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

Development and Validation of Localized General and Violent Reconviction Scores for the South Dakota Level of Service Inventory-Revised

Prepared for the South Dakota Unified Judicial System, Court Services

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 Level of Service Inventory-Revised and the Violence Risk Appraisal Guide 2

The Importance of Localization and Revalidation 3

Item Weighting 5

Outcome-Specific Scores 6

Project Structure 7

Validation of the Existing VRAG and LSI-R 9

Validation Data and Methods 9

Findings 10

Predictive Accuracy 10

Performance Across Subgroups 11

Calibration 12

Implications for Tool Development 13

Localizing the LSI-R 14

Study Population 14

Information Available for Development 15

Reconviction Outcomes 16

Assessment Items Considered 17

Criminal History, Current Offense, and Prior Convictions 18

Needs and Circumstances 19

Age 19

Predictive Performance 19

Performance of the Localized LSI-R 20

Fairness Assessment 22

How Fairness Was Evaluated 23

Findings by Race 24

Comparison of Updated and Existing LSI-R Score 26

Sources of Improved Predictive Performance 27

Discussion: Implications for South Dakota 30
Using the Scores and Next Steps 31
Implementation and Ongoing Evaluation 33
Recommended Next Steps 33
Limitations 34
Conclusion 36
References 38
Technical Appendix 41
A1. Statistical Modeling Specifications 41
A2. Items Without Weight in Either Model 43
A3. Complete Point Values 43
A4. Classification Error Rates at Alternative Thresholds 45

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.

Currently, South Dakota Court Services uses the Level of Service Inventory-Revised (LSI-R) to assess individuals supervised on probation and has also used the Violence Risk Appraisal Guide (VRAG) to screen for violence risk. This project sought to examine how effectively these instruments predict general and violent reconviction and whether predictive accuracy can be improved by adapting the LSI-R using South Dakota's own assessment and outcome data. The study aim was to strengthen the information available to Court Services and treatment courts while building on assessment practices already established across the state.

This 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. South Dakota was one of three states contributing data to develop state-specific assessment tools, with its participation focused on evaluating and refining instruments used in statewide probation practice and treatment court screening.

Primary Findings

The project first validated the VRAG and the LSI-R as currently scored in South Dakota using assessments of adult men completed by Court Services between March 2021 and March 2025 (Ursino, 2026). The primary findings from that evaluation are summarized below.

- *The VRAG demonstrated negligible accuracy for violent reconviction, while the LSI-R demonstrated borderline moderate accuracy for violent reconviction and weak accuracy for general reconviction.*
- *Both tools overpredicted violent reconviction, with predicted risk exceeding observed recidivism rates by 31% for the VRAG and 50% for the LSI-R on average. Further, the LSI-R overpredict general reconviction by 12%.*

The study findings indicated that neither tool's existing scoring structure provides adequate prediction of violence risk, and both would substantially overstate the likelihood of violent reconviction if applied in treatment court screening. Since the broader LSI-R captured more relevant risk information than the VRAG, its items were retained and used to develop a localized General and Violent Reconviction scores, which were weighted using South Dakota assessment and administrative data. Performance was evaluated and the primary findings are summarized below.

- *The localized General Reconviction tool score demonstrated moderate and the Violent Reconviction score indicated strong predictive accuracy.*
- *Localized scores were well calibrated, with predicted and observed rates differing by 7% for general reconviction and 9% for violent reconviction on average.*
- *Notably, the two tools draw on different information to predict risk. The Violent Reconviction tool is driven by current offenses, dissatisfaction with a marital or equivalent relationship, residential instability, and substance use, while the General*

Reconviction Tool is driven by prior adult convictions, violations during prior supervision, substance use, and a poor attitude toward supervision.

- *Relative fairness was also demonstrated across racial subgroups. The General Reconviction tool demonstrated moderate accuracy for both White and Non-White individuals, and the Violent Reconviction score demonstrated strong accuracy for White individuals and moderate accuracy for Non-White individuals.*

Together, the findings indicate that Alaska can obtain substantially improved violence risk prediction with an updated, locally weighted assessment developed without adding items or changing data collection procedures. Further, because the two tools 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. Overall, the updated tool will provide South Dakota agencies with more accurate prediction of general and violent reconviction risk for all the supervised population.

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 and individuals whose current circumstances 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, from those for whom participation may not be appropriate or responsive to treatment court programming.

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. This consideration is particularly relevant in South Dakota, where both of the instruments used to assess risk were developed in other jurisdictions and adopted for local use.

The Level of Service Inventory-Revised and the Violence Risk Appraisal Guide

The Level of Service Inventory-Revised (LSI-R) was developed in Canada in 1995 using on a sample composed primarily of White, male probationers (Andrews & Bonta, 1995; Vose et al., 2008). The LSI-R combines static criminal history information with dynamic needs that can change over time. In total, the LSI-R contains 54 items across ten subcomponents: Criminal History, Education/Employment, Financial, Family/Marital, Accommodation, Leisure/Recreation, Companions, Alcohol/Drug Problems, Emotional/Personal, and Attitudes/Orientation. Each item contributes one point when the risk factor is present, and the items are summed to produce a total score used to assign a risk level (Andrews & Bonta, 1995). The relative ease of scoring has contributed to its popularity, as the LSI-R is among the most widely used assessments, having been adopted by an estimated 900 correctional agencies across North America and is used in prisons, jails, and community corrections (Vose et al., 2008).

The use of the Violence Risk Appraisal Guide (VRAG) is more limited in scope, as it is often used as a screening instrument to obtain more targeted information about individuals who may pose an elevated risk of violence. The VRAG contains 12 items that rely primarily on

static, unchangeable indicators, which makes it relatively easy to administer (Rice et al., 2013). Like the LSI-R, the VRAG has been widely adopted across correctional populations, despite initially being developed in a clinical setting. The LSI-R and VRAG, therefore, are both widely used assessments that have simplistic scoring schemes and a relatively limited number of items. Thus, it is important for local jurisdictions to evaluate the accuracy of these tools to ensure that they demonstrate accurate prediction within their populations.

The Importance of Localization and Revalidation

Despite their ease of use, differences in the populations, policies, and supervision practices across jurisdictions can affect how accurately these instruments predict recidivism. This phenomenon is referred to as predictive shrinkage, where an assessment demonstrates lower accuracy when applied to populations or settings that differ from those used in its initial development (Desmarais et al., 2022; Hamilton et al., 2022, 2025). Fundamentally, predictive shrinkage occurs when the relationships between assessment items and reoffending risk identified in the original development population do not hold as strongly in other jurisdictions.

