Validating the Minnesota Level of Supervision/Case Management Inventory (LS/CMI): Interim Report

April, 2026

Prepared for:

The Minnesota Department of Corrections (DOC)

Completed by:

The Advancing Research in Corrections (ARC) Lab

School of Criminology & Criminal Justice

University of Nebraska – Omaha

Authors:

Zachary Hamilton, Ph.D.

Professor and ARC Lab Director

John Ursino, MA

& Yujin Kim, MA

Table of Contents

EXECUTIVE SUMMARY

The Level of Service/Case Management Inventory (LSCMI) has been used to assess risk and need of the Minnesota community corrections population since the early 2000's. During this time, the tool has been used by County and State agency officials to identify individuals' level of recidivism risk, identify those most in need of programming, and direct supervision in the community. However, sufficient information regarding its functionality and effectiveness in the state of Minnesota was lacking. In 2025, the Minnesota Department of Corrections contracted with the Advancing Research in Corrections (ARC) Lab at the University of Nebraska – Omaha to evaluate the use of the LS/CMI in Minnesota. The proposal outlined four project deliverables that examine the tool's 1) usability, 2) reliability, 3) validity in predicting recidivism, and 4) functionality.

This study provides findings from the four outlined objectives. Using a mixed methods approach, we gathered both qualitative and quantitative information on the tool's use and functionality. Regarding quantitative aims, study analyses linked LS/CMI assessments completed between January 1, 2012, and March 31, 2022, and reconviction outcomes. Demographics and case characteristics were also included to measure the LS/CMI's performance with regard to discrimination, calibration, and equity. The analytic sample included 131,899 assessments, where outcome base rates were observed to be roughly 34% for general and 7% for violent reconvictions at three-years following assessment administration.

The qualitative component consisted of five web-based focus groups with managers and line staff from forty-three agencies. Participants described how the tool is taught, scored, documented, and integrated into supervision across a decentralized system that includes Community Corrections Act agencies, the Department of Corrections, and county probation offices. The focus groups provided context for interpreting statistical results and helped explain how local practices and conditions impact scoring and classification. Study findings trace predictive validity at the item, domain, score, and risk level category levels. Predictive bias was also examined across sex and race/ethnicity groups using iterative logistic regression models. Finally, we identified how closely the tool's prediction of risk matches observed recidivism rates and further tested whether recalibration of the tool can reduce observed error.

Key Findings

Usability of the tool as described by staff

- Focus group participants value the LS/CMI as a common language for risk and needs but also describe wide differences across counties regarding training depth, scoring interpretation, cut points, use of trailer tools, and documentation practices.
- Trailer tools are routinely provided to fill in LS/CMI prediction gaps and often govern classification overrides.
    - Risk classification is further complicated for intrastate transfers, as receiving counties often do not share the same instruments or risk-level thresholds.
- There was staff concern that sections of LS/CMI are ambiguous and, in turn, not scored reliably.
- Case-load size and structure vary significantly across county agencies.
- Some felt that all counties should be required to use the same, standardizing risk-level category scoring ranges.
    - However, many staff were strongly against adopting a singular set of ranges, describing concerns regarding resource constraints.

Inter-Rater Reliability (IRR)

- IRR testing indicates *excellent agreement* for total LS/CMI scores and items.
    - These findings suggest that staff are well-trained in delivering the tool consistently and not the source of potential performance issues.
    - Needs-based items produce more variability in IRR scoring compared to criminal history items but remain well within the bounds of strong reliability.

Predictive Power, Bias, Classification, and Calibration

- Regarding prediction, the LS/CMI demonstrated *moderate-to-weak* performance.
- Bias tests were statistically significant, yet modest when examining the tool's prediction by sex and by race/ethnicity.
- Domain-specific findings indicate that criminal history and antisocial patterns carry the strongest predictive power.
    - Many of the LS/CMI needs items add limited predictive value.

- Calibration analyses indicated that the LS/CMI consistently overpredicts risk. These findings indicate that the tool provides higher rankings of risk than warranted.

Conclusion and Recommendations

Together, study findings indicate that Minnesota's use of the LS/CMI provides an *adequate* prediction of recidivism risk; however, noted inconsistencies across counties and tool limitations reduce its effectiveness statewide. Although recalibration of cut points and probabilities may resolve some concerns, we found consistent issues related to usability, bias, and standardization. Accordingly, we recommend the following actions to strengthen risk assessment practices in Minnesota:

Recommendations

- **Pursue targeted modifications to the LS/CMI** to increase predictive value and/or reduce performance inequities.
    - Consider weighting or removing items based on predictive strength and equity.
- **Evaluate alternative risk assessment tools** that offer greater flexibility, improved calibration, with built-in gender responsive and violence specific models.
    - Notably, more modern assessment tools have combined "multi-band" models of prediction that would reduce the need for trailer tools.
- **Strengthen statewide standardization and consistency** by establishing uniform guidance for scoring, documentation, and risk category thresholds across counties.
    - However, strengthening standardization should also consider county resource constraints and population characteristics.
    - Standardization should also include clear expectations for narrative justification within the S^3 system to support interpretability during intrastate transfers.
- **Enhance training, proficiency, and quality assurance** efforts through regular interrater reliability checks, mandatory refresher trainings, and verification that all agencies are using current manuals and scoring guidance.
    - Continued proficiency testing and standardized training helps to ensure that the high levels of agreement observed are maintained in routine practice.

- **Reduce reliance on trailer assessment tools or standardize how they are used**, particularly where local policies default to the most restrictive outcome, contributing to inconsistent supervision decisions across counties.
- **Explore development of a Minnesota-specific risk assessment tool** using existing administrative data.
    - A locally developed tool could improve predictive accuracy, reduce bias, better reflect the probation population, and generate long-term cost savings compared to proprietary instruments.

In sum, the LS/CMI provides an acceptable foundation for risk assessment in Minnesota, and current findings suggest agencies have worked thoughtfully within the tool's notable limitations. While adequate for continued use based on national standards (CSG, 2022), its current design and implementation constrain consistency and equity at the state level. Given agencies' continued support for evidence-based practices, we strongly recommend Minnesota consider updates to its assessment tool, training, and/or standardization that could better support the quality of practice already in place and allow risk classification to more effectively inform supervision and resource allocation.

INTRODUCTION

Risk and needs assessment (RNA) tools have become essential instruments in modern correctional practice, fundamentally transforming how criminal justice agencies supervise individuals and allocate treatment resources. The development of the Risk-Need-Responsivity (RNR) model in the 1980s revolutionized corrections by establishing that effective intervention requires three key components: assessing an individual's likelihood of reoffending (risk), identifying changeable factors that contribute to criminal behavior (needs), and addressing individual barriers to successful intervention (responsivity; Andrews & Bonta, 2010). Incorporating the RNR model into supervision helped to move corrections away from subjective professional judgment and toward structured, evidence-based assessment practices designed to improve outcomes while promoting fairness and consistency in decision-making.

Today, nearly every correctional agency in the United States employs some form of RNA tool to guide supervision intensity, program placement, and case management decisions (Singh et al., 2018). These instruments serve dual critical purposes: first, they provide a standardized framework for evaluating recidivism risk relative to the broader supervised population, ensuring consistent practices across geographic regions. Second, they inform the development of individualized supervision plans, including contact frequency and targeted programming to address criminogenic needs (Desmarais et al., 2018).

However, the adoption and implementation of RNA tools present significant challenges. Research has documented the phenomenon of 'predictive shrinkage,' where assessment tools demonstrate reduced accuracy when applied to populations different from those used in their original development (Hamilton et al., 2025). Factors contributing to shrinkage include demographic differences, variations in jurisdictional statutes, differing law enforcement priorities, and inconsistent implementation practices. Additionally, growing evidence suggests that many contemporary RNA tools exhibit bias, systematically overclassifying women and racial/ethnic minorities into higher risk categories than their actual recidivism rates warrant (Hamilton et al., 2022; Skeem & Lowenkamp, 2016). These concerns have prompted increased scrutiny of assessment practices and calls for regular validation studies to ensure tools function equitably and accurately at the local level.

Background

The Level of Service/Case Management Inventory (LS/CMI) represents one of the most widely adopted RNA tools in North America. Developed in the Canadian provinces of Ontario and Manitoba, the LS/CMI evolved from its predecessors, the Level of Supervision Inventory (LSI) and the Level of Service Inventory-Revised (LSI-R), to provide a comprehensive assessment instrument that combines risk prediction with case management guidance (Andrews et al., 2004). The LS/CMI employs a dichotomous scoring system (0=not present, 1=present) administered through structured interviews, supplemented by information from correctional databases, family members, and employers. With a scoring range of 0 to 43, the LS/CMI reflects consolidation and refinement of its 54-item predecessor, the LSI-R (Bonta & Wormith, 2018). This contemporary version distinguishes itself by incorporating additional risk/need factors, assessments of incarceration experiences, and additional considerations of responsivity and protective factors. A case management section integrates risk, need, and responsivity factors to formulate programming and treatment recommendations (Andrews et al., 2004; Bonta & Wormith, 2018).

Research on the LS/CMI has generally demonstrated moderate predictive validity across diverse settings. However, a meta-analysis by Olver and colleagues (2014) examined thirty years of research on the LSI-R and LS/CMI, finding evidence of predictive shrinkage, where the tool demonstrated moderate-to-strong predictive validity in the Canadian providences in which it was developed, but slightly weaker accuracy when generalized to the rest of Canada, and substantially weaker prediction in U.S. samples (Olver et al., 2014). Moreover, the same research identified concerning patterns of differential prediction across demographic groups, with some studies documenting weaker performance for racial and ethnic minorities and females (Olver et al., 2014).

LS/CMI Implementation in Minnesota

A key aspect of assessment provision is standardization, or the ability to provide similar assessment and interventions across a jurisdiction. However, Minnesota's correctional landscape presents unique complexities that distinguish it from many other jurisdictions. The state operates three distinct probation systems across 87 counties: Community Corrections Act (CCA) agencies serving approximately 71% of the adult probation population, the Minnesota Department of

Corrections (DOC) supervising 18%, and County Probation Offices (CPO) handling non-felony cases comprising 11% of supervised individuals. This county-based system results in unique local standards for supervision and programming, leading to a varied landscape of practices across the region. To provide some historical context, Minnesota initially adopted the LSI-R in the early 2000's before transitioning to the LS/CMI in 2012. Historically, there has been no statewide policy governing risk assessment use; practices have varied considerably across the state's county-based systems. In 2013, the Minnesota Department of Corrections (MNDOC) introduced Division Directive 203.016, mandating consistent use of a risk/needs tool within DOC-supervised probation to determine risk levels and intervention targets. This directive requires that individuals convicted of felonies undergo LS/CMI assessment within 90 days of probation assignment. There is also variation in the use of pre-screening instruments. The DOC previously used the Wisconsin Risk/Needs Assessment to establish initial risk levels, while other agencies adopted different tools. Still others, such as Hennepin County, developed their own pre-screener.

Despite this policy framework, comprehensive analyses of the LS/CMI's performance in Minnesota remain limited. However, prior evaluations have revealed that LSI-R's needs items demonstrated weak predictive accuracy in a Minnesota sample (Duwe & Rocque, 2016), and more recent studies revealed that, among previously incarcerated individuals, the LS/CMI similarly exhibited weak predictive accuracy for violent, non-violent, and felony recidivism (Duwe, 2024).

Previous Validation Efforts and Identified Concerns

In 2012, the Hennepin County Department of Community Corrections and Rehabilitation compared the performance of the LSI-R to the LS/CMI prior to the statewide transition. Researchers analyzed roughly 10,000 LSI-R assessments administered solely in Hennepin County between 2009 and 2012. Researchers found that the LS/CMI performed with moderate predictive accuracy. However, overclassification of females was apparent, with High-Risk females significantly less likely to recidivate than High-Risk males. Similar disparities emerged between Black individuals scored as Moderate and High-Risk compared to White individuals within the same risk levels (Wildermuth et al., 2021).

These early concerns were amplified by a comprehensive 2023 report from the Council of State Governments (CSG) examining probation and supervision services in Minnesota. The report highlighted several operational challenges warranting attention:

1. **Inconsistent Implementation**: LS/CMI assessments were conducted for 70% of felony probationers from 2018 to 2020, but only 32% for gross misdemeanors and 18% for misdemeanors, with considerable county-to-county variation.
2. **Lack of Formalized Case Management**: No standardized process existed for translating assessment results into case management plans across jurisdictions.
3. **Supplemental Assessment Proliferation**: Some counties felt compelled to augment LS/CMI assessments with additional tools such as the Domestic Violence Inventory (DVI), STABLE, and the Women's Risk-Need Assessment (WRNA), suggesting perceived inadequacies in the LS/CMI's coverage.
4. **Data Standardization Problems**: Lack of uniformity in information collection sent to the Statewide Supervision System (S^3), along with varying operational definitions, has hindered the standardization of assessment findings and efficient utilization of results/scoring.
5. **Racial/Ethnic Disparities**: High rates of revocation among Black and American Indian individuals on probation persisted even after accounting for relevant offense factors, suggesting that documented predictive variability in the LS/CMI for these populations (Olver et al., 2014) may also be present in Minnesota (CSG, 2023).

Based on these findings, CSG recommended either adopting a single tool for standardized use across the state, validated on Minnesota samples, or agreeing upon supplemental screeners alongside the LS/CMI, with mandatory validation and re-validation every five years (CSG, 2023). The convergence of limited prior validation research, documented concerns regarding predictive accuracy and bias, and Minnesota's uniquely decentralized correctional structure creates a need for comprehensive evaluation of the LS/CMI's performance in this context. To provide the outlined evaluation, the Minnesota Department of Corrections (MnDOC) contracted with the University of Nebraska's Advancing Research in Corrections (ARC) Lab, outlining deliverables that extend beyond simple predictive validity testing, to encompass multiple

dimensions of assessment quality, including reliability, usability, functionality, and equity across demographic groups.

