Preventing predictive shrinkage via the STRONG-R assessment method. *Justice Quarterly*, *42*(7), 1279–1299. https://doi.org/10.1080/07418825.2024.2386637

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Latessa, E. J., Schweitzer, M., Lux, J., Papp, J., McCafferty, J., & Lugo, M. (2019). The OST & FROST revisited: Testing the reliability and validity of the Offender Screening Tool (OST) and the Field Reassessment Offender Screening Tool (FROST). University of Cincinnati Corrections Institute.

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency, 52*(1), 77–93.

National Association of Drug Court Professionals. (2018). Adult drug court best practice standards: Volume I (Rev. ed.).

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.

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0

Ursino, J., & Hamilton, Z. (2026). Validation of the Arizona Offender Screening Tool (OST) and Field Reassessment Offender Screening Tool (FROST) – Research brief (Doc. #2026-002). Advancing Research in Corrections Lab, University of Nebraska Omaha.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35.

Technical Appendix

This appendix documents the statistical specifications and development decisions needed to understand and reproduce the localized OST. The main report describes these procedures in practitioner terms. The specifications should be read alongside the substantive interpretation in the main report. Outstanding verification items remain marked where the source materials do not resolve them.

A1. Statistical Modeling Specifications

Coding and data preparation. Each candidate item was coded so that higher values indicate greater risk. OST items are recorded on a −1 to 1 scale, where 1 indicates a risk factor, 0 its absence, and −1 a protective response. Before estimation, responses of −1 were recoded to missing and, with other missing predictor values, replaced with the item median; in the operational scorecard these responses score zero. Highest grade completed retained its −1 to 2 coding, with −1 indicating a bachelor's or other undergraduate degree. Age at assessment was winsorized and sign-reversed so that younger age corresponds to greater risk. Cases missing the outcome were excluded. Race was retained for evaluation only and excluded from model construction.

Estimation and tuning. Each model was estimated as a ridge (L2 penalized, $\alpha = 0$) logistic regression (Hoerl & Kennard, 1970). The penalty parameter λ was selected by five-fold cross-validation within the training sample over a grid of 200 log-spaced values from 10^{4} to 10^{-4}, using binomial deviance as the criterion. The value minimizing mean cross-validated deviance (λ.min) was selected.

Models for both outcomes were computed using a penalized ridge logistic regression, a statistical approach that estimates how strongly each assessment item is associated with reconviction while reducing the risk of assigning too much importance to individual factors. Unlike the OST's original scoring structure, which assigns predetermined point values to items, this approach uses observed reconviction outcomes to determine how much each factor should contribute to the overall risk estimate. Ridge regression also addresses a common challenge in risk assessments, where many items capture related characteristics and therefore, provide overlapping information. Without accounting for this overlap, a model may assign unusually large weights to certain items or produce weights that change substantially across samples. Ridge regression reduces this problem by applying a penalty that pulls item weights toward zero, limiting the influence of less informative factors while allowing stronger predictors to retain greater influence (Hoerl & Kennard, 1970; Friedman et al., 2010). Importantly, all candidate items remain in the model, but their contributions are adjusted to produce more stable predictions.

The appropriate strength of this penalty was determined using a process called five-fold cross-validation, which evaluates how well a model performs on data that were not used to estimate its weights. The training sample was divided into five groups. The model was developed using four groups and then evaluated on the remaining group to assess how well it predicted reconviction among individuals whose records were not used to estimate that version of the model. This process was repeated five times, with each group serving as the evaluation group once. Different penalty strengths were compared, and the one producing the best average predictive performance across the five evaluations was selected. This approach helps identify a model that performs consistently across different portions of the data rather than one that fits the

training records particularly well but performs poorly on new cases. The separate test sample was then reserved for the final evaluation of the selected model.

Direction constraint. Coefficients from the initial model were inspected, and predictors with negative coefficients were removed. The model was re-estimated with the retained predictors using the same cross-validation procedure, and λ.min was again selected. This procedure was completed separately for the General and Violent Reconviction models.