Assessment accuracy can also diminish over time, even when the instrument initially performs well in a particular jurisdiction. Similar to the predictive shrinkage that occurs from adopting a tool developed from a different sample, assessment accuracy can diminish as population characteristics and agency context gradually change over time. Consequently, adopting a new instrument requires evaluating its performance using local data, and periodic revalidation is necessary to determine whether its scores continue to provide accurate and useful information for correctional decision-making (Bucklen et al., 2021).

Research examining the LSI-R across jurisdictions reinforces the importance of evaluating its performance in the populations where it is used. A meta-analysis of 128 studies

involving more than 130,000 individuals found that Level of Service instruments demonstrated greater predictive accuracy using samples in Canada, where the tool was developed, than in U.S. samples, suggesting that the instruments' performance may be influenced by the populations and correctional contexts in which they are applied (Olver et al., 2014). Within the U.S., the LSI-R's performance has been found to be inconsistent across populations, assessment practices, and outcomes (Kroner et al., 2020).

The need for local validation becomes more important when assessments are used to inform decisions involving different types of reoffending. Differences in accuracy are particularly consequential for less common, but serious outcomes, such as violence, where relatively low rates of reoffending make it difficult to identify individuals who are likely to commit a violent offense. When assessments are not designed to predict low base-rate outcomes, the instrument may still be able to consistently distinguish between individuals at greater risk of violent reoffending from those at lower risk, while simultaneously overestimating the true likelihood that violent reoffending will occur (Steyerberg, 2019). Consistent with these concerns, the Level of Service instruments generally predict violent recidivism less accurately than general recidivism (Olver et al., 2014) and its ability to predict violent reoffending is inconsistent across samples (Campbell et al., 2009).

These findings also raise questions about the information used to predict violent reoffending and how that information is incorporated into assessment scores. While the VRAG is a violence-specific assessment, it relies on a small set of 12 predictors selected for their associations with violent outcomes. However, prior research has shown that incorporating a wide range of risk factors greatly enhances predicting infrequent outcomes (Hamilton et al., 2024). Moreover, the limited set of items contained in the VRAG make the tool unable to be

locally optimized, as larger ranges of items are needed to identify the strongest predictors for local populations. While still relatively limited, the LSI-R assesses a wider range of risk factors that can be evaluated and weighted specifically for violent recidivism, potentially improving prediction without requiring agencies to collect additional assessment information.

However, remaining limitations in outcome-specific prediction have direct implications for how the LSI-R is used by local and state agencies, particularly when assessment results inform treatment or program eligibility and other decisions involving public safety. Unfortunately, the LSI-R was primarily designed to assess general risk and criminogenic needs, rather than to produce separate estimates of violent reconviction. Under its existing scoring structure, each item contributes one point to the total score. This approach does not account for differences in how strongly individual items predict recidivism.

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 structure is unable to capture the unique predictive value that specific items have for local populations. For treatment courts and other case management decisions, 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.

Item Weighting

Local revalidation and refinement offer a practical means of addressing these limitations while preserving South Dakota's existing assessment infrastructure. Rather than assuming that each LSI-R item should contribute one point to the total score, an empirically

derived scoring structure can assign weights based on observed relationships with recidivism among individuals supervised in South Dakota. 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. Evaluating these locally derived scores alongside the existing LSI-R and VRAG provides a basis for determining whether the available assessment information can more accurately support the state's supervision and treatment court decisions. This approach builds on South Dakota's investment in structured assessment while ensuring that resulting scores better reflect its current correctional population and the outcomes most relevant to agency decision-making.

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 supervision intensity decisions, 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 South Dakota's existing LSI-R 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 South Dakota and Alaska using the LSI-R and Arizona using the Offender Screening Tool (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 South Dakota state-specific assessment. The pooled assessments will be described in a separate report. The next section presents the validation of the existing VRAG and LSI-R. 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 tools, performance across demographic groups, and the translation of the statistical models into a scoring system suitable for agency use.

Validation of the Existing VRAG and LSI-R

Before developing the localized assessment, the research team validated the VRAG and the LSI-R as currently scored in South Dakota (Ursino, 2026). The validation assessed how accurately the VRAG and LSI-R predicted reconviction within three years of assessment, whether accuracy differed by 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 two samples drawn from South Dakota Court Services, both consisting of adult men who completed the VRAG or LSI-R between March 2021 and March 2025. The VRAG sample included 1,469 individuals, and the LSI-R sample included 7,234 cases. Because the samples differ in size and composition, the findings should be interpreted as evidence of each tool's performance within its own sample. The VRAG, a violence-specific instrument, was evaluated for violent reconviction only, while the LSI-R was evaluated for both general and violent reconviction, each within three years of assessment. 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). Because violent reconviction is uncommon, the validation also reported the area under the precision-recall curve (PR-AUC) for that outcome, which is interpreted relative to the base rate of the outcome. A PR-AUC near the base rate indicates little improvement over chance.

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 over or underestimating their actual likelihood of reconviction.

Findings

The following section describes findings from study analyses. As described, we examined the overall predictive performance of the VRAG and LSI-R, but also whether that performance varies across racial 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. For violent reconviction, the VRAG demonstrated negligible accuracy (AUC = 0.54), and the LSI-R reached the threshold of

moderate accuracy (AUC = 0.64). For general reconviction, the LSI-R demonstrated weak accuracy (AUC = 0.57), predicting the broader outcome with less accurately than violent reconviction. The PR-AUCs for violent reconviction were 0.11 for the VRAG and 0.17 for the LSI-R. Thus, the VRAG did not meaningfully distinguish individuals who were later reconvicted of a violent offense from those who were not, and the broader LSI-R performed modestly better for violent reconviction but provided limited information regarding general reconviction. Out of the nine tests conducted, predictive discrimination was identified to be a negligible effect twice, a weak effect five times, and a moderate effect twice.