Scope of Current Project

This evaluation represents a comprehensive examination of the LS/CMI assessment system as implemented across Minnesota's adult probation population. The project seeks to address the gaps identified in prior research while providing actionable recommendations to enhance the usability, reliability, accuracy, and functionality of risk and needs assessment practices in Minnesota. The evaluation encompasses multiple objectives that provide an understanding of the LS/CMI's performance in Minnesota's correctional context.

1. First, we assessed the LS/CMI's predictive validity for the Minnesota probation population, examining overall discrimination, calibration, gender and racial/ethnic bias.
2. Second, we assessed the usability of the LS/CMI by evaluating current training practices and the consistent use of the tool across Minnesota's probation systems. Through interviews with key stakeholders, we sought to identify practical barriers that impact accurate and efficient assessment administration, evaluate the adequacy of training and ongoing support systems, and understand how contextual factors, such as caseload pressures and access to collateral information sources, affect assessment quality.
3. Based on these findings, the project provides evidence-based recommendations to improve not only the LS/CMI instrument itself, but also the broader system of training, quality assurance, and implementation support necessary for effective risk and needs assessment practices.
4. Finally, the project sought to enhance the functionality of the LS/CMI by examining its intended use for the probation population and developing recommendations to improve how assessment results are translated into case management actions, program referrals, and supervision planning decisions. This functional assessment recognizes that even technically valid and reliable assessment tools may fail to improve correctional outcomes if their results are not effectively integrated into practice.

LS/CMI VALIDATION METHODOLOGY

The current evaluation employed a comprehensive mixed-methods approach to assess the LS/CMI's performance across four critical dimensions: usability, reliability, predictive validity, and functionality. The quantitative portions of the study use administrative data provided by the Minnesota Department of Corrections and included LS/CMI item-level responses, domain scores, total scores, and risk level categories for all electronically available assessments completed between January 1, 2012, and March 31, 2022. These assessment records were linked to demographic information (gender, race/ethnicity, age), case characteristics (offense type, supervision type, jurisdiction), and recidivism outcomes defined as any new conviction within three years of assessment. Qualitative data was obtained through semi-structured interviews with LS/CMI users and agency administrators representing diverse jurisdictions, probation systems, experience levels, and roles. Reliability information was collected via an internal proficiency exercise completed by MNDOC staff as part of their ongoing efforts to enhance reliability training.

Usability

The usability assessment of Minnesota's LS/CMI takes on heightened importance given the state's decentralized supervision structure. The usability of risk-needs assessment tools shapes their effectiveness in correctional practice. *Usability* encompasses the practical realities of assessment implementation: a) the clarity of item wording and scoring guidelines, b) feasibility of completing assessments within existing time and resource constraints, c) adequacy of training to ensure consistent interpretation, and d) degree to which assessment results provide actionable information that meaningfully informs supervision and programming decisions.

This usability assessment employed qualitative methods to capture experiences and perspectives of those directly involved in LS/CMI administration and use. By systematically examining both line-level implementation challenges and system-level organizational factors, this component of the evaluation provides essential context for interpreting the quantitative validity and reliability findings, while generating actionable recommendations for improving assessment design, training, and integration into probation practice.

Conducting Focus Groups

Qualitative evaluation data were collected through semi-structured focus groups with Minnesota probation personnel. In total, we conducted five sessions via Zoom, including three groups with county probation agency managers and two with line-level staff who regularly score the LS/CMI. We ensured the sample included both experienced assessors and newer users to gain a representative sample of those working in Minnesota's probation system. In total, representatives from 43 different agencies attended the sessions. Sessions typically lasted 60 minutes, were recorded with verbal consent, and transcribed.

The interview protocol was developed by UNO to align questions with both research objectives and agency priorities. It covered assessment administration workflows, time requirements, competing demands, information availability, and recurring item-specific scoring challenges. It also examined training and support resources, including initial and booster trainings, scoring guidance, and quality assurance processes. In addition, the discussion addressed how risk levels and domain scores are integrated into supervision intensity, case planning, and program referrals. Finally, we explored variation across jurisdictions in implementation practices, supplemental assessments, available resources, and local policies, as well as participants' views on the LS/CMI's strengths, limitations, alignment with clinical judgment, and potential improvements specific to Minnesota practice. Open-ended questions allowed participants to describe their experiences in their own words, and targeted probes were used when participants raised themes of particular relevance, such as whether item-scoring difficulties reflected ambiguous wording, limited information, or cultural considerations.

Qualitative analysis followed thematic analysis procedures (Braun & Clarke, 2006). Researchers first read all transcripts to build familiarity and note initial observations. Next, each line of the transcript was read and coded by researchers to capture participants' experiences, perspectives, and suggestions. Related codes were grouped into candidate themes, which were later refined. Representative quotations are included in the report to illustrate the range of perspectives. The usability findings provided essential context for interpreting quantitative validity and functionality results and generate practical recommendations for assessment design, training, and workflow integration.

Inter-Rater Reliability

Inter-rater reliability (IRR) refers to the extent to which different assessors assign consistent scores when evaluating the same individual using the LS/CMI. While predictive validity evaluates whether scores are associated with future recidivism, reliability evaluates whether those scores are produced consistently across users. In decentralized systems such as Minnesota's probation structure, reliability is especially important because variation in scoring practices can produce artificial differences in classification that are unrelated to a probationer's risk.

To examine reliability, we analyzed a dedicated interrater reliability dataset consisting of LS/CMI assessments completed independently by multiple assessors on the same cases. These data were generated through a statewide proficiency test administered by the MNDOC between January 20 and 30, 2026. Assessors completed a standardized scenario consisting of a recorded interview and written supplemental materials developed and vetted by LSCMI master trainers. The assessment was administered through the Statewide Supervision System (S^3) Quality Assurance platform and required approximately 2.5 hours to complete. To preserve independence of scoring, assessors were instructed to complete the exercise individually without discussion or consultation, and scoring guidance was withheld until after the testing period concluded.[1]

Reliability for all forty-three LS/CMI items was examined, along with the total score. Interrater agreement was estimated using linear weighted kappa statistics, which quantify the extent to which two assessors provide the same ratings after accounting for agreement that would be expected by chance. Kappa values range from negative values to 1.00. A value of 1.00 indicates perfect agreement, a value of 0 indicates agreement no better than chance. Consistent with common interpretive conventions, kappa values below 0 are considered poor, 0 to 0.20 slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and above 0.80 excellent agreement (Landis & Kock, 1977). These analyses provide an empirical assessment of scoring consistency across assessors and serve as an important complement to both the usability and predictive validity findings presented in this report.

[1] Assessors could potentially view prior scores in the system if instructions were not strictly followed. While this posed a possible risk to scoring independence, it was deemed comparable to informal score sharing, and similar limitations were unavoidable given the absence of proctored testing.

Predictive Validity

Predictive validity represents a risk-need assessment's ability to accurately forecast which individuals will reoffend. A predictively valid tool enables correctional agencies to allocate intensive supervision and treatment resources to those at highest risk of reoffending while reducing unnecessary intervention for lower-risk individuals, improving public safety outcomes and promoting efficient resource utilization. However, validity is often context dependent. A tool developed and validated in one jurisdiction, with a particular population, may demonstrate substantially different performance when applied in another setting (Hamilton et al., 2022).

This study evaluated the LS/CMI's performance across multiple outcome definitions, follow-up periods, and levels of analysis, including items, domains, total scores, and risk categories. Using administrative data linking LS/CMI scores to general and violent recidivism, predictive validity was assessed across these dimensions and stratified by gender, race/ethnicity, and case type to examine differential prediction. Multiple metrics were used to capture distinct aspects of validity and variation across administration contexts. In addition, logistic regression models were used to test for bias, including whether the strength of prediction differed across groups and whether certain groups were systematically over- or under-classified. Together, these analyses provide a comprehensive assessment of the LS/CMI's accuracy and equity, and whether modifications may be needed to better fit Minnesota's probation population.

Predictive accuracy is typically measured via the discrimination statistic – the Area Under the Curve (AUC), which quantifies a tool's ability to assign a higher risk score to individuals who reoffend. Ultimately, the AUC represents the *probability* that someone who reoffends will be scored as higher risk than someone who does not, with values ranging from 0 to 1. A value of 0.5 is representative of random chance (i.e., predictions of recidivism are no better than random chance), with lower values suggesting that the tool performs worse than randomly assigning risk scores. AUC values are commonly reported in effect size ranges, where values below 0.55 are considered 'negligible', 0.56 to 0.63 are considered 'small', 0.64 to 0.70 are 'moderate', and values 0.71 and above are 'strong' (Rice & Harris, 2005).

Functionality

Functionality addresses how well the tool works in practice once implemented. An assessment tool's functionality encompasses how effectively it serves its intended purposes within the correctional system's operational context. Even a tool that demonstrates moderate predictive validity may function poorly if its risk level categories are mis-calibrated to the population distribution or supervision resources. This concept relates to the use of the risk categories in practice, where domains (or treatment needs) may fail to correspond with available services and treatment programs, or item weighting may fail to account for differential predictive strength among items.

The central questions are whether a) risk levels align with appropriate supervision intensity and available programs, b) domain scores meaningfully guide needs identification and referral, c) risk categories are calibrated to Minnesota's population distribution and resource constraints, and d) the current scoring approach makes efficient use of item-level predictive information for this population. These questions are especially important in Minnesota for four reasons. First, the LS/CMI was developed using Canadian samples and has not been locally calibrated to Minnesota's probation population. Second, prior evaluations in Minnesota have found inconsistent links between assessment results and service provision. Third, most counties rely on supplemental assessments to inform override decisions, particularly in DUI and sex offense cases. Fourth, earlier evaluations in Minnesota have identified possible overclassification in specific demographic groups.

The functionality assessment proceeded at item, domain, and risk-category levels. Item-level and domain-level associations with recidivism were evaluated using univariate and multivariate logistic regression to identify components that contribute most strongly to prediction and to test for differential functioning across demographic groups. Risk-category performance was examined by describing population distributions and comparing recidivism rates across adjacent categories. Alternative cut-point configurations were simulated to evaluate whether different thresholds would create greater distinction in reoffending risk between risk level categories and, in turn, improve resource allocation.

First, we assess discrimination by reporting AUCs for each domain, and we identify potentially problematic items within each domain. Specifically, we identify items with AUC values less than 0.50 (i.e., worse than chance prediction). We also report false positive rates

(FPRs), which identify items that are frequently scored as a '1' for individuals who ultimately do not reoffend. Notably, FPRs are key indicators of overclassification, and prior evidence suggests that this may be a particular concern for females and non-White individuals. Next, we describe how cases are distributed across risk level categories using the standard LS/CMI cut points and summarize predictive accuracy. Further, we examine the utility of risk-level categories by calculating odds ratios (ORs). Notably, odds ratios (ORs) indicate the magnitude of the association between risk classification and recidivism. With the Low-Risk/Very Low-Risk group used as the reference category, an OR above 1.00 indicates that a higher risk group has greater odds of recidivating. For instance, an OR of 1.50 indicates that the higher risk group has 50% greater odds of recidivism than the Low-Risk/Very Low-Risk group. Additionally, ORs represent effect sizes where 1 to 1.4 are negligible, 1.5 to 2.4 are small, 2.5 to 4.2 are moderate and 4.3 or greater are considered large effects.

Even when the tool adequately distinguishes between levels of risk, mis-calibrated categories or weak connections between domain scores and available programming can limit its practical usefulness for case planning, resource allocation, and decision-making. Calibration analyses compare predicted and observed recidivism rates overall, where increases in risk scores should be associated with an equitable increase in recidivism rates. Overall score performance was summarized via the calibration Root Mean Squared Error (RMSE), which represents the distance between predicted probabilities and observed outcome rates. Smaller values indicate better performance, and a value of zero indicates perfect agreement between LS/CMI predicted and observed rates. We also assess calibration using the Maximum Calibration Error (MCE), which indicates the maximum error observed at any risk score level, identifying the worst-performing segment of the risk assessment tool.

Taken together, this mixed-methods design provides a comprehensive assessment of the LS/CMI as it is currently implemented in Minnesota, combining practitioner insight with rigorous quantitative evaluation across usability, reliability, predictive validity, and functionality domains of performance. Qualitative findings offer important context for understanding how the tool is applied in practice, reflecting a high level of engagement and thoughtful use among practitioners navigating a complex and decentralized system. At the same time, the quantitative analyses establish a clear baseline of performance, including evidence of predictive accuracy and variation across key dimensions of use. By integrating these approaches, the evaluation is

designed not only to assess how the LS/CMI is functioning, but to also identify opportunities where refinements to the tool and its implementation could better support the level of practice already demonstrated across Minnesota's probation system.

FINDINGS

In completing the evaluation, we initially conducted interviews with key staff and stakeholders to gain insight into how the LS/CMI fits the needs of Minnesota's local agencies. We used *identified areas of concern* described by administrators and prior evaluations to guide our investigation of the tool's functionality and formulate recommendations. In the following sections, we first provide descriptive information on Minnesota's probation population. Next, we highlight key themes that emerged from the stakeholder and staff interviews. We then provide findings regarding the predictive validity of the LS/CMI in Minnesota and examine differences among an array of outcomes, as well as initial and reassessments. Finally, we provide more detail as to the functionality of the tool by providing insight into how items and domains within the assessment function, as well as describe the tool's calibration, as it relates to assigning individuals to risk level categories.

Descriptive statistics of the study sample are provided in Table 1, where the sample is composed of mostly male (77.1%) and White (62.8%) individuals. Moreover, drug offenses account for most of the sample's current convictions, closely followed by violent offenses (27.3% and 24.6% respectively). The average LS/CMI score is 20.5 out of a possible maximum score of 43. Finally, 33.8% of the sample received a new conviction within three years of their initial LS/CMI assessment, and only 7.3% received a new violent conviction.