Evaluation. Final models were applied to the test sample to generate predicted probabilities. AUC was estimated with 95% confidence intervals using the DeLong method (DeLong et al., 1988). Calibration was assessed by grouping predicted probabilities into deciles and comparing mean predicted probability with the observed rate in each decile (Steyerberg, 2019). The ECE was calculated as the mean absolute difference between predicted and observed rates across deciles, weighted by decile size, with a 95% confidence interval from 1,000 bootstrap resamples of the test sample. Classification thresholds for subgroup error rates were identified using Youden's J (Youden, 1950).

A2. Items Without Weight in Either Model

Eleven candidate measures received no weight in either final model and do not appear on the scorecard: OST Vocational item 5, Education items 1 and 2, Alcohol item 3, Attitude items 2, 3, 4, and 7, Criminal Behavior item 3, sex offender status, and the count of prior felony convictions. Information on prior felonies remains represented through OST Criminal Behavior item 6 and the post-conviction designation item. Items scored zero for only one outcome are shown in Table A1. Exclusion from the operational scorecard should not be interpreted as evidence that these measures have no association with reconviction. Rather, these items did not

provide sufficient additional predictive information once the other factors in the models were taken into account.

A3. Complete Point Values

Table A1. Point Values by Item and Outcome

Domain	Item	Response Scored	General	Violent
Current Offense and Status	Current offense is a felony	Yes	16	0
Current Offense and Status	Current offense is undesignated	Yes	4	0
Current Offense and Status	DUI status	Has DUI status	4	0
Current Offense and Status	Domestic violence status	Has DV status	0	12
Criminal Behavior	Q1. Age of first arrest	16 or younger	0	1
Criminal Behavior	Q2. Prior juvenile adjudications and adult convictions	3 or more	8	6
Criminal Behavior	Q4. Present offense represents an increase in seriousness	Yes	5	3
Criminal Behavior	Q5. Post-conviction designation	Felony or undesignated	10	10
Criminal Behavior	Q6. Any previous felony convictions	Yes	3	0
Criminal Behavior	Q7. Present conviction is violence related	Yes	0	2
Criminal Behavior	Q8. At least one previous violence-related conviction	Yes	2	6
Criminal Behavior	Q9. Two or more previous violence-related convictions	Yes	0	3
Criminal History Counts	Prior misdemeanor convictions	Per conviction	3	4
Criminal History Counts	Prior prison sentences	Per sentence	3	0

Domain	Item	Response Scored	General	Violent
Criminal History Counts	Prior violent convictions	Per conviction	2	1
Education	Q3. Reading ability	Below 6th grade level	0	2
Vocational	Q1. Current verified employment	Unemployed	3	0
Vocational	Q2. Unemployed 30+ days three or more times in past 3 years	Yes	6	4
Vocational	Q3. Fired or asked to resign in past 3 years	Yes	0	1
Vocational	Q4. Current financial situation	Unable to meet obligations	3	3
Family / Social	Q1. Raised primarily by a biological parent	No	1	2
Family / Social	Q2. Lived in family home until age 16	No	0	1
Family / Social	Q3. Immediate family member has a criminal record	Yes	0	1
Family / Social	Q4. Current family relations	Negative	1	3
Family / Social	Q5. Childhood history of domestic violence	One or more incidents	1	1
Family / Social	Q6. Current companion / romantic relationship	Problematic or no companion	2	5
Family / Social	Q7. Current peer and associate relationships	Negative / antisocial	5	2
Family / Social	Q8. Associates with a prosocial person regularly	No	2	2
Residence	Q1. Current residential arrangement	Negative and/or antisocial	2	1
Residence	Q2. Stability of residence	Moved two or more times in last 6 months	2	1
Alcohol	Q1. Alcohol-related arrests	2 or more	2	4
Alcohol	Q2. Effect of alcohol use in month before offense	Problem / interferes with life	0	2
Drug	Q1. Illegal drug use in past year	Yes	5	3