Table 1. AUCs for the VRAG and LSI-R

Group	VRAG (Violent)	LSI-R (Violent)	LSI-R (General)
Total	0.54	0.64	0.57
White	0.51	0.64	0.56
Non-White	0.63	0.61	0.57

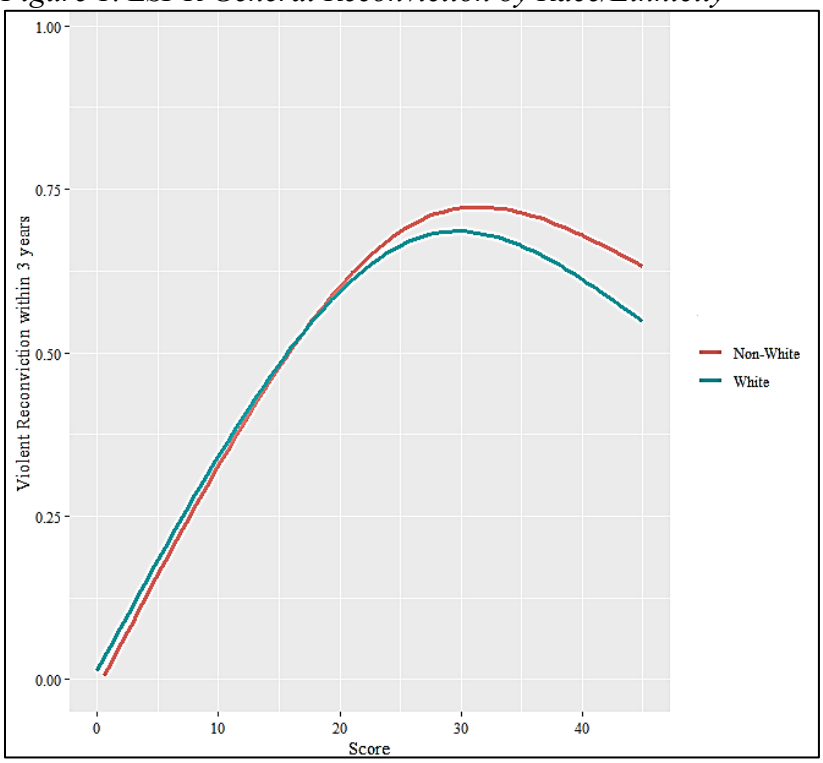
Note: PR AUC for violent reconviction using the total sample were 0.11 and 0.17 for the VRAG and LSI-R respectively.

Performance Across Subgroups

Differences by race were larger for the VRAG than for the LSI-R. The VRAG demonstrated weak accuracy for Non-White individuals (AUC = 0.63) and accuracy slightly better than chance for White individuals (AUC = 0.51). For violent reconviction, the LSI-R demonstrated moderate accuracy for White individuals (AUC = 0.64) and weak accuracy for Non-White individuals (AUC = 0.61), a smaller gap in the opposite direction. For general reconviction, the LSI-R demonstrated weak and nearly identical accuracy for White (AUC = 0.56) and Non-White (AUC = 0.57) individuals.

Figure 1 illustrates the relationship between LSI-R scores and general reconviction for White and Non-White individuals. The weak and nearly identical AUCs for the two groups indicate that higher LSI-R scores were associated with only modest increases in the likelihood of general reconviction, and that this relationship was similar across racial groups.

Figure 1. LSI-R General Reconviction by Race/Ethnicity



Calibration

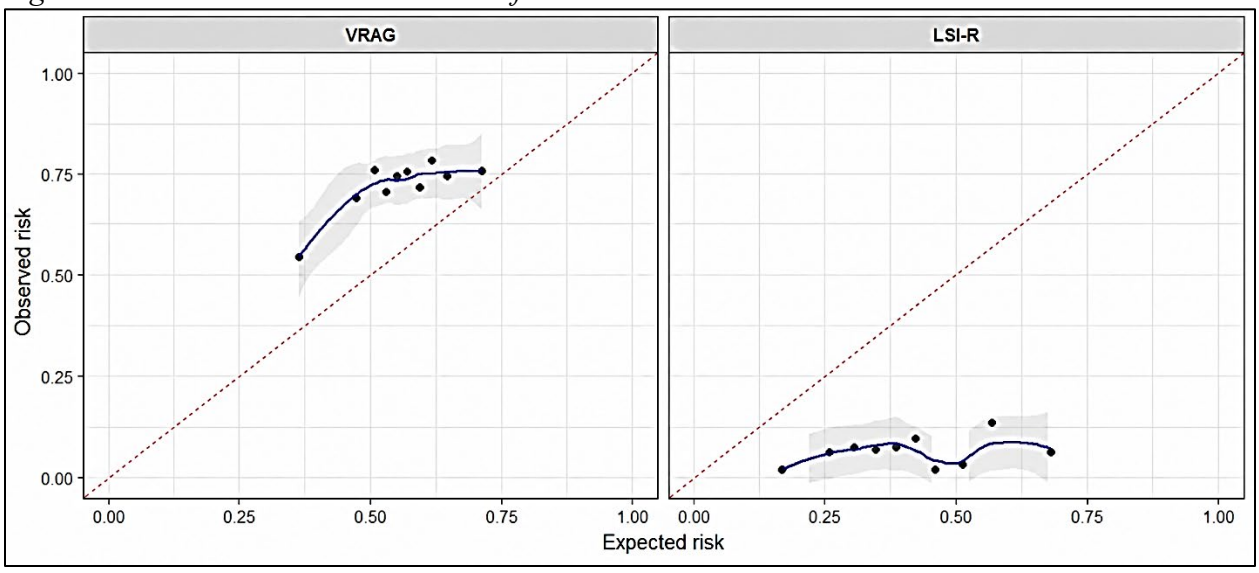
Table 2 displays ECEs trends, showing that both tools' substantially overpredicted violent reconviction, with predicted risk exceeding observed rates by about 31 percentage points for the VRAG and 50 percentage points for the LSI-R. Calibration was considerably better for the LSI-R's prediction of general reconviction, where predicted and observed rates differed by about 12 percentage points on average. Even so, errors of this size, and particularly the overprediction of violent reconviction, may lead to inappropriate supervision and treatment decisions with tool scores providing inflated estimates of risk.

Table 2. ECEs for the VRAG and LSI-R

Tool	ECE
VRAG	0.31
LSI-R (Violent)	0.50
LSI-R (General)	0.12

Figure 2 presents the calibration results for both tools for violent reconviction. The diagonal dotted line represents perfect calibration, where predicted probabilities match observed reconviction rates. For the LSI-R, the calibration trend remains below the diagonal, indicating substantial overprediction of violent reconviction. Conversely, VRAG underpredicted risk. Among individuals assigned a LSI-R-predicted probability of 0.50, for example, only about 10% were reconvicted of a violent offense. Consequently, individuals may be assigned to higher risk categories than their observed likelihood of violent reconviction would warrant, potentially affecting supervision intensity and treatment decisions.

Figure 2. VRAG and LSI-R Calibration for Violent Reconviction



Implications for Tool Development

The evaluation found that neither instrument demonstrated sufficient predictive performance to support decisions specifically concerning violent recidivism risk. The VRAG's negligible discrimination and overprediction indicated that it should not serve as the primary

basis for supervision placement or classification decisions. The LSI-R ranked individuals more accurately for violent reconviction, suggesting that a broader risk-needs assessment captures more locally relevant information about violence than a short violence-specific screener, but its predicted probabilities substantially exceeded observed violent reconviction rates. Its existing total score also provided only weak discrimination for general reconviction.