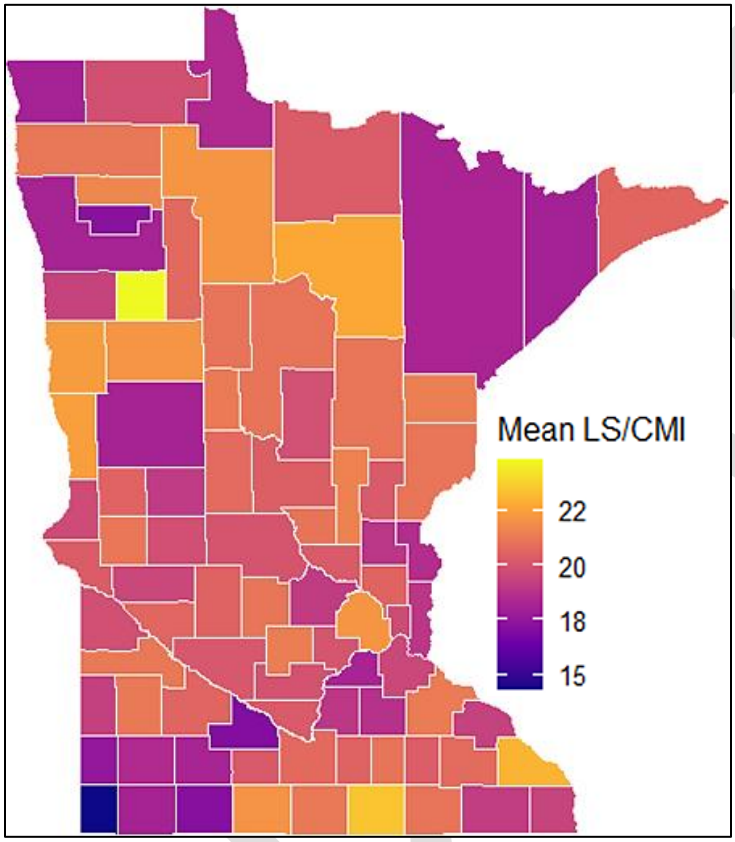
Table 1. Descriptive Statistics (N=131,899)

Descriptive	%/Mean
Male	77.1%
Race/Ethnicity	
White	62.8%
Black	19.7%
American Indian	7.7%
Hispanic	5.2%
Asian	2.1%
Other/Unknown	2.5%
Current Offense	
Violent	24.6%
Drug	27.3%
DWI	16.5%
Property	12.5%
Other	19.1%
LS/CMI Scores	
Total LS/CMI Score	20.5
Criminal History Score	4.7
Needs Score	16.1
Outcomes	
Reconviction - 3 yr.	33.8%
Violent Reconviction - 3 yr.	7.3%

Figure 1 provides a visual (heat map) of how risk scores are distributed across Minnesota's counties, with higher/lower average LS/CMI score by county indicated by color, where lighter colors mean higher average scores. This visualization allows for a quick comparison of how average assessed risk varies across different regions of the state. Several regional patterns are visible, where counties in the southwest portion of Minnesota generally present lower average risk scores, as indicated by the darker purple shading. Similarly, several counties in the northeast and parts of the northwestern region of the state also display comparatively lower average scores. In contrast, counties surrounding the Twin Cities metropolitan area fall in the middle of the distribution, with moderate risk levels that typically

range from approximately 18 to 21 on the LS/CMI scale. Higher average scores appear more frequently in counties located in the central southern and northern regions. These counties are represented by lighter orange and yellow shades. Overall, the map highlights meaningful geographic variation in the average LS/CMI risk levels and may reflect differences in local probation populations, case mix, supervision practices, and the level of supervision and service resources likely needed across jurisdictions.

Figure 1. Average LS/CMI Score by County



Usability Findings

Five focus groups were conducted with Minnesota probation professionals to examine current practices around the Level of Service/Case Management Inventory (LS/CMI). Participants included both line staff (users) who administer and score the tool daily, and administrators who oversee implementation, quality assurance, and policy decisions. The sessions revealed that, while the LS/CMI is valued as a common language for risk assessment and case planning, significant variability exists across counties in training quality, scoring practices, cut points, and how risk classifications map to supervision resources.

Training and Proficiency Are Uneven

Training practices vary considerably across Minnesota counties, particularly following the COVID-19 pandemic. Annual refresher training is common, with some counties offering semiannual or domain-focused sessions. However, participants described wide variation in the depth and utility of these boosters. Many participants found boosters useful, others characterized them as repetitive and failing to address key issues regarding ambiguity in scoring specific items. A common characteristic of trainings deemed as “helpful” or “very helpful” was a group discussion regarding why tool administrators chose to score each item with a given response. This helped to clarify ambiguity and standardize interpretation of some items.

Smaller and rural agencies report greater reliance on state-provided trainers for initial and ongoing education. Others attend training courses hosted in nearby counties. A key gap identified was the absence of proficiency testing, as statewide assessments that once ensured scoring consistency were discontinued *before* the COVID-19 pandemic and have not been reinstated. One administrator estimated it took approximately three years to bring their large county to roughly 90% proficiency, underscoring the extended timeline required to achieve scoring reliability, as well as agency managers’ strong emphasis on reaching high levels of reliability within their agencies. Tool users echoed these sentiments, with one expressing *"refreshers are hugely helpful*" while others expressed worry about remaining inconsistencies in scoring between agents trained 10 years ago and those trained within the last few years.

Ambiguous Items Create Scoring Challenges

Scorers consistently identified specific LS/CMI items that generate confusion and inconsistent interpretation. The companions/peers domain emerged as particularly problematic, with users struggling with item redundancy, the one-year lookback period, and how to apply scoring criteria within Minnesota-specific contexts. One issue regularly discussed by participants was how to score questions regarding drug use. Following the legalization of THC in Minnesota, some agents chose to not include THC when scoring drug-related questions, and others only count use when there are specific court orders regarding its restriction. It was common for tool users to report settling on when/how to score these items informally within their local agencies, with one participant stating:

> *"There may be inconsistent scoring across counties, but here, we go to each other when we are unsure, and by doing that we kind of 'standardize' how questions are scored and that makes things pretty consistent."*

That informal peer consultation is a useful practice as it helps staff resolve ambiguous items quickly and builds shared interpretations that improve standardization within agencies. However, it is unlikely to improve cross-agency standardization.

Negatively phrased wording in some LS/CMI items was repeatedly flagged as a source of scoring confusion, especially in the criminal history domain, with multiple participants highlighting the question regarding "No engagement in pro-social activities". It also came to light in one of the focus groups that some agencies exclude the extant offense when scoring "ever incarcerated upon arrest" while others do not. Participants also highlighted employment-related edge cases that do not fit cleanly into the tool's categories, including how to value part-time work and how to score per-capita payments received on reservations. Items that probed probationers' time spent in "structured" versus "unstructured" activities were similarly described as ambiguous among agents in smaller counties, where formal and structured socialization was generally less common relative to informal or unstructured socialization. Unfortunately, the restricted range of LS/CMI items (0/1), often creates *gray area situations* that require administrator discretion and inconsistencies.

Cross-County Variability in Versions, Cut-Points, and Risk-to-Supervision Mapping

A significant finding was the extent of variation in LS/CMI implementation across Minnesota counties. Staff reported toggling between a 2004 manual and a 2018 supplement, with many expressing a desire for a single unified handbook that takes Minnesota-specific contexts into account. Participants generally believed that this contributes to discrepancies in how items are interpreted and scored across counties, with one participant noting:

> "*It would be really nice to have an 'LS/CMI-MN' for Minnesota with all the updated versions in a new handbook, so people aren't flipping between them, and we are all training the same thing each year. I literally have three or four of the 2018 updates with notes on each one that don't line up in some places because the scoring guidelines for some questions changed.*"

When discussing how LS/CMI scores translate to programming and supervision practices, respondents generally expressed the helpfulness of the tool, with the caveat that resource restrictions may limit how agents can turn tool recommendations into practice. Relatedly, many respondents described how their agencies' cut points are partially defined based on available supervision resources. Participants described how counties use different cut points to define risk levels (low, medium, high, very high), and many further adapt how these risk classifications map to supervision intensity based on local capacity and resources. While administrators generally expressed support for Minnesota-specific validation and standardized cut-points, they cautioned that supervision intensity is ultimately constrained by caseload sizes, personnel and resource limitations, and funding realities.

Adding to this variability, Hennepin County modified the standard LS/CMI scoring by removing several items in response to documented equity concerns identified through a local validation study. That study found that four items related to educational attainment, criminal family or spouse, and attitudes toward the sentence or offense exhibited racial bias. Specifically, these items were weakly related to recidivism but strongly correlated with race, functioning as proxies for race rather than valid indicators of individual risk. Black clients scored higher on these items, on average, despite limited predictive utility, contributing to disproportionate classification into the high-risk category and high rates of false positive errors (Wildermuth et al., 2021). As a result, Hennepin County scores these items "zero" for all individuals and

adjusted risk classification cut points to reduce racial disproportionality while maintaining predictive accuracy. However, these changes substantially alter the way the tool is scored across counties, where those assessed in other counties have a greater likelihood of being classified as high-risk compared to those assessed in Hennepin County. Therefore, these modifications altered the total score distribution and risk thresholds relative to jurisdictions using the unmodified instrument, complicating direct comparisons of risk scores and classifications across counties.

Overrides and "Trailer" Tools Are Routine for Specific Populations

Supplemental assessment tools, commonly referred to as "trailers", are used routinely alongside LS/CMI for certain offense types and populations. One common theme across focus groups was the tendency of the LS/CMI to score individuals charged with DUI, domestic violence, and sex offenses as lower risk than appropriate, often resulting in overrides. The most frequently used trailers include the Impaired Driving Assessment (IDA) and Drivers Risk Inventory (DRI) for DUI offenses, stable-static tools for sex offenses, Ontario Domestic Assault Risk Assessment (ODARA), and Domestic Violence Inventory (DVI) for domestic violence cases, and the Women's Risk Need Assessment (WRNA) for women. Many agencies combine results from multiple tools to govern supervision decisions.

Upward departures, where the LS/CMI risk levels were manually increased, were frequently discussed by participants for DUI cases, sex offenses, and some other person offenses, even when the tool indicates low risk. However, practices regarding when and how to document override decisions, and who must approve overrides vary across sites. This non-standardized use of trailer tools creates significant challenges during intrastate transfers. Those receiving transferred cases reported difficulty in interpreting trailer scores when they do not routinely use said tools in their own counties. Generally, there was a perceived lack of common understanding regarding what trailer assessments measure and how their results should inform supervision decisions, complicating case management continuity across county lines.

Reassessment Cadence and Capturing Dynamic Change

Current reassessment practices typically require a first reassessment at 6-to-12 months post-initial assessment, with annual reassessments thereafter. Some agencies conduct additional reassessments when supervision levels change or new offenses occur. Low-risk individuals are often reassessed less frequently than higher-risk cases. Users indicated that reassessments are

more accurate when they are able to maintain the same clients on their caseload over time, as this continuity allows them to build rapport and trust. However, this practice of maintaining consistent agent-client pairings is not standard, as some counties transfer cases between staff, which disrupts the relationship-building, a process perceived to facilitate more honest and complete assessment responses. Among those agencies that transfer cases between staff, tool users generally reported that maintaining clients for the duration of their supervision would be beneficial.

Both administrators and line staff expressed a desire for mechanisms to capture short-term change and desistance between these annual checkpoints. Participants cited ACUTE-style rapid checks as useful for detecting meaningful shifts in risk factors, particularly for higher-risk profiles where circumstances can change quickly. This desire was reported both in terms of maintaining public safety if a client demonstrates increasingly risky behavior, as well as responsiveness to clients' progress in addressing their needs. Notably, with some exceptions, the "public safety" focus was more commonly voiced by agency managers, while a "responsivity" focus was more typically voiced by agents.

Case Planning and Client Engagement

Participants commonly reported that the strength of their current LS/CMI implementation is its use in case planning and client engagement. Consistent with best practices, nearly all agents reported that they share assessment results with clients and use the tool's domains to collaboratively set goals and identify treatment needs. Participants emphasized that the LS/CMI helps avoid over-programming low-risk individuals, a key principle of evidence-based practice. Most agency administrators reported an expectation of agents to review scores with clients and use the assessment to frame responsivity needs and strengths. This practice appears to build rapport and help clients understand the rationale behind supervision requirements.

Moreover, some agencies reported using structured case management platforms, such as the Responsibility Assignment Matrix (RACI) and Epics, to organize and track intervention delivery alongside risk assessment results. These platforms help agents systematically document services provided, monitor compliance with case plan requirements, and maintain accountability for addressing identified criminogenic needs. Among departments using such models, participants reported that integration of such case management tools with the LS/CMI can

improve the translation of risk and need scores into concrete supervision activities. However, the use and depth of these case management systems vary across counties, with few reporting their full integration in daily tasks.

Quality Assurance, Posting, and Documentation Practices Vary

Quality assurance practices for LS/CMI assessments were also reported to differ substantially across counties. Some use second reviewers, or designated staff, to post notes on assessment scoring into the statewide system (S^3), while, in other counties, agents post their own work. Posting can be time-consuming, but participants emphasized it is critical for statewide visibility. When assessments are not posted to S^3, they remain "invisible" to other counties' post-transfers.

Documentation practices also vary, as some counties require narrative notes on every risk-scored item to explain the rationale, while others provide minimal documentation. Users broadly agreed that detailed notes in S^3 are important to provide essential context for understanding initial assessments after intrastate transfers. When receiving a transferred case, agents rely heavily on these notes to understand the scoring rationale and the client's circumstances at the time of assessment. In some instances, only scores (not full assessments) are visible in S^3, which many felt limited the utility of information in case planning decisions.