Domain	Item	Response Scored	General	Violent
Drug	Q2. Illegal drug use in month before offense	Yes	3	0
Drug	Q3. Self-perception of drug use	Any use / interferes with life	7	0
Mental Health	Q1. History of mental illness	History of illness or treatment	2	3
Mental Health	Q2. Current mental health functioning	Moderate to serious problems	1	1
Health	Q1. Health status	Interferes with probation	1	0
Health	Q2. Health status	Interferes with employment	1	1
Attitude	Q1. Attitudes supportive of crime	Yes	3	3
Attitude	Q5. Poor attitude toward authority figures	Yes	0	2
Attitude	Q6. Motivation to improve life	Poor	1	1
Demographics	Highest grade completed	Less than 9th grade	8	4
Demographics	Highest grade completed	9th through 12th grade	4	2
Demographics	Highest grade completed	HS diploma, GED, some college, or certificate	0	0
Demographics	Highest grade completed	Bachelor's or other undergraduate degree	-4	-2
Demographics	Sex	Male	3	5
Demographics	Age at assessment	See age table in main report	0 to 5	0 to 5

Responses not listed, including the absence of each risk factor, score zero.

A4. Classification Error Rates at Alternative Thresholds

To provide a starting point for examining these errors, initial thresholds were identified using Youden's J, a statistical method that balances the ability to identify reconvictions against correctly identifying individuals who are not reconvicted (Youden, 1950). This produced reference cutoffs of 26% predicted probability for General Reconviction and 4.3% for Violent Reconviction. These thresholds are not recommended classification or eligibility cutoffs. Rather,

they provide a consistent reference point for examining how classification errors vary across groups.

Table A2 presents the Violent Reconviction Youden threshold alongside three more selective alternatives in the test sample. It reports the percentage flagged, the percentage reconvicted among those flagged, and false-positive and false-negative rates for the full sample and separately for men and women. The main report explains why the Youden cutoff produces the observed error rates and why no single cutoff is automatically appropriate for Arizona's agencies. These figures are intended to support local discussion of the consequences of alternative thresholds, rather than establish an operational classification rule.

Table A2. Violent Reconviction Classification Error Rates at Alternative Thresholds, Test Sample

Threshold	Group	Flagged (%)	Reconvicted Among Flagged (%)	False Positive Rate (%)	False Negative Rate (%)
Youden (0.04)	All	25.3	16.9	22.3	20.2
	Men	28.2	16.6	25.0	19.7
	Women	16.2	18.4	13.8	22.7
Top 20% (0.09)	All	20.0	19.0	17.1	29.0
	Men	22.5	18.7	19.4	28.1
	Women	12.2	21.1	10.0	33.2
Probability of 0.10 or higher	All	18.1	19.6	15.4	33.9
	Men	20.4	19.3	17.5	32.5
	Women	10.7	21.5	8.7	40.2
Top 10% (0.16)	All	10.0	23.1	8.1	56.9
	Men	11.5	22.4	9.5	55.7
	Women	5.2	27.9	3.9	62.5

Note. Thresholds were applied to the Violent Reconviction model's predicted probabilities in the test sample (N = 35,789; base rate 11.1%). Values in parentheses are the predicted-probability cutoffs. The top 20% and top 10% thresholds classify the highest-scoring 20% and 10% of the full test sample. Flagged is the share of each group classified above the threshold.

A5. Concentration of Reconviction Across Scores

Examining how reconvictions are distributed across risk scores helps illustrate how effectively the models distinguish individuals with different levels of risk and how much of the observed reconviction is concentrated among those receiving higher scores. In the General Reconviction test sample, the highest-scoring 10% accounted for 18.7% of all reconvictions and had an observed rate of 48.1%; the highest-scoring 20% and 30% accounted for 34.4% and 48.9%, respectively, with observed rates of 44.4% and 42.0%. The lowest-scoring 30% accounted for 11.0% of reconvictions and had a 9.5% observed rate. These results show concentration of events at higher scores. Violent reconviction was considerably more concentrated. The highest-scoring 10% accounted for 43.1% of all violent reconvictions and had an observed rate of 23.1%, over twice as high of the 11.1% observed base rate; the highest-scoring 20% and 30% accounted for 71.0% and 82.6%, respectively, with observed rates of 19.0% and 14.7%. The lowest-scoring 30% accounted for only 4.6% of all violent reconvictions, illustrating the concentration of these outcomes among higher-scoring individuals.