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 LSI-R 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 LSI-R

Because the initial validation established the need for outcome-specific scoring, the next phase used South Dakota 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 LSI-R was developed using records for 7,234 adult men assessed with the LSI-R by South Dakota Court Services between March 2021 and March 2025. 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. 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.

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 South Dakota Study Population

Characteristic	Training Sample	Test Sample	Total
Individuals (N)	5,788	1,446	7,234
White (%)	52.4	52.4	52.4
Non-White (%)	47.6	47.6	47.6
General Reconviction within 3 years (%)	56.7	56.9	56.7
Violent Reconviction within 3 years (%)	18.0	17.6	17.9

Information Available for Development

The project did not begin with a new list of risk factors. Instead, we evaluated information South Dakota already collects through the LSI-R and its supervision records. Candidate measures came from three existing sources:

- *LSI-R assessment items covering criminal history, education and employment, finances, family and marital relationships, accommodation, leisure and recreation, companions, alcohol and drug problems, emotional and personal functioning, and attitudes.*
- *Current offense information, including whether the current offense is violent and the number of present offenses.*
- *Criminal history and demographic information, including counts of prior felony, misdemeanor, and violent convictions and age at assessment.*

Most LSI-R items record whether a risk factor is present or absent. Several items are rated by the assessor on a graded scale, and for South Dakota these were retained in graded form rather than collapsed to present or absent, so that each step toward greater need adds points. Participation and performance at school or work and peer interactions at school or work use four levels. Dissatisfaction with a marital or equivalent situation, unsatisfactory accommodation, current alcohol problems, and current drug problems use three levels. The number of present offenses was grouped into five ranges: 0 to 4, 5 to 8, 9 to 12, 13 to 16, and 17 or more. Age was entered as years below 42, so that each year younger than 42 adds risk and ages 42 and older contribute nothing. Cases missing an outcome 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. The same General and Violent Reconviction outcomes used to evaluate the LSI-R's initial performance were used to develop the new tools. Among individuals with complete

three-year follow-up, 56.7% experienced General Reconviction and 17.9% experienced Violent Reconviction.

Assessment Items Considered

Sixty-one candidate measures were evaluated for each outcome, including LSI-R items and measures of current offenses, criminal history, and age. 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, an individual with nine present offenses also meets the LSI-R criterion of three or more present offenses and is likely to have prior adult convictions. Although each measure may be associated with reconviction, they often capture similar information. We used penalized 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 47 measures that received weight in at least one model, by area, and the number that carry points in each model.

Table 4. Scorecard Items by Area

Area	Number of Items	Retained General	Retained Violent	Content
Age and Prior Record	4	4	3	Age; prior felony, misdemeanor, and violent convictions
Criminal History	9	6	7	Current violent offense, number of present offenses, prior adult convictions, early arrest, escape, institutional misconduct, supervision violations, record of violence
Education and Employment	8	6	7	Reliance on a pension or assistance, employment history, educational attainment, school

				suspension, performance and peer interactions at school or work
Family and Marital	3	3	1	Marital, parental, and other family relationships
Accommodation	3	2	3	Satisfaction with accommodation, address changes, high-crime neighborhood
Leisure and Recreation	2	2	1	Participation in organized activities and use of leisure time
Companions	3	3	1	Criminal acquaintances and friends, anti-criminal friends
Alcohol and Drug Problems	7	5	6	Lifetime and current substance problems and their effects on law violations, family, school or work, and health
Emotional and Personal	5	3	3	Mental health interference, treatment history, psychological assessment
Attitudes and Orientation	3	3	1	Attitudes toward crime, convention, and supervision
Total	**47**	**37**	**33**	

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 an area with many retained items may contribute less to the final score than an area 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 considered.

Criminal History, Current Offense, and Prior Convictions

Criminal history is consistently among the strongest predictors of future offending (Gendreau et al., 1996). The localized LSI-R draws on this information in several forms: the LSI-R Criminal History items, which record the extent and timing of prior offending and prior responses to supervision and incarceration; current offense information, including whether the offense is violent and the number of present offenses; and counts of prior felony, misdemeanor, and violent convictions. These measures' content overlap, and the modeling approach described in the next section outlines the predictive value each provides.

Needs and Circumstances

The Education and Employment, Family and Marital, Accommodation, Leisure and Recreation, Companions, Alcohol and Drug Problems, Emotional and Personal, and Attitudes and Orientation areas 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

Age is among the most consistently observed predictors of offending, with risk generally declining across adulthood. Age cannot be changed through programming and is included as it provides useful predictive information.

Predictive Performance

Once the General and Violent Reconviction models were finalized using the training sample, they were applied to the test sample. Performance was evaluated using discrimination and calibration, with results compared against the existing LSI-R 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 the 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 LSI-R

Table 5 summarizes discrimination and calibration for the localized version of South Dakota's tool using the test sample. General Reconviction achieved an AUC of 0.68, within the moderate range, and Violent Reconviction achieved an AUC of 0.71, a strong effect size. The Violent Reconviction score ranked individuals more accurately than the General Reconviction score despite modeling a less common outcome. The corresponding ECEs were 0.09 for violent reconviction and 0.07 for general reconviction.

Table 5. Test Sample Performance

Outcome	AUC	ECE
General Reconviction	0.68	0.07
Violent Reconviction	0.71	0.09

Although the models demonstrated moderate-to-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. Predicted

probabilities were well-aligned with observed rates, although modest deviations occurred across the distribution. The largest discrepancy occurred near the top of score range, where the predicted probability was 18.9% compared with an observed reconviction rate of 14.9%, a 5% difference. The highest-risk group also exhibited some overprediction, with a predicted probability of 45.1% compared with an observed rate of 41.4%.