Resources and Caseload Realities

It was noted that specialized caseloads exist in many counties, including sex offense, domestic violence, drug court, enhanced, and mental health caseloads, but the sizes and targets for these caseloads vary widely. Some counties have agents who supervise only high-risk cases, while others organize caseloads around specific offense types or service needs. Participants emphasized that this variation often reflects broader differences in local capacity and available programming. In particular, smaller and rural sites identified significant resource gaps, especially in leisure and prosocial programming options, which complicate efforts to match identified needs to appropriate services. Here too, administrators repeatedly stressed that mapping risk levels to supervision intensity is fundamentally bound by local capacity, funding, and staffing constraints.

Under-Capture of Mental Health, Trauma, and Gender-Responsive Factors

A consistent concern across focus groups was that severe mental health needs often yield surprisingly low LS/CMI scores. Staff described encountering clients with significant mental

health issues who nonetheless score as low or low-moderate risk because the core LS/CMI items do not adequately capture these clinical dimensions. To address this gap, agencies use overrides or administer supplemental screening tools such as the WRNA. Participants emphasized that trauma histories and gender-responsive factors, particularly for women, need better visibility in the assessment process to appropriately inform supervision and case planning decisions. Unfortunately, WRNA recommendations can provide a stark contrast to that of the LS/CMI, making it difficult to translate assessment results into a case plan when the two instruments offer competing risk-based recommendations.

Standardizing Risk-Level Categories Is Both Desired and Contentious

Our data collection revealed some support among staff for a Minnesota-specific validation of the LS/CMI and adoption of standardized cut-points to improve fairness and portability across counties. Many participants view standardization to ensure similar cases are classified consistently, regardless of county. However, the prospect of standardization also generates concern, where some staff worry about political optics, particularly around how differential reoffending rates across jurisdictions would be defined and communicated. Others expressed concern about resource capacity. If standardized labels were to shift a greater proportion of people into higher-risk categories, some counties may not have the supervision resources to meet those needs. Some administrators, however, cautioned that risk classification should be separated from available supervision intensity. They emphasized that labels should be driven by validated predictive accuracy rather than being "*constrained by biggest caseload limitations*", though others acknowledged that supervision delivery will inevitably reflect local resource realities.

Differences Between Administrators and Line Staff

While administrators and line staff shared many concerns, their perspectives differed in emphasis. Administrators focused on system-level issues including the need for statewide validation, consistency across counties, and resource and caseload constraints. They prioritized reinstating proficiency testing and emphasized the importance of addressing transfer and posting issues that affect statewide coordination. Line staff emphasized day-to-day practical challenges such as ambiguous items (particularly peers/companions and negative phrasing), the value of walkthrough discussions for clarifying scoring decisions, and their reliance on detailed notes to

document rationales. Users were more likely to highlight specific gaps such as mental health needs not being captured by the assessment and resource constraints that directly affect their ability to match clients with appropriate services.

Summary of Focus Group Findings

Overall, the focus groups indicated that the LS/CMI is widely used across Minnesota probation agencies and is generally viewed as a useful framework for structuring risk assessment and case planning. Participants consistently described the tool as providing a shared language for discussing risk and criminogenic needs and as a practical mechanism for engaging clients in goal setting and supervision planning. At the same time, the discussions revealed substantial variability in how the tool is implemented across counties, particularly in training practices, scoring interpretation, risk classification cut points, and how risk levels translate into supervision intensity.

Several operational themes emerged across groups. Participants described uneven training and the absence of statewide proficiency testing as contributing to potential scoring inconsistencies. Staff also reported ongoing challenges interpreting certain items, particularly in domains with ambiguous wording or changing policy contexts. In addition, counties reported using different versions of LS/CMI manuals, locally defined risk thresholds, and a variety of supplemental assessment tools for specific offense types potentially impacting the tool's consistent application. Quality assurance practices, documentation standards, and reassessment procedures also differed across jurisdictions. Together, these findings show that the LS/CMI remains a common statewide framework, but its implementation often *requires agencies to work around tool limitations,* training structure, and local capacity. At the same time, focus groups reflect a strong commitment to evidence-based practice and suggest that Minnesota's agencies are well positioned to achieve greater consistency and utility with a more refined assessment approach.

Inter-Rater Reliability (IRR)

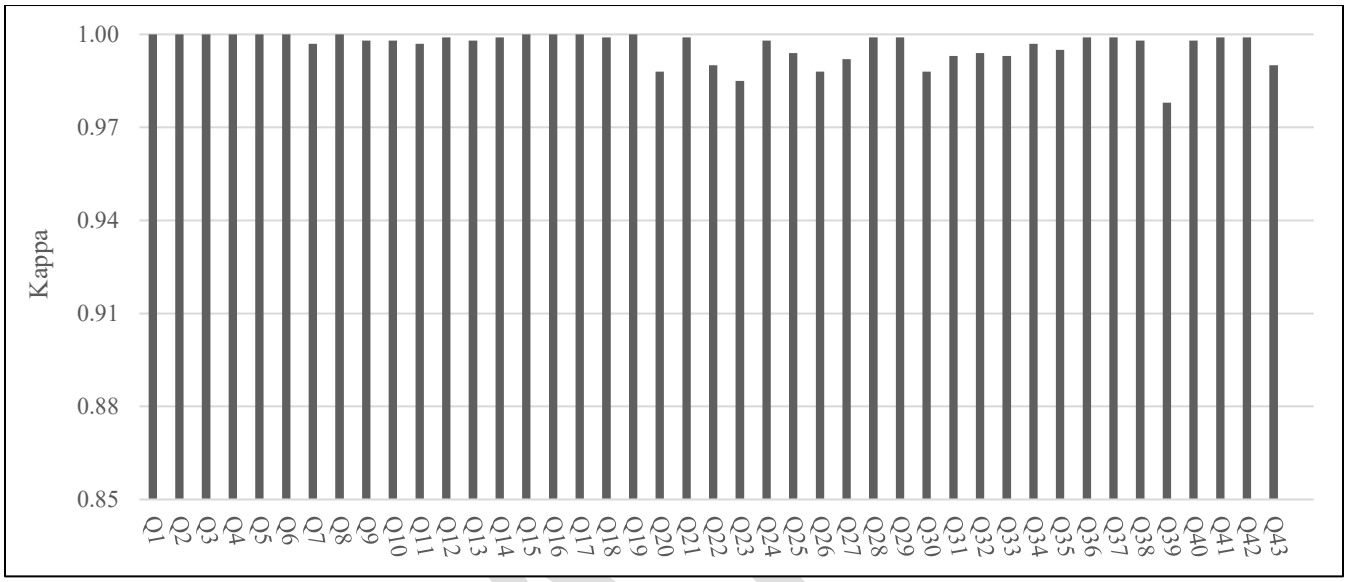
Building on the emergent themes from the focus groups, we examined whether the LS/CMI is scored consistently across assessors. This topic is especially important given the focus group findings describing uneven training, ambiguity in some items, and variation in scoring practices across jurisdictions. IRR refers to the degree of consistency between assessors when scoring the same case(s). Once completed, IRR findings are used to evaluate whether an assessment tool is applied in a dependable and standardized way across raters. In this context, strong IRR indicates that LS/CMI scores are not heavily dependent on who completes the assessment, which supports confidence in the instrument's consistency. This section provides an overview of the IRR findings, with the results for the entire sample displayed in Figure 2.

IRR analyses indicate that, overall, LS/CMI total scores demonstrate strong agreement across assessors. The average agreement statistic (Kappa) across the entire sample was 0.93, indicating *excellent* reliability. At the individual item level, agreement was highly uniform, with kappa values ranging from 0.97 to 1, suggesting near perfect agreement among raters. Agreement was strongest for clearly defined and historically anchored items, particularly within the criminal history domain, where Kappa values approached perfect agreement ($\kappa \approx 0.98$ to 1.00). These items rely on objective records and therefore produce highly consistent scoring across users. In contrast to concerns raised in focus groups regarding potential ambiguity in certain needs-based items, reliability results indicate that even items requiring interpretation of peer associations, leisure engagement, attitudes, and behavioral patterns demonstrated near perfect agreement, with Kappa values above 0.97.

Agency-level analyses similarly demonstrated consistently high agreement, with all jurisdictions producing Kappa coefficients in the excellent range for total score (0.83 to 0.97), and item-level reliability never falling below 0.90. This suggests that, under controlled conditions, assessors apply scoring criteria in a highly consistent manner across jurisdictions. Despite the high reliability demonstrated, the number of participating raters varied substantially across jurisdictions, with three jurisdictions contributing as few as two raters and twenty-three jurisdictions contributing ten or fewer, while five jurisdictions had greater than fifty raters participated. This uneven distribution means that reliability estimates for jurisdictions with few

raters should be interpreted cautiously, as Kappa statistics derived from very small rater pools may be less stable than those generated from larger samples.[2]

Figure 2. Item-Level Reliability for Entire Sample



Summary of IRR Findings

Overall, the IRR findings indicate that Minnesota assessors apply the LS/CMI with a high degree of consistency, even considering the implementation concerns raised in the focus groups. This strong agreement observed here reflects a meaningful accomplishment and suggests that practitioners have developed a dependable scoring process within the current system. At the same time, these results also suggest that inconsistency in outcomes is less likely to stem from assessor error alone and more likely to reflect limitations in the tool itself, or in how it is structured for Minnesota practice. Put differently, the findings point to a system in which staff appear to be using the instrument as intended, suggesting that improvements in performance will be difficult to gain through improved user compliance or consistency.

[2] Item-level reliability results for each jurisdiction can be found at https://arc-data-dashboard.shinyapps.io/LSCMI_Dashboard/.

Predictive Validity Findings

This section presents findings examining the LS/CMI's predictive validity in the Minnesota probation system. In this section, we present four sets of analyses. First, we report the AUC values predicting general and violent reconvictions using a common three-year follow-up, consistent with MNDOC's operational definition of recidivism. We report the overall AUC for both outcomes, as well as specific estimates by gender and race/ethnicity. Further, we present statistical models using criminal history and the needs scales separately. The intent of this analysis was to identify if one scale performs unevenly or drives most of a greater proportion of the prediction, which may reveal sources of bias inherent within LS/CMI scoring.

Second, we examine prediction bias across gender and race/ethnicity subgroups using iterative logistic regressions for general and violent recidivism. These models include LS/CMI scores, a group indicator for sex or race/ethnicity, and an interaction term of the LS/CMI score and the group indicator. These models follow the industry standard *Cleary Method*. Using this approach, intercept bias is tested via the group indicator. A significant group effect identifies inequality in the tools risk prediction across groups. The interaction term between LS/CMI score and group is then used to test slope bias. Here, a significant interaction indicates that the relationship between the risk score and the outcome differs by group. In other words, a one-point increase in LS/CMI score may be associated with a larger increase in the likelihood of reoffending for one group than for another. Given the large sample size, interpretation centers primarily on odds ratios (ORs) rather than statistical significance alone.[3] In large samples, very small differences can reach conventional significance thresholds despite limited practical importance. Accordingly, ORs are used to assess the substantive magnitude of effects. ORs close to 1.00 indicate little-to-no meaningful difference in the likelihood of reoffending, while values that depart further from 1.00 reflect stronger effects.

Third, we examine predictive validity over time, as prior research has indicated that gradual changes in state demographics and departmental policies can affect a tool's accuracy (Hamilton et al., 2022). Subsequently, identifying periods when accuracy shifts can point to policy or practice changes that warrant closer review. Finally, we build on the focus group findings by testing differential validity between initial assessments and reassessments,

[3] Statistical significance was evaluated using p values, with $p < .05$ treated as significant.

comparing initial assessments completed at the start of a supervision sentence with those completed after an intrastate transfer. As indicated, focus group members suggested that reassessments can be more accurate once officers have built rapport with the individual, whereas assessments following intrastate transfers may be less accurate.

Predictive Accuracy Findings

Table 2 presents the AUC values for the LS/CMI in predicting general and violent reconvictions. For the total sample, the LS/CMI demonstrated moderate predictive validity for general reconvictions (AUC=0.65) and a *weak* effect when predicting violent reconvictions (AU=0.63). Predictive accuracy for general reconviction was equitable between males and females (AUC=0.65), and a negligible difference in predictive validity found for violent reconvictions (Male AUC=0.63, Female AUC = 0.64). Pertaining to race/ethnicity, the LS/CMI demonstrated moderate predictive accuracy for Hispanic individuals when predicting general (AUC-0.66) and violent (AUC=0.65) reconviction. Moderate predictive validity was also observed for White individuals when predicting general reconviction (AUC=0.66). For the remaining sample, AUCs indicated weak effect sizes, with values for general reconvictions ranging from a low of 0.60 for American Indians, to 0.62 for Black individuals. For violent reconviction, the values range from 0.59 for American Indians to 0.62 for Black individuals.

Table 2. AUCs for General & Violent Reconviction by Gender & Race/Ethnicity

Outcome	Total	Male	Female	White	Black	Hispanic	American Indian	Other/Unknown
General	0.65	0.65	0.65	0.66	0.62	0.66	0.60	0.65
Violent	0.63	0.64	0.63	0.63	0.62	0.65	0.59	0.63

Next, we examined differences between the predictive validity of criminal history and needs domains, identifying minimal differences (AUC=0.63 & 0.64, respectively). Notably, the criminal history scale only possesses eight assessment items, while the needs scale comprises thirty-five, indicating that, while the needs-based items may be helpful for case management, these items provide limited incremental value in predictive validity, above and beyond, the criminal history scale.

Prediction Bias

Here, we first present performance by gender and race/ethnicity and then present formal analysis to determine the magnitude of the bias detected. Figure 3 shows LS/CMI performance

by gender for general and violent reconviction. The figure shows slope bias, where trend lines separate as risk scores increases. This pattern indicates that increasing risk scores are more strongly associated with increasing recidivism for males than females, leading females to be increasingly overclassified as their assessed risk increases. For general reconviction, the gap is greatest in the middle of the scale, identifying a substantial 11% difference in recidivism probability for males and females on the LS/CMI. For violent reconviction, the gap is greatest at the higher end of the risk scale, also producing a 11% reoffending differential, despite receiving the same assessment score.

Notably, the LS/CMI is not specifically designed to predict violent recidivism. When comparing the two plots, importantly the violent recidivism trends both indicate flatter slopes. This finding is indicative of a poor performing scale, where there is only a 5% increase in the prediction of violent recidivism as the scale moves from the low risk range (scores of 5 to 10 using provider recommended cut points) to medium (scores of 11 to 19), and another 5% when scores increase from the medium to the high-risk range.

Figure 3. LS/CMI Performance by Gender

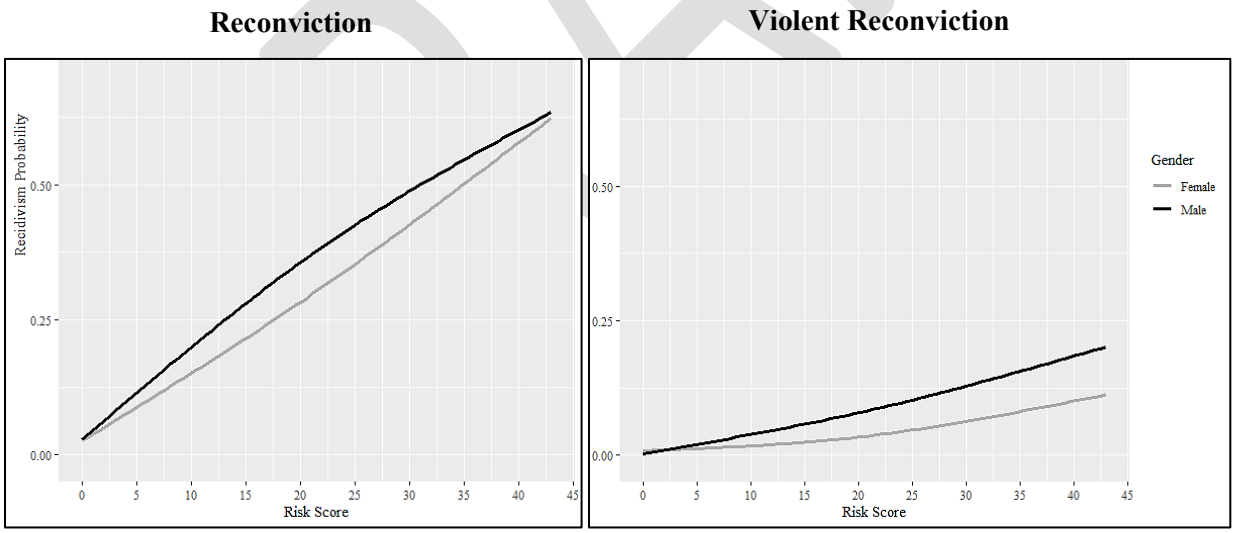


Figure 4 indicates a similar pattern for White individuals compared to Non-White individuals. For general reconviction, the greatest separation between groups appears in the low-risk end of the scale, where Non-White individuals are overclassified by about 8%. That separation remains consistent as risk scores increase and ultimately narrows as risk scores reach 35. For violent reconviction, the pattern is reversed. Group separation is limited at lower risk

levels but becomes more pronounced at the high-risk end of the scale, where Non-White individuals are overclassified by roughly 11% in the very high-risk range. Again, the flatter trend of the violent recidivism plot indicates poor predictive discrimination.

Figure 4. LS/CMI Performance by Race/Ethnicity

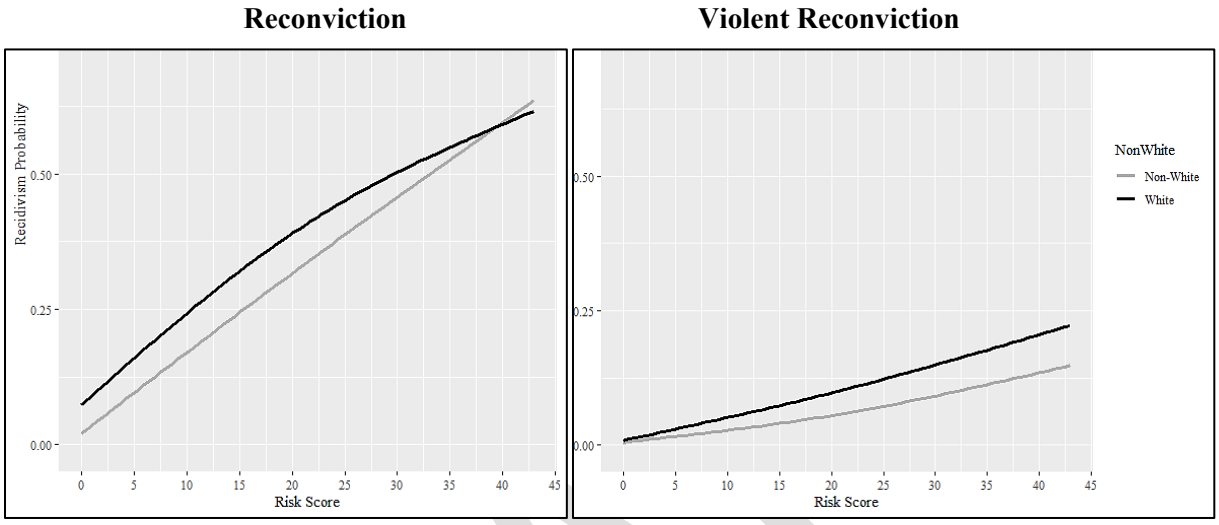


Table 3 shows the results from the final general and violent reconviction models. When examining the general reconviction models, Model 2 found a significant and negative coefficient for females for both general (OR = 0.63; z = -11.2) and violent reconviction (OR = 0.37; z = -10.1), indicating that females were reconvicted less often than males with the same score (i.e., intercept bias). In other words, *the LS/CMI systematically over-predicts risk for females*. However, the interaction in Model 3 is only marginally significant for general reconviction and the odds ratio is close to 1.00 (OR = 1.01), indicating a negligible degree of slope bias. The interaction for violent reconviction produced an OR of 0.89, indicating that *each one-unit increase in risk score has an 11% smaller effect on the odds of violent reconviction for females compared to males.* Essentially, there is greater bias across risk scores for males compared to females when predicting violent recidivism.

Table 3: Logistic Regression Models Examining Predictive Bias for Females

Outcome/Measures	Final Model		
General Reconviction	**Est.**	**z-value(SE)**	**OR**
LS/CMI	0.65	85.2(.001)	0.92
Female	-0.46	-11.2(.041)	0.63
Female*LS/CMI	0.02	3.8(.001)	1.01
Violent Reconviction			
LS/CMI	0.06	46.9(.001)	1.06
Female	0.01	-10.1(.098)	0.37
Female*LS/CMI	-0.01	1.5(.003)	0.89

Table 4 provides bias analyses for the Non-White group compared to White individuals. Again, we identify significant intercept bias for both general reconviction (OR = 1.78; z = 17.2) and violent reconviction (OR = 2.04; z = 11.3). Moreover, the Non-White indicator remains significant in the final model for violent reconvictions, indicating that the LS/CMI produces both intercept bias in predicting general reconviction among Non-White individuals and slope bias when predicting violent reconvictions. Different from the comparison of males and females, these findings demonstrate that Non-White individuals have larger risk scores, on average, indicating overclassification when predicting recidivism generally. However, the gap between risk trends decreases as scores increase. Yet, when examining violent recidivism, slope bias is identified, where Non-White individuals are overclassified when compared to White individuals and the gap between trend lines increases as risk scores increase.

Table 4: Logistic Regression Models Examining Predictive Bias for Non-White Individuals

Outcome/Measures	Final Model		
General Reconviction	**Est.**	**z-value(SE)**	**OR**
LS/CMI	0.07	79.3(.001)	1.07
Non-White	0.57	17.2(.033)	1.78
Non-White*LS/CMI	-0.01	-9.4(.001)	0.99
Violent Reconviction			
LS/CMI	0.06	34.8(.001)	1.06
Non-White	0.71	11.3(.043)	2.04
Non-White*LS/CMI	-0.01	-1.7(.002)	0.96

Figure 5 displays the predictive value of the LS/CMI criminal history and needs domains by sex. As illustrated, the divergence (gap between trend lines) in reoffending probability between males in females is consistent when examining the criminal history score, which is indicative of intercept bias. Specifically, males are roughly 5% more likely to reoffend than

females with the same criminal history score. In the needs score panel, the gap between male and female recidivism probabilities increases as risk scores increase, highlighting the item types with greater slope bias. In this instance, males are roughly 11% more likely to reoffend than females with the same LS/CMI needs scores.

Figure 5. Criminal History & Needs Domains Reoffending Probability by Gender

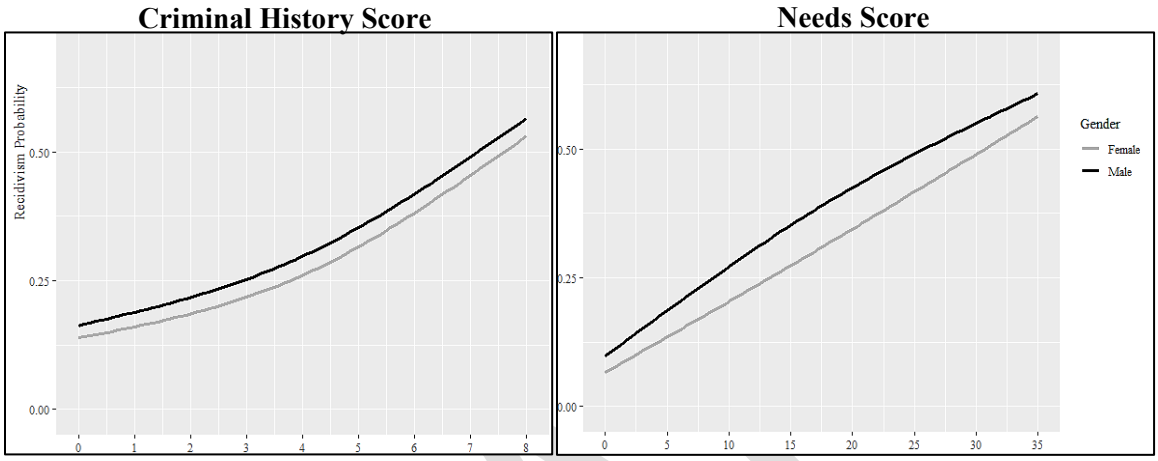
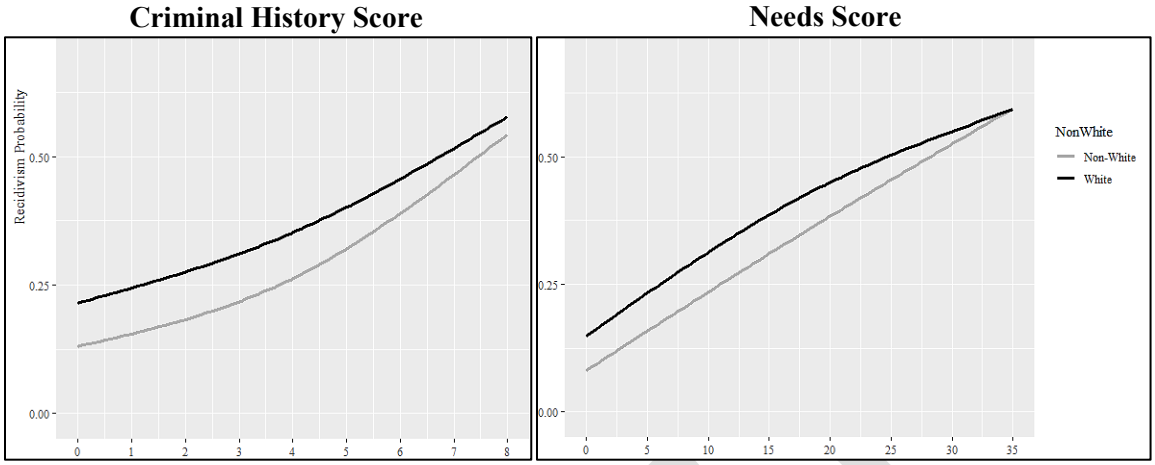


Figure 6 compares White and Non-White individual's criminal history and need scores. For this comparison, the criminal history domain demonstrated intercept bias, where White individuals are more likely to reoffend than Non-White individuals with the same criminal history score.[4] Interestingly, the needs domains panel also shows intercept bias and a smaller degree of slope bias, where increasing needs were more strongly associated with increasing reoffending risk for Non-White individuals compared to White individuals. Overall Figures 5 and 6 demonstrate consistent bias, where moderate levels of overclassification are present for both females and non-White individuals.

[4] See Appendix A for plots displaying each racial/ethnic category.

Figure 6. Criminal History & Needs Domains Reoffending Probability by Race/Ethnicity



Overall, these results indicate consistent evidence of differential prediction across both gender and race/ethnicity. Females and Non-White individuals generally experience lower observed recidivism rates than their counterparts with the same LS/CMI scores, resulting in patterns of overclassification. While the magnitude of these differences varies across outcomes and points along the risk scale, the findings are consistent across descriptive and model-based analyses. While not shown here, the patterns of bias seen in criminal history and needs scales are more severe when examining violent reconviction. Taken together, these results indicate that the LS/CMI performs unevenly across groups, with both intercept and slope bias present in key comparisons.