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

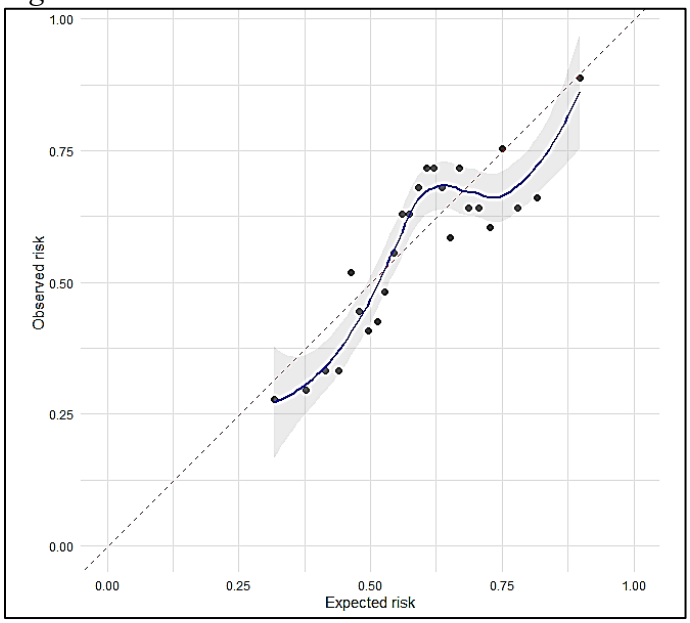
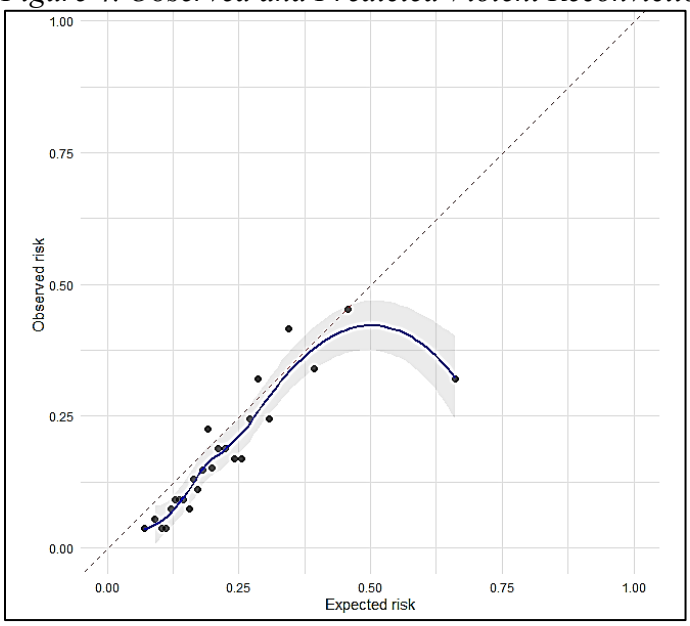


Figure 4 presents the corresponding results for Violent Reconviction. Despite this outcome's lower base rate, predicted probabilities were well-aligned with observed rates, although some deviations remained at the upper end of the risk distribution. The largest discrepancy occurred in the ninth decile, where the predicted probability was 26.4% compared with an observed reconviction rate of 33.8%. The highest-risk decile exhibited modest overprediction, with a predicted probability of 41.4% compared with an observed rate of 37.6%. This is in part due to the small number of individuals in test sample scoring at the high-end of the risk scale.

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



Fairness Assessment

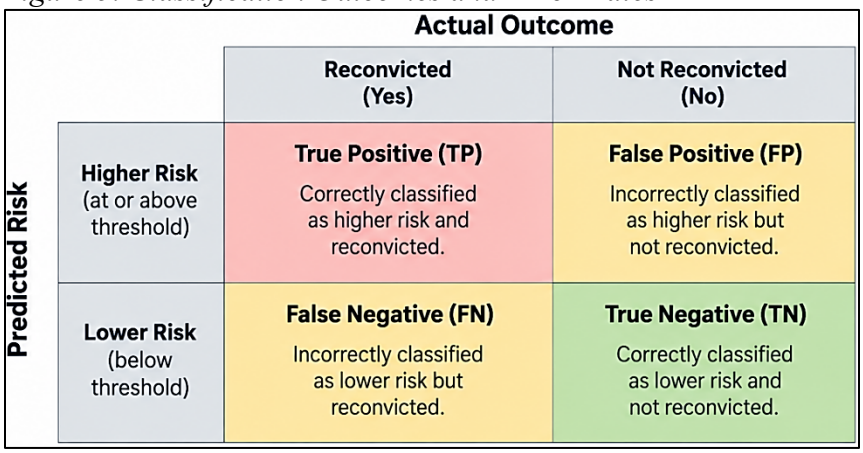
An assessment can perform well for the overall population while performing differently for particular groups. For that reason, the localized LSI-R was also evaluated separately by race. The purpose of this analysis was not simply to determine whether groups received different average scores. Groups can 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 tool fairness. Because the South Dakota data included only adult men, sex could serve neither as a predictor nor as a basis for subgroup comparison. 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 racial 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 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 0.55 predicted probability for General Reconviction and 0.14 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. 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 Race

Differences in predictive accuracy by race were modest. The General Reconviction tool demonstrated moderate accuracy for both White (AUC = 0.68) and Non-White (AUC = 0.67) individuals, a difference that is modest and too small to suggest a meaningful gap. The Violent Reconviction score demonstrated strong accuracy for White individuals (AUC = 0.72) and moderate accuracy for Non-White individuals (AUC = 0.69). This 3% gap is also modest. However, future evaluations should further monitor group differences to ensure fairness remains

Observed reconviction rates were higher for Non-White individuals for both general (59.3% versus 44.1%) and violent (19.7% versus 15.4%) reconviction, and classification error rates at the reference thresholds differed across groups. For General Reconviction, Non-White individuals had both a higher false-positive rate (46% versus 34%) and a higher false-negative rate (39% versus 31%) than White individuals. For Violent Reconviction, false-negative rates were identical (45% for both groups), while White individuals had a higher false-positive rate (29% versus 20%).