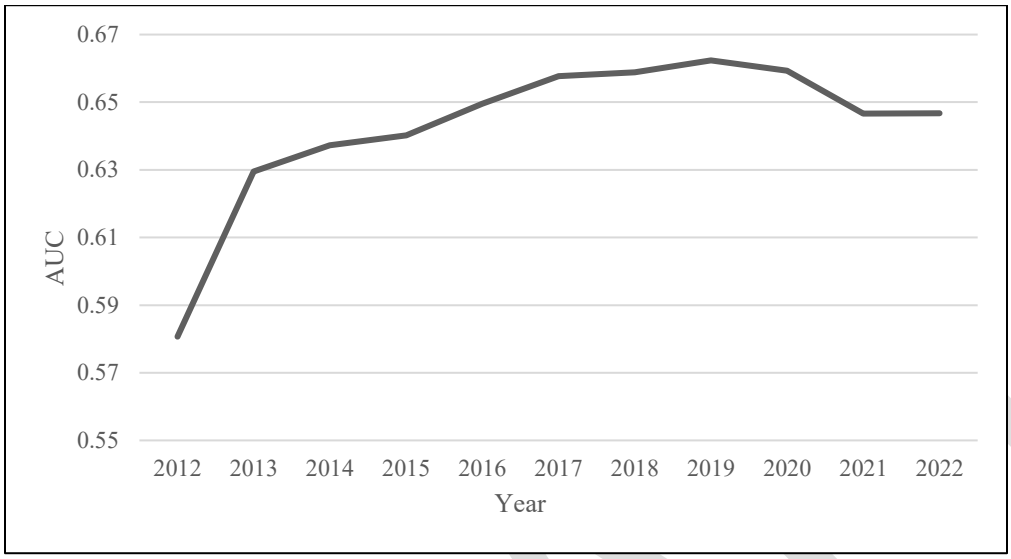
Predictive Accuracy Over Time

Next, we examined how the predictive accuracy of the LS/CMI has evolved over the course of its provision in Minnesota. Figure 7 represents the average annual AUC in predicting general reconviction[5]. Beginning in 2012, the average AUC was 0.58, falling within the 'negligible' effect size range. While notable, a higher degree of inaccuracy is expected soon after tool adoption as agencies and staff adapt to its use and scoring. However, the average accuracy sharply increased the following year and steadily increased thereafter until 2019, reaching an average AUC of 0.66. Notably, predictive accuracy began to decline after 2019, echoing concerns voiced by stakeholders that state-wide training and inter-rater reliability assessments

[5] We note that the annual trends stop at 2022, as a three-year follow-up window is needed to assess recidivism, which extends the tracking period from 2022 through the end of our sample frame in 2025.

were discontinued with the onset of the COVID-19 pandemic. While accuracy was no longer declining in 2021 and 2022, it has not yet increased to pre-COVID levels and remains at 0.64, the floor of the "acceptable" accuracy range.

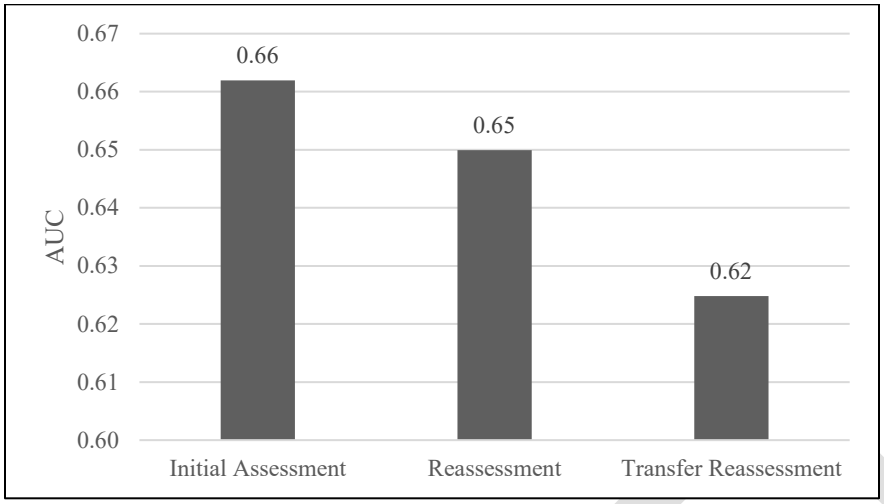
Figure 7. Average AUC Over Time for General Reconviction Within Three Years



Predictive Accuracy by Assessment Type and Referral Source

In response to concerns raised during the focus groups, we compared the accuracy of initial assessments, reassessments, and first assessments after an intra-state transfer. Figure 8 displays the average AUCs for each assessment type. Little variation was found regarding accuracy between initial assessments (AUC=0.66) and reassessments (AUC=0.65), with both demonstrating moderate effect sizes. However, there is a notable decrease in predictive accuracy for an initial assessment that occurs after an intra-state transfer, with predictive accuracy falling to a weak effect size range (AUC=0.62). Information gathered from the focus groups indicate that, in the initial assessment the agent is gathering all the necessary information, while, upon transfer, the agents are not conducting another interview but are instead using information from the notes in S3 to score a new assessment. This revised process leaves room for mis-calculation, failing to account for the dynamic changes of occurring since the initial assessment. These findings also validate concerns raised by stakeholders during focus groups indicating that intrastate transfers impact the accuracy of the tool's prediction, potentially due to a lack of rapport and unfamiliarity between client and officer.

Figure 8. AUC by Assessment Type



Summary of Predictive Validity Findings

Overall, the LS/CMI demonstrated moderate predictive validity for general reconvictions and weak predictive validity for violent reconvictions in the Minnesota probation population. Predictive accuracy was largely consistent across gender groups and varied modestly across racial and ethnic groups, with lower predictive discrimination observed for some subgroups. Viewed in context, this level of performance suggests that Minnesota agencies have been able to obtain reasonably strong results from the instrument in practice, particularly given the complexity of the population and the limitations of an assessment not originally calibrated to this setting. When examining criminal history and needs scales separately, both presented similar levels of predictive validity despite the substantially larger number of items included in the needs scale.

At the same time, findings indicate a need for tool improvement. Analyses identified bias for females and Non-White individuals. Predictive accuracy also changed over time, increasing during the early years following implementation and declining slightly after 2019. Moreover, validity was lower for assessments conducted following intrastate transfers, echoing concerns voiced by focus group participants. When taking predictive validity and IRR findings together, the results suggest that Minnesota's probation agencies are scoring the LS/CMI reliability and appear to be aware of some of the contexts in which its validity is more limited, such as after intrastate transfers. When considered alongside the strong IRR results, the observed bias patterns are not consistent with a problem of inconsistent scoring but are indicative of limitations in the

tool itself. Overall, the findings suggest staff implementation of the tool is effective, however, current instrument limitations only allow the LS/CMI to max-out at the borderline, or acceptable, level of predictive accuracy in Minnesota.

Functionality Findings

To further examine the effects of the tool, we provided a breakdown of its elements. This section examines how the LS/CMI functions at the item, domain, and risk level category levels. It also examines calibration – how risk scores translate into to actual reoffending rates – and highlights how calibration error can impact effective resource usage through classification. While all analyses were conducted separately for general and violent reconvictions, we report the former for most examinations, as results were consistent for both outcomes. We display AUCs as a measure of discrimination for each domain and report FPRs to identify items that may contribute to overclassification by labeling individuals as higher risk despite reduced rates of reoffending.

Next, we display results regarding risk-level categories. Importantly, Minnesota does not use a single set of standardized risk levels across all counties. Therefore, we first calculated ORs for each county using their county-specific cut points and risk level categories and report the average and range of ORs. Next, we apply the risk level category scheme used by the DOC to see how applying a standard set of risk levels would impact classification performance. Finally, we evaluate tool calibration and present recalibrated tool findings as a proof of concept to potential improvements to the LS/CMI in Minnesota.

Domain and Item Performance

Figure 9 displays the predictive validity of each domain separately for general and violent reconviction.[6] The strongest domains were the Criminal History (8 items) and Antisocial Pattern domains (2 items), with the latter being the only domain to demonstrate a moderate predictive value (violent reconviction AUC=0.64). While each item in the Criminal History domain possessed adequate predictive values (AUCs ranging from 0.53 to 0.56) [7], items one, two, and six displayed high FPRs (0.88, 0.78, & 0.81 respectively).[8]

Interestingly, the Pro-criminal Attitude (4 items), Family/Marital (4 items), and Education/Employment (9 items) domains displayed modest improvements in violent

[6] Supplemental information on item-level analyses can be found at https://arc-data-dashboard.shinyapps.io/LSCMI_Dashboard/.
[7] AUCs below 0.55 are considered Negligible, 0.56 – 0.63 are Small, 0.64-0.70 are Moderate, and 0.71 and Above are Strong (Rice & Harris, 2005).
[8] False positive rates for violent reconviction were 0.90, 0.83, and 0.81 respectively.

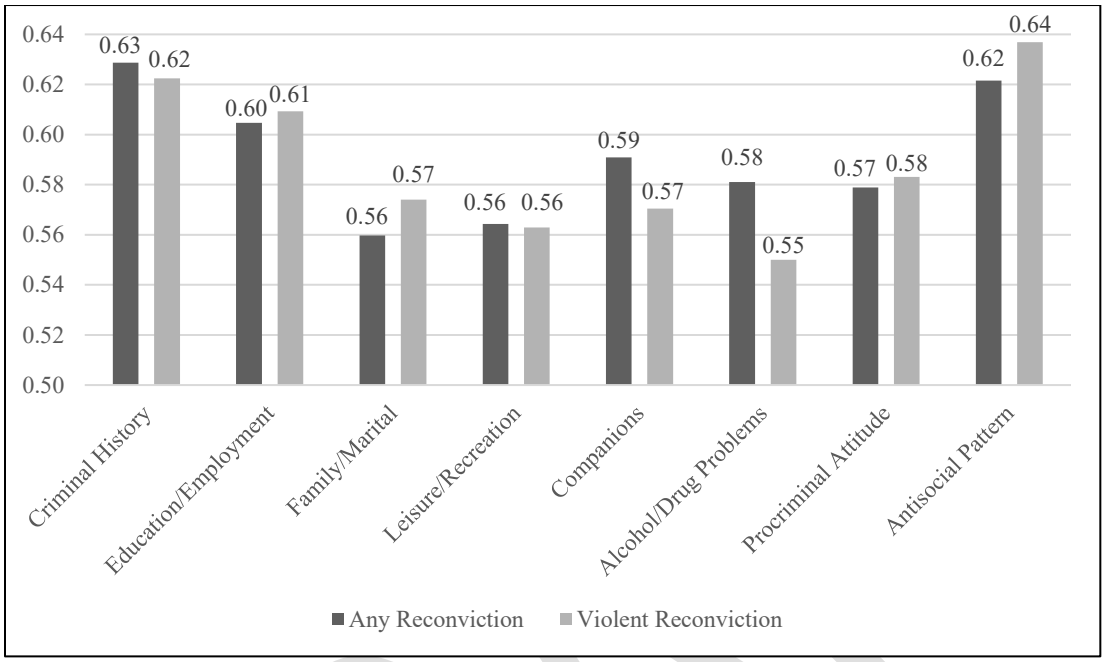
reconviction compared to general prediction. Notably, only the Alcohol/Drug Problems (8 items) domain fell into the 'negligible' effect size range when predicting violent reconviction (AUC=0.55), and item twenty-nine, "Drug problem ever", displayed a high false positive rate, especially among females (FPR=0.80). Certain items also produced large differences in performance across racial groups. For example, item twelve in the Education/Employment domain, "Less than regular grade 10 or equivalent", showed very low false positive rates among White individuals but very high rates and low predictive value among Non-White individuals, suggesting that this item possesses substantial prediction bias. The remaining two domains, Leisure/Recreation (2 items) and Companions (4 items) produced equitable predictive value between the two outcomes but also contained items with high FPRs. Specifically, item twenty-two, "Absence of recent participation in an organized activity", in the Leisure/Recreation domain displayed a high FPR for Non-White individuals (FPR=0.78), and item twenty-four, "Some criminal acquaintances", in the Companions domain, which was highlighted in focus groups as being ambiguous, also demonstrated a high FPR for all groups (0.81).

Overall, these findings suggest that the LS/CMI's predictive power is driven by a small number of domains, particularly Criminal History and Antisocial Pattern, while most other domains contribute limited incremental accuracy. This pattern is consistent with prior research on the LSI-R, where evaluations in Washington State found that criminal history alone was nearly as predictive as the full instrument, and that many domains added little incremental value, and that the tool demonstrated diminished predictive accuracy for violent reoffending (Barnoski & Aos, 2003). These findings contributed to a shift away from reliance on the full instrument and reflect a broader dynamic commonly described as predictive shrinkage (Hamilton et al., 2022). However, domains and items with comparatively stronger AUCs often exhibit high false positive rates, indicating a tendency toward overclassification rather than precise risk differentiation. Importantly, several items produce disproportionate false positives for females and Non-White individuals, signaling that predictive accuracy at the domain level can mask meaningful sources of bias at the item level.[9] Taken together, these results indicate that domain specific AUCs

[9] The items were "Less than grade 10 or equivalent", "Absence of recent participation in an organized activity", "Some criminal acquaintances", and "Drug problem ever".

should be interpreted cautiously, as modest predictive gains frequently coincide with equity concerns that are highly consequential in applied decision-making contexts.

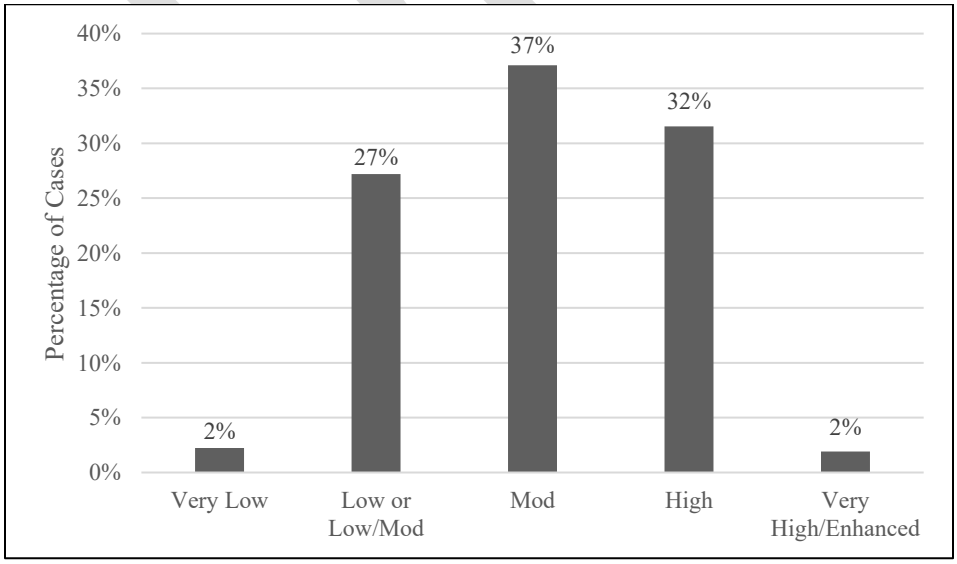
Figure 9. Reconviction & Violent Reconviction AUC by Domain



Risk-Level Categories

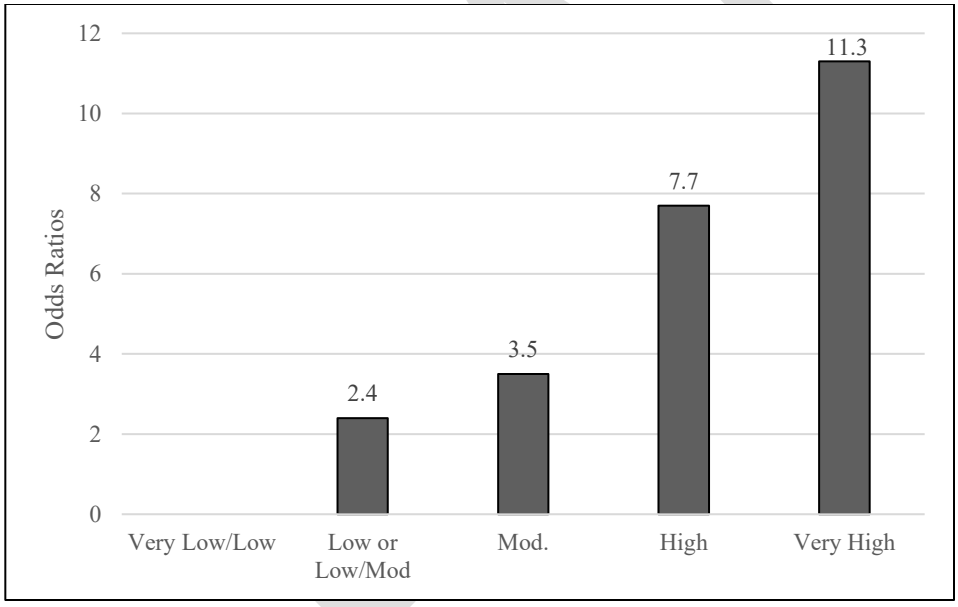
Next, we examined risk-level categories. In doing so, we first identified classifications and thresholds across Minnesota's counties. Figure 10 presents the proportion of cases falling into each category. Notably, few cases are classified as Very Low or Very High-Risk (2%).

Figure 10. County-Specific Risk-Level Categories Case Distribution



Due to the low sample size of the Very Low-Risk category, when calculating odds ratios, we combined the "Very Low" and "Low" categories to be used as the reference in our comparisons. Figure 11 displays the risk category odds ratios. Notably, the figure shows promising classification results, where the odds of being reconvicted within three years of assessment increases with each category, or "steps-up". The odds of reoffending were 2.4 times greater for those classified as Low/Moderate compared to those classified as Very Low/Low. Those classified as Moderate had 3.5 greater odds of reoffending. Notably, the increase in reoffending probability between the Low-Moderate and Moderate groups is much smaller compared to the difference between the High and Very-High risk categories, indicating that the LS/CMI has trouble distinguishing between lower risk groups. While the range of effect sizes across counties was quite large, ORs remained consistently above 1, consistent with a strong classification scheme.i[10]

Figure 11. Odds Ratios for County-Specific Risk-Level Categories



Next, we evaluated what would happen if counties applied the DOC's risk category cut points. The Minnesota DOC uses a graduated supervision framework that classifies individuals into Low Phase 2 (0-10), Low Phase 1 (11-19), Moderate Phase 2 (20-24), Moderate Phase 1

[10] The ORs ranged from 1.1 to 7.5 for Moderate Risk, 2.8 to 15.1 for High Risk, and 7.3 to 15.1 for Very High Risk.

(25-31), and High (32+). These levels are tied to assessed risk and need, with the Field Services Contact Standards indicating that each category corresponds to specific minimum contact requirements. Higher intensity categories require more frequent monthly contacts and home visits, while lower intensity categories involve less frequent contact and more flexibility in whether supervision occurs through office, virtual, or other approved formats. Compared with the current approach, the distribution becomes less consistent, with more cases concentrated in the lowest and highest categories. The share of cases in the lowest group increases from 2% to 13%, while the combined share in the two highest categories remains essentially unchanged. The practical implication is that standardization would primarily shift a meaningful number of individuals into the lowest supervision tier, and increase the number classified as highest risk. This could reduce supervision intensity for some low-risk cases and improve consistency across counties, but it also underscores the need to confirm that the expanded lowest category still aligns with comparable recidivism risk statewide.

Figure 12. Standardized Risk-Level Category Case Distribution

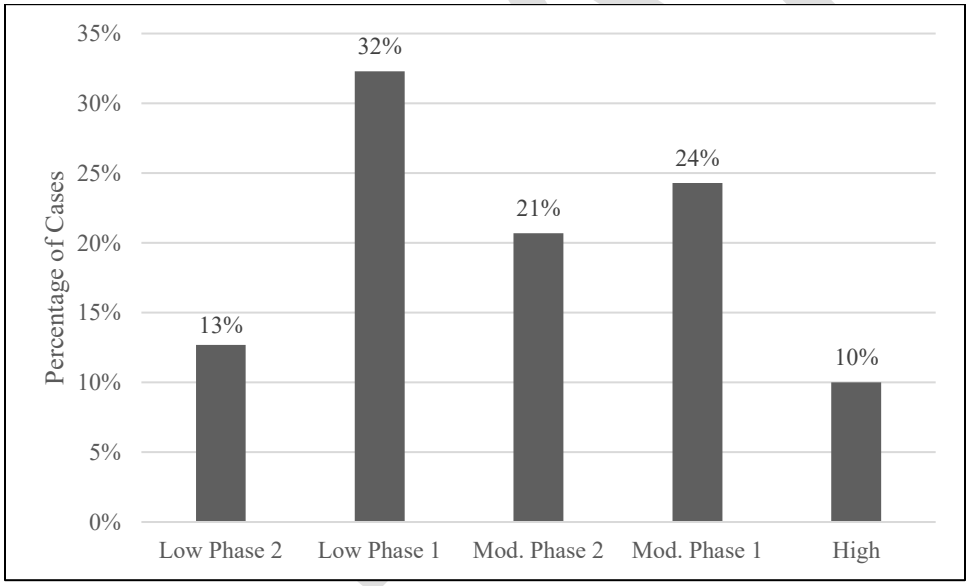
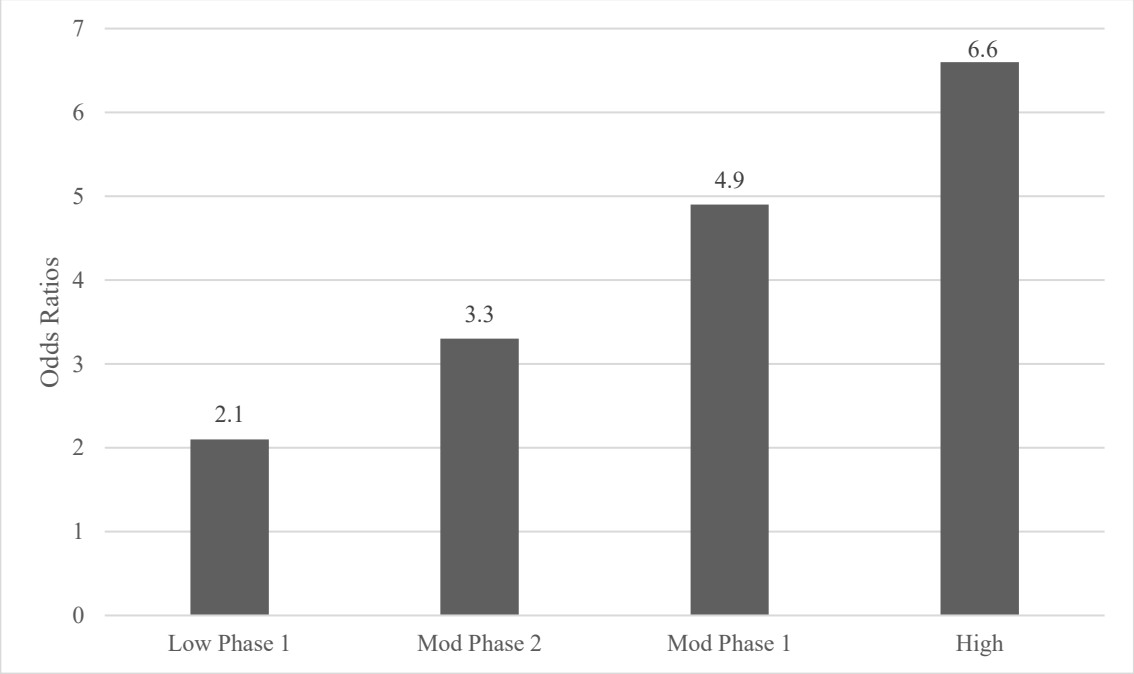


Figure 13 displays the odds ratios for the entire sample, using Low Phase 1 as a reference. Here too, we see the increasing "stair-step" pattern indicative of a good classification scheme. Notably, while there is a more uniform increase in ORs with each step up in risk-level category, the increase in odds ratios between each is less distinct under this version when compared to county-specific schemes, indicating a diminished distinction of those who reoffend

among risk level categories. However, the differences in odds ratios between the standardized and county-specific schemes are small for low and moderate risk categories, particularly for the Low Phase 1 (2.1 versus 2.4) and Moderate Phase 2 (3.3 versus 3.5) groups.

Figure 13. Odds Ratios for Standardized Risk-Level Categories (Entire State Sample)



Evaluation results indicate that LS/CMI's predictive performance varies across domains and items. Criminal History and Antisocial Pattern demonstrated the strongest discrimination, while several other domains showed weaker predictive value. At the item level, a number of indicators produced high false positive rates, and several items showed notable differences in performance across demographic groups. These patterns suggest that domain-level predictive statistics may mask substantial variation in performance at the item level.

Analyses of risk level categories show that the odds of reconviction generally increase across categories, indicating that the LS/CMI classifications distinguish between higher and lower risk individuals. However, the magnitude of these differences varies across counties due to the use of different cut points and category structures. Applying the standardized LS/CMI classification scheme produces a different distribution of cases, with a larger share classified in the lowest and highest supervision categories, with less predictive discrimination between risk levels.

Calibration Findings

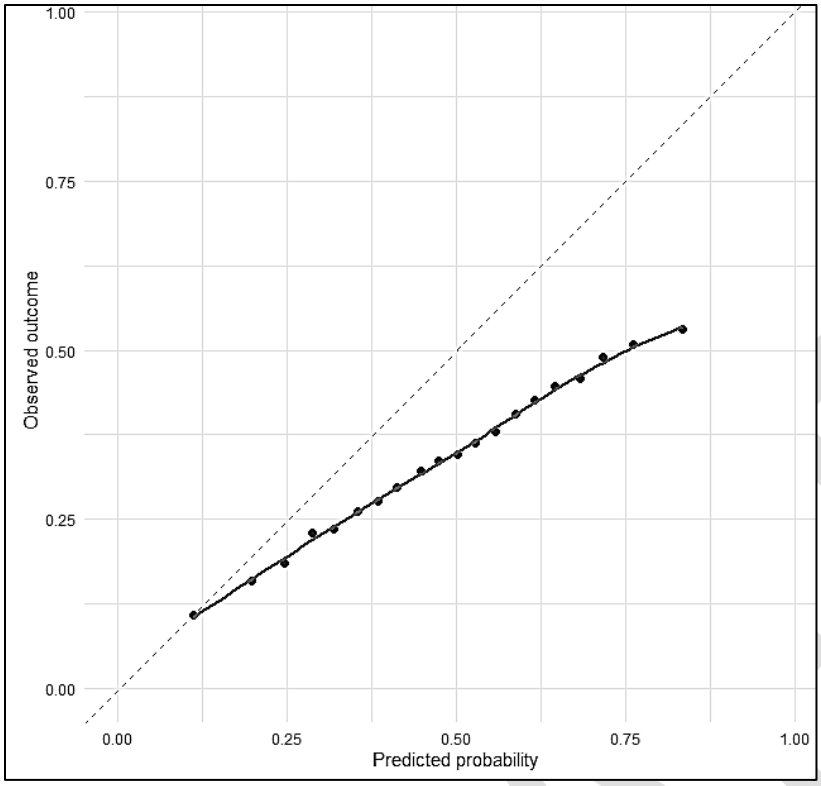
This section examines calibration of the LS/CMI in our sample. Unlike the AUC, which focuses on predictive discrimination (i.e., ranking those who reoffend as higher risk than those who do not), calibration examines whether increases in risk scores are associated with an equitable increase in reoffending rates. Calibration plots compare the expected rate of reoffending to the actual rate of reoffending.[11] [12] These plots display a diagonal dotted line, representing perfect calibration, where, for example, a reoffending probability of 0.75 is associated with an actual reoffending rate of 75%. We also report the performance RMSE and calibration MCE, describing statistical performance of the tools' calibration.