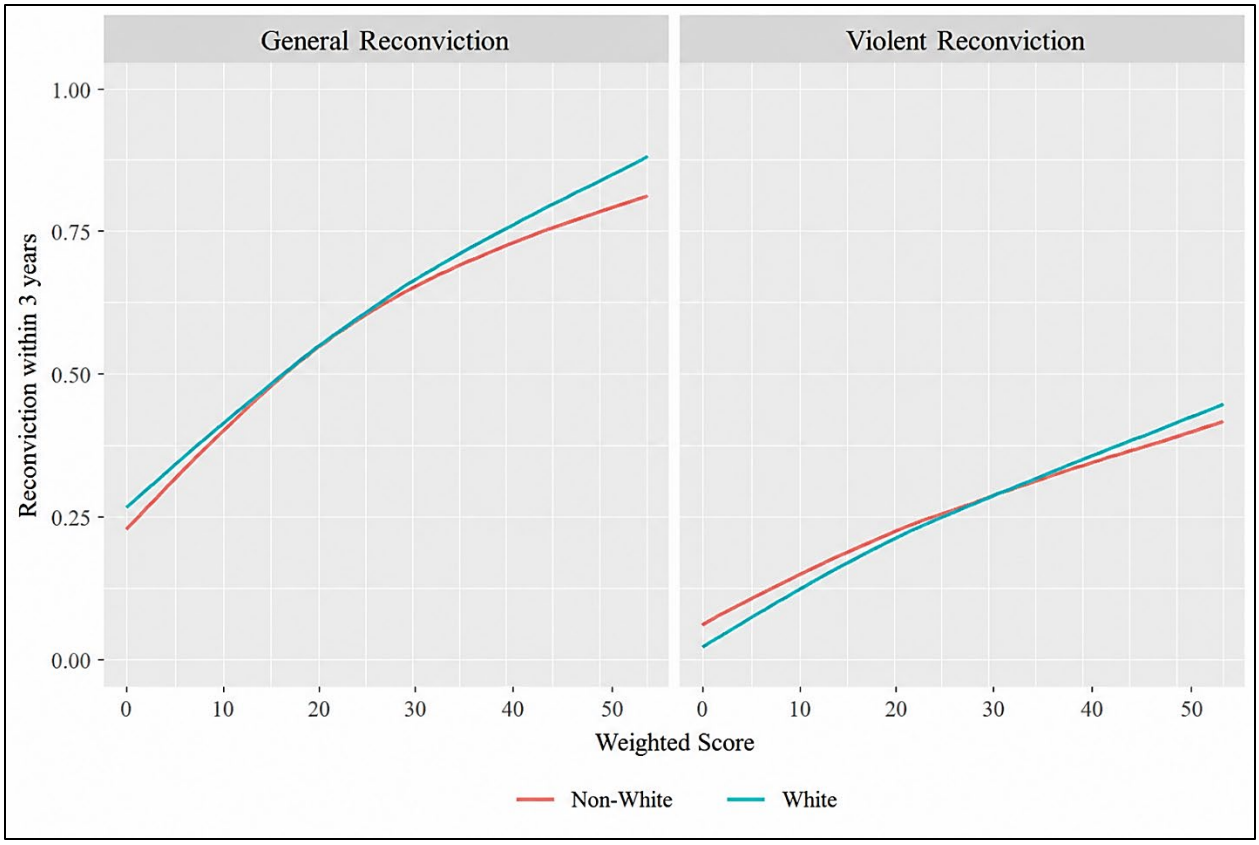
Table 6. Performance by Race, Test Sample

Outcome	Group	N	Reconviction (%)	AUC	FPR	FNR
General	White	758	44.1	0.68	0.34	0.31
General	Non-White	688	59.3	0.67	0.46	0.39
Violent	White	758	15.4	0.72	0.29	0.45

Violent	Non-White	688	19.7	0.69	0.20	0.45

Figure 6 illustrates the relationship between weighted scores and predicted reconviction probabilities across racial categories. For General Reconviction (left panel), the trendlines are closely aligned through the lower and middle score ranges and are nearly identical between scores of about 15 and 30. Some separation emerges at higher scores, where White individuals receive higher predicted probabilities at equivalent scores, reaching approximately 88% compared with 81% for Non-White individuals at the highest scores shown. For Violent Reconviction (right panel), Non-White individuals receive slightly higher predicted probabilities at the lowest scores (about 6% compared with 2% at a score of zero), the curves cross near a score of 30, and White individuals receive slightly higher estimates at the upper end (about 45% compared with 42%). Together, the panels demonstrate broadly consistent relationships between weighted scores and estimated risk across racial categories, with differences of about 7% or less concentrated at the extremes of the score range, where fewer individuals score and probability estimates are less stable.

Figure 6. Predicted Reconviction Probabilities by Weighted Score and Race/Ethnicity



Overall, the models demonstrated comparable ability to distinguish individuals by risk across racial categories, and both groups fell within the moderate-to-strong range on both outcomes. Differences in classification errors at a common threshold can be more pronounced than differences in discrimination, which highlights the importance of examining classification performance across groups rather than relying solely on overall predictive accuracy when evaluating the instruments.

Comparison of Updated and Existing LSI-R Score

The existing LSI-R 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 validation (Tables 1 and 2) used all available LSI-R assessments. For a direct comparison, Table 7 evaluates the existing and localized LSI-R using the same test sample for both outcomes. The localized models improved both predictive accuracy and calibration. For General Reconviction, AUC values increased from 0.57 to 0.67, moving from weak to moderate accuracy, while ECE decreased from 0.12 to 0.07. For Violent Reconviction, AUC increased from 0.64 to 0.71, moving from moderate to strong accuracy, while ECE decreased from 0.50 to 0.09. Improvements in calibration are particularly important for violent reconviction, as the existing LSI-R overstated violent reconviction by about 50 percentage points on average in the validation. Continued use of the existing scoring would therefore substantially overestimate individuals' likelihood of violent reoffending, potentially resulting in unnecessarily restrictive supervision and treatment decisions.

Table 7. Comparison of Localized and Existing LSI-R Scores

Outcome	Original AUC	Localized AUC	Original ECE	Localized ECE
General Reconviction	0.57	0.68	0.12	0.07
Violent Reconviction	0.64	0.71	0.50	0.09

Sources of Improved Predictive Performance

The final models show which items are the strongest predictors for both outcomes. Age and the number of present offenses contribute to both scores, but the remaining measures differ by outcome. Age adds up to 24 points to each score for the youngest individuals, and the number of present offenses adds up to 36 General and 92 Violent points, making it the largest single contributor to the Violent Reconviction score. These patterns provide context for understanding why separate outcome-specific models were retained.

For General Reconviction, criminal history and supervision history carry heavy weight, led by any prior adult conviction and a charge or suspension during prior community supervision. Substance use that contributed to law violations is the largest single needs item,

followed by a poor attitude toward supervision or sentence, reliance on a pension or public assistance, severe interference from mental health problems, and current unemployment. Criminal acquaintances and friends contribute to General Reconviction but not to violence. Their weights describe predictive associations after accounting for other retained items.

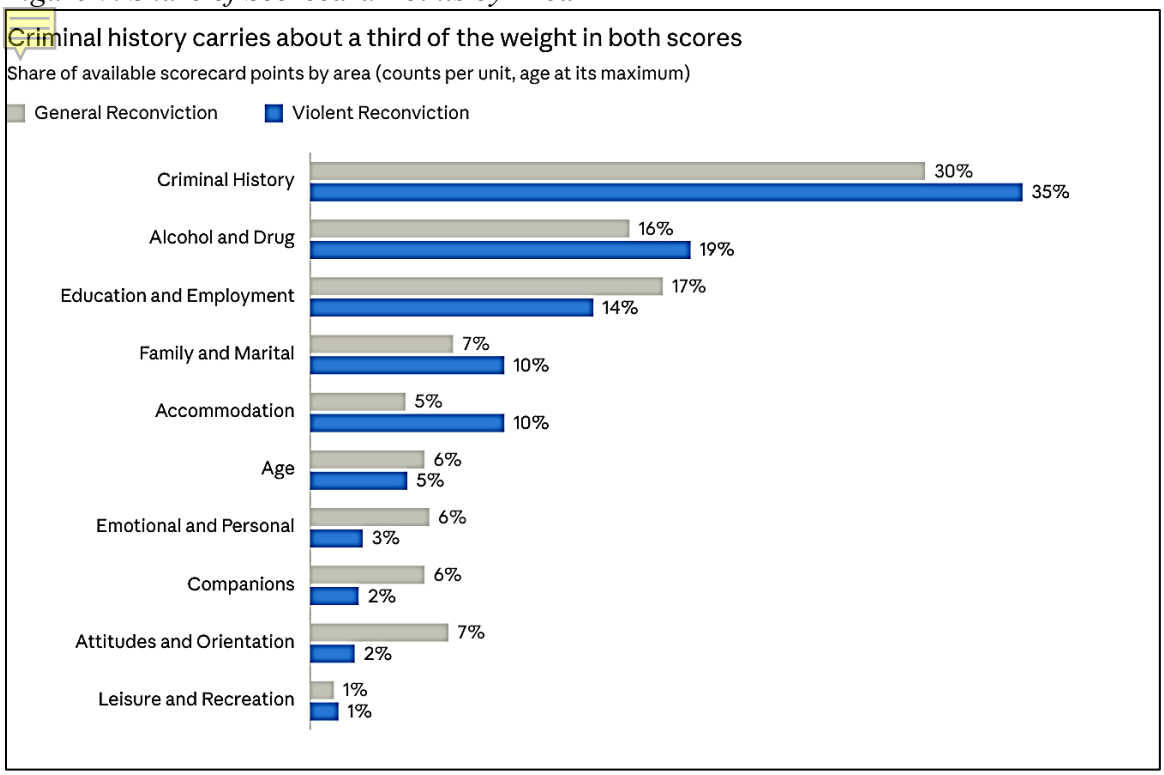
For Violent Reconviction, relationship and residential instability contribute more than they do to general reconviction. Dissatisfaction with a marital or equivalent situation can add up to 48 points, and three or more address changes in the last year adds 27. Substance use that interfered with school or work, contributed to law violations, or contributed to family problems also carries substantial weight. Prior violence contributes through an official record of assault or violence, a current violent offense, and 5 points for each prior violent conviction. Severe mental health interference and escape history, which add substantially to the General Reconviction score, carry no points for the violence scale.

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 areas highlight the value of developing separate scoring models for general and violent reconviction rather than relying on a single measure of risk. Figure 7 illustrates the relative contribution of each area to the General and Violent Reconviction scores, measured as its share of the points available on the scorecard. Criminal history carries the greatest weight in both models, accounting for about a third of the available points. Education and employment factors and attitudes carry greater weight in the

general model, whereas family and marital relationships, accommodation, and alcohol and drug problems contribute more prominently to violent reconviction.

Figure 7. Share of Scorecard Points by Area



A charge or suspension during prior community supervision contributes 28 points to General Reconviction and 6 to Violent Reconviction, while an official record of assault or violence contributes 23 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 South Dakota

The findings support a practical strategy for improving correctional assessment: using locally observed outcomes to reassess how existing information is combined. South Dakota currently collects extensive information through the LSI-R 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, and a short violence-specific screener such as the VRAG may not capture enough locally relevant information to fill that gap. 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 LSI-R 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 South Dakota 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 South Dakota 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 South Dakota treatment court eligibility decisions, the resulting models have broader applications. Using information already collected statewide, the research team developed updated LSI-R models predicting both general and violent reconviction. Given their improved predictive performance relative to the existing LSI-R and the VRAG, we recommend that South Dakota consider adopting the updated models statewide to improve risk estimation across the probation population while also supporting treatment court eligibility decisions.

Regarding next steps, we suggest the courts and probation agencies determine which populations and decisions the updated models will best be used and establish appropriate risk classification thresholds. South Dakota 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 LSI-R will require South Dakota 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, South Dakota 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 Court Services' 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, South Dakota should routinely monitor risk classifications and compare predicted probabilities with observed reconviction rates. Once sufficient follow-up data have accumulated, the updated LSI-R 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 LSI-R while supporting its continued accuracy and usefulness in correctional decision-making.

Limitations

While the localized LSI-R was developed from South Dakota'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 adult men assessed by South Dakota Court Services 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.

Adult men only. Because the South Dakota data included only adult men, the assessment's performance for women is unknown. The scores should not be applied to women until their performance has been evaluated in a sample that includes them.

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, including American Indian individuals. 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 South Dakota'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 South Dakota'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 South Dakota stakeholders can further improve the models' accuracy, fairness, and practical utility over time.

Conclusion

South Dakota's 2013 Public Safety Improvement Act committed the state to evidence-based probation supervision guided by validated risk and needs assessment. That commitment created a valuable foundation in the expanded use of the LSI-R, but instruments adopted from other jurisdictions require local validation and periodic revalidation as populations, supervision practices, and intended uses evolve. The initial validation found that the VRAG provided negligible accuracy for violent reconviction and that the LSI-R, although more accurate for violence, substantially overstated its likelihood and provided only weak accuracy for general reconviction. The validation results demonstrate why localization is not simply a matter of adopting a widely used assessment. Assessment must be periodically reevaluated using recent data and refined to support the decisions it is intended to inform.

The localized LSI-R demonstrates the potential value of deriving outcome-specific risk scores from information South Dakota already collects. The General and Violent Reconviction scores represent substantial improvement from the current tools. The improvement in predictive

accuracy supports continued development of a two-score approach rather than reliance on a single general risk measure or a short violence screener for decisions involving violence.

Moving forward, South Dakota 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 LSI-R lies not simply in more accurate risk prediction, but in helping South Dakota 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. (1995). *The Level of Service Inventory-Revised*. Multi-Health Systems.

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.

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*(6), 567–590.

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.

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., Ursino, J., & Kigerl, A. (2024). Advancing pretrial assessments: Development of the Personal Recognizance Interview & Needs Screen (PRINS). *Journal of Criminal Justice, 92*, 102183.

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.

Kroner, D. G., Morrison, M. M., & Lowder, E. M. (2020). A principled approach to the construction of risk assessment categories: The Council of State Governments Justice Center Five-Level System. International Journal of Offender Therapy and Comparative Criminology, 64(10–11).

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.).

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment, 26*(1), 156–176.

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.

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the VRAG and SORAG: The Violence Risk Appraisal Guide–Revised (VRAG-R). *Psychological Assessment, 25*(3), 951–965.

S.B. 70, 88th Legislative Assembly (S.D. 2013). An Act to improve public safety. https://mylrc.sdlegislature.gov/api/Documents/Bill/24139.pdf?Year=2013

South Dakota Unified Judicial System. (2025). *Drug court handbook* (Beadle County). https://ujs.sd.gov/media/d3udun0b/beadle-county-drug-court-application.pdf

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. (2026). *Validating the South Dakota Violence Risk Appraisal Guide (VRAG) – Research brief* (Doc. #2026-001). Advancing Research in Corrections Lab, University of Nebraska Omaha.

Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation, 72*(3), 22–29.

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 LSI-R. The main report describes these procedures in practitioner terms. The specifications should be read alongside the substantive interpretation in the main report.

A1. Statistical Modeling Specifications

Coding and data preparation. Each candidate item was coded so that higher values indicate greater risk. Dichotomous LSI-R items were coded 1 when the risk factor was present and 0 otherwise. Graded rated items were retained in graded form and reverse-coded where needed so that higher levels indicate greater need: participation and performance and peer interactions at school or work (0 to 3), and dissatisfaction with a marital or equivalent situation, unsatisfactory accommodation, current alcohol problems, and current drug problems (0 to 2). Age was entered as max(0, 42 − age). The number of present offenses was entered as an ordinal rank from 0 (0 to 4 offenses) to 4 (17 or more). Prior felony, misdemeanor, and violent convictions were entered as counts. Cases missing the outcome were excluded, and missing predictor values were replaced with the item median. 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, 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 LSI-R's original scoring structure, which assigns one point to each item, 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

Items scored zero for only one outcome are shown in Table A1. Exclusion from the operational scorecard means 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

Area	Item	Response Scored	General	Violent
Age and Prior Record	Age	Per year younger than 42	1	1
Age and Prior Record	Prior felony convictions	Per conviction	1	0

Age and Prior Record	Prior misdemeanor convictions	Per conviction	1	1
Age and Prior Record	Prior violent convictions	Per conviction	1	5
Criminal History	Current offense is violent	Yes	0	16
Criminal History	Any prior adult convictions	Yes	28	13
Criminal History	Three or more present offenses	Yes	19	19
Criminal History	Number of present offenses	5 to 8	9	23
Criminal History	Number of present offenses	9 to 12	18	46
Criminal History	Number of present offenses	13 to 16	27	69
Criminal History	Number of present offenses	17 or more	36	92
Criminal History	Arrested under age 16	Yes	3	0
Criminal History	Escape history from a correctional facility	Yes	12	0
Criminal History	Ever punished for institutional misconduct	Yes	0	1
Criminal History	Charge laid or probation/parole suspended	Yes	28	6
Criminal History	Official record of assault or violence	Yes	0	23
Education and Employment	Reliant on pension	Yes	21	18
Education and Employment	Currently unemployed	Yes	19	9
Education and Employment	Frequently unemployed	Yes	14	0
Education and Employment	Never employed for a full year	Yes	0	1
Education and Employment	Less than regular grade 10	Yes	11	10
Education and Employment	Suspended or expelled at least once	Yes	0	2
Education and Employment	Participation / performance	Per level (0 to 3)	1	5
Education and Employment	Peer interactions	Per level (0 to 3)	2	5
Family and Marital	Dissatisfaction with marital or equivalent situation	Per level (0 to 2)	14	24
Family and Marital	Non-rewarding parental relationships	Yes	1	0
Family and Marital	Non-rewarding relationships with other relatives	Yes	1	0
Accommodation	Unsatisfactory accommodation	Per level (0 to 2)	0	6

Accommodation	Three or more address changes in last year	Yes	15	27
Accommodation	High crime neighborhood	Yes	5	9
Leisure and Recreation	Absence of recent participation in organized activity	Yes	4	7
Leisure and Recreation	Could make better use of time	Yes	1	0
Companions	Some criminal acquaintances	Yes	9	0
Companions	Some criminal friends	Yes	5	0
Companions	Few anti-criminal friends	Yes	10	12
Alcohol and Drug Problems	Drug problem, ever	Yes	10	16
Alcohol and Drug Problems	Alcohol problem, currently	Per level (0 to 2)	0	3
Alcohol and Drug Problems	Drug problem, currently	Per level (0 to 2)	4	0
Alcohol and Drug Problems	Substance use contributed to law violation	Yes	30	25
Alcohol and Drug Problems	Substance use contributed to family or marital issues	Yes	9	19
Alcohol and Drug Problems	Substance use interfered with school or work	Yes	10	27
Alcohol and Drug Problems	Substance use interferes with health	Yes	0	1
Emotional and Personal	Mental health moderately interferes with life	Yes	0	3
Emotional and Personal	Mental health severely interferes with life	Yes	21	0
Emotional and Personal	Previously received mental health treatment	Yes	1	0
Emotional and Personal	Currently receiving mental health treatment	Yes	0	2
Emotional and Personal	Psychological assessment indicated	Yes	3	8
Attitudes and Orientation	Supportive of crime	Yes	1	0
Attitudes and Orientation	Unfavorable attitude toward convention	Yes	1	0
Attitudes and Orientation	Poor attitude toward supervision or sentence	Yes	27	11

A4. Classification Error Rates at Alternative Thresholds

To provide a starting point for examining classification 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 55% predicted probability for General Reconviction and 14% 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 three alternative thresholds for the Violent Reconviction score in the test sample: classifying the highest-scoring 20% of individuals, classifying everyone with a predicted probability of 0.10 or higher, and classifying the highest-scoring 10%. It reports the percentage flagged, the percentage reconvicted among those flagged, and false-positive and false-negative rates separately for White and Non-White individuals. The thresholds illustrate the trade-off between the two types of error. The probability threshold of 0.10 flags most individuals in both groups and misses few violent reconvictions, but at the cost of false-positive rates above 70%. The top 10% threshold reduces false-positive rates below 10% but misses most individuals who were later reconvicted of a violent offense. At each threshold, a larger share of Non-White individuals was flagged and false-positive rates were higher for Non-White individuals, while precision among those flagged was higher for White individuals at the top 20% and top 10% thresholds. No single cutoff is automatically appropriate for South Dakota'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 (%)
Top 20% (0.23)	White	17.0	38.6	12.4	57.4
	Non-White	23.2	32.6	19.4	61.6

Probability of 0.10 or higher	White	75.2	18.8	72.1	8.2
	Non-White	82.0	23.0	78.6	4.3
Top 10% (0.31)	White	8.5	42.1	5.8	76.7
	Non-White	11.6	32.4	9.8	80.9

Note. Thresholds were applied to the Violent Reconviction model's predicted probabilities in the test sample (N = 1,446; base rate 17.6%). 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.