Figure 14 displays the calibration plot of LS/CMI scores predicting general reconvictions. Findings show that the LS/CMI trend line is below the dashed diagonal, indicating that the tool consistently *overpredicts* reoffending risk among Minnesota probationers. For example, Figure 14 shows that a reoffending probability of 0.75 was associated with an actual reoffending rate of about 51%, an overprediction of 24 percentage points. Generally, the tool struggles most with accuracy at the top end of the scale. For instance, while it may only slightly overpredict for low-risk individuals, it misses the mark by a much wider margin—roughly 24 percentage points—when it identifies someone as having a very high probability of reoffending. Moreover, the RMSE is 0.21, which means the model's predicted reoffending probabilities are off by about 21 percentage points on average. The MCE is 0.38, indicating a 38% mismatch, on average, between predicted risk and observed recidivism at the high end of the risk scale. Notably, overprediction increases as predicted risk increases, with negligible calibration error at the bottom end of the scale and the most substantial miscalibration occurring at the high-risk end of the scale.

Figure 14. Calibration Plot for LS/CMI Scores

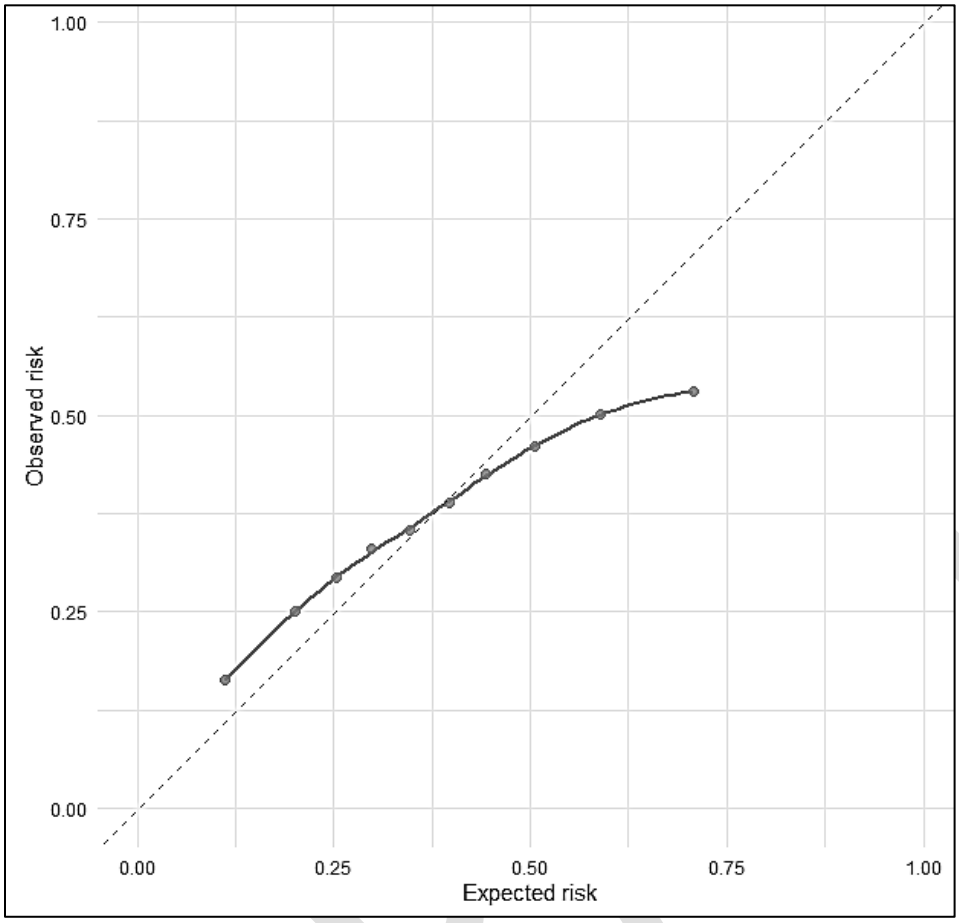
[11] Re-calibration does not impact the scoring of the tool, and therefore, the AUCs remain the same. While scoring remains unchanged, matching expected and observed reoffending has significant implications for classification.

[12] See Appendix B for how recalibration affects risk-level category classification.



Next, we re-calibrated the assessment, adjusting the predicted probabilities of reoffending to more closely match the observed rates of reoffending. This recalibration adjusts the overall level of predicted risk up (or down) without changing how individuals are ranked by risk. In other words, individuals' risk scores remain the same, but the probabilities attached to those scores are shifted to better reflect the actual reoffending rates observed in the data. Figure 15 displays LS/CMI expected versus observed reoffending rates after recalibration. Predicted probabilities were recalibrated by fitting a logistic regression model that adjusts only the intercept, leaving the relative ordering of risk scores unchanged. This shifts all predicted risks up, or down, so that the average predicted probability aligns with the observed outcome rate of the sample. Notably, the trend line more closely mirrors the perfect calibration line, and consistent overprediction is no longer an issue. After re-calibration, the RMSE falls to 0.06, indicating a 15-percentage point reduction in error when compared to the raw LS/CMI scores. Moreover, the MCE is 0.14, a 24% reduction in the maximum calibration error. We provide additional recalibration details in Appendix B, demonstrating how recalibrating tool predictions can substantially impact supervision caseloads.

Figure 15. Calibration Plot After Re-Calibrating Scoring Probabilities



Overall, calibration findings indicate that, while the LS/CMI is capable of distinguishing between individuals at relatively higher and lower risk of recidivism, the probabilities attached to those scores are not well aligned with observed outcomes in Minnesota. These calibration analyses shows that the tool systematically overestimates the likelihood of reoffending across much of the LS/CMI scoring distribution. As a result, individuals may be classified into supervision levels based on risk estimates that exceed their actual likelihood of recidivism. This type of miscalibration has practical implications for probation systems, as inflated risk predictions can lead to unnecessary supervision intensity and inefficient allocation of limited resources.

The recalibration results demonstrate that this issue can be substantially improved without altering the relative ranking of individuals by risk. By adjusting the predicted probabilities to better reflect observed recidivism rates, the assessment produces estimates that

more closely align with real outcomes in the Minnesota probation population. The resulting reductions in RMSE and MCE indicate a meaningful improvement in calibration accuracy. These findings highlight the importance of periodically evaluating and recalibrating risk assessment tools when they are applied to new populations or over time, ensuring that predicted risk levels remain consistent with observed patterns of recidivism and support more efficient case management decisions.

Summary of Functionality Findings

Overall, the functionality analyses indicate that the LS/CMI produces a coherent and usable classification structure in Minnesota, with risk categories that generally differentiate levels of recidivism. At the same time, patterns observed across items, domains, and calibration suggest that this performance is achieved despite notable inefficiencies, including high false positive rates, variation across counties, and systematic overprediction of risk. Viewed together, these findings suggest that Minnesota agencies are using the tool effectively and extracting meaningful information from it in practice. However, findings also indicate that there is considerable opportunity to improve both precision and alignment with observed outcomes through refinements to the current assessment approach.

CONCLUSIONS AND RECCOMENDATIONS

Minnesota's use of the LS/CMI is extensive and unstandardized. The focus groups describe a common language for risk and needs that is applied within a somewhat piece-meal set of training practices, scoring guidance, risk category cut points, documentation routines, and trailer tool usage. Participants report that many agents enter detailed item and domain rationales in the statewide S^3 system, but that many assessments appear with only scores or minimal notes, constraining interpretability during transfers. They also report that most offices have access to an older (2018) manual while some still rely on even older guidance that defines scoring differently. As a result, those with identical scores can be mis-classified if supervised in different counties, especially when local policy ties supervision to the more conservative outcomes from a trailer instrument such as ODARA, DVI, IDA, or WRNA.

Study reliability findings indicate that LS/CMI scoring consistency is strong across assessors in this exercise. The total score demonstrated excellent agreement and individual items displayed high reliability, indicating agreement well beyond chance. These results suggest that, when assessors apply the instrument under proficiency training conditions, scoring rules can be implemented with a high degree of precision. The strength of agreement across both objective criminal history items and more interpretive needs-based items indicates that the instrument itself is capable of being scored consistently, and that Minnesota probation staff are proficient at overcoming the scoring challenges highlighted in focus groups. Although excellent reliability was demonstrated, ongoing monitoring remains important as small differences in scoring can shift individuals across locally defined risk thresholds. Continued proficiency testing and standardized training will help ensure that the high levels of agreement observed in this sample are maintained in routine practice.

When examining accuracy via risk score discrimination, the LS/CMI is found to provide borderline weak-to-moderate performance for general and violent reconvictions. When the trajectory of accuracy is examined across application years, the assessment's initial performance was rated as weak in 2012 but then rose steadily to the acceptable-moderate range by 2019. Accuracy declined slightly following the onset of the pandemic with no clear rebound through 2022. When assessment types were compared, initial assessments and reassessments are similar, but first assessments following an intrastate transfer are less accurate. These patterns match the

qualitative findings that continuity and documentation support better scoring, while transfers complicate interpretation.

Bias analyses identify systematic differences regarding predictive equity. Visual and model-based tests show overclassification for females, where they reoffend less often than men with the same LS/CMI score. Similar issues are identified for violent outcomes, where the rate of female overclassification grows as LS/CMI scores increase. For race and ethnicity, the models show both intercept and slope bias in key comparisons, where Non-White individuals are consistently overclassified compared to White individuals.

Domain-level results indicate that criminal history and antisocial pattern domains carry the strongest predictive value. However, several items within the criminal history domain show high false positive rates, and a set of needs items show weak and inequitable performance across demographic subgroups and contribute relatively little to risk classification. These domain and item patterns help explain why total scores discriminate better than many specific needs items, even when those needs remain important for case planning. As a result, after accounting for criminal history and antisocial pattern domain scores, the incremental contribution from other domains is small and varies substantially by subgroup.

We also note that the LS/CMI domains were conceptualized in the 1980s, at a time when psychometric analysis and scale development was limited. Notably, prior studies have identified the limitations of the Central Eight domains and proper methods of scale construction (Mei et al., 2023). For example, domains must possess a minimum of three items to be an adequate scale, where the Antisocial Pattern scale contains only two and many other scales do not provide sufficient construct validity. Furthermore, despite many studies identifying the limitations of their scales (Kitzmiller et al., 2022; Palmer & Hollin, 2007; Schmid et al., 2005; Schmidt et al., 2017), the LS/CMI developers and providers have yet to update the tool within the last 20 years.

Overall, county-specific risk categories separate outcomes in an increasing, stair-step pattern, indicative of an appropriate classification scheme. As a part of the current study, we applied the classification scheme used by the Department of Corrections to all agencies in the state. With this schematic, classification shifts more individuals towards low risk, yet distinctions in reoffending probability between risk levels are slightly diminished. These results underscore that category design is a major driver of how scores translate into daily supervision. Category sizes are sometimes skewed, with some counties having many clients clustered at the lowest

levels while others cluster clients in moderate risk categories. Changes to thresholds therefore shift supervision contact standards or programming requirements even when underlying scores are the same. In practice, this means that the probability that someone within a given risk category recidivates can be substantially different between counties, complicating the common language of risk assessment that stakeholders highlighted as a strength of the LS/CMI in focus groups.

Calibration results were central to understanding local contextual effects. The raw LS/CMI probabilities consistently overpredict observed outcomes for Minnesota probationers, which means the LS/CMI scores and risk categories do not adequately reflect probationers' risk, especially at higher risk levels. Following our recalibration, prediction error is reduced. If adopted, the changes outlined would improve risk level prediction. Unfortunately, under the standard provider thresholds, more than half of the population would be classified as High or Very High-Risk with fewer than 2% classified as Very Low. However, our recalibration efforts adjust the scale probabilities, so risk level categories more closely resemble a probation population, reducing the proportion of High and Very High-Risk individuals.[13]

Overall, these findings present a consistent picture across qualitative and quantitative components. Minnesota's probation agencies have developed a thorough approach to using the LS/CMI, applying it as a shared framework for risk assessment and case planning while navigating variation in training, resources, and local practice. The strong reliability results and improvement in predictive accuracy over time suggest that practitioners are able and committed to applying the tool with a high degree of consistency and extracting meaningful information in practice. In this context, the observed level of predictive performance reflects a system that has become adept at using risk assessment to support evidence-based case management. However, that progress is constrained by the tool's modest predictive performance, over-estimation of risk, and evidence of predictive bias. Moreover, the need for trailer assessments to address tool shortcomings, as well as the difficulty of applying a single standardized classification scheme across counties, further constrains how effective the LS/CMI can be within the operational reality of Minnesota's probation system.

[13] See Appendix B.

Recommendations

Based on study findings, we provide several recommendations for future use of risk-needs assessments, including the LS/CMI. Regarding the overall level of predictive accuracy, the LS/CMI's prediction of general reconviction currently meets minimum standards. This finding provides evidence that the tool can continue to be used for the Minnesota probation population. However, with recent declines in the predictive accuracy of the tool, it is recommended that, at a minimum, recalibrated cut points be used to adjust risk categories to better improve prediction with the Minnesota population. Further, we recommend continued efforts be provided for proficiency training and greater quality assurance provisions provided to retain interrater reliability.

With that said, qualitative findings demonstrate areas of concern with the functionality and usability of the tool, especially regarding confusion and inefficient item coverage. Further, the LS/CMI displayed concerning levels of sex and race/ethnicity biases. Our examination revealed few needs domains to be predictive and provided limited incremental improvement outside of the criminal history domain. As indicated, a common issue among underperforming tools is predictive shrinkage, where tools developed with differing populations and jurisdictional distinctions will demonstrate decreased effects when applied in a new location. As part of best practices, researchers/developers are encouraged to adjust item weights to improve prediction, reduce bias, and provide distinct models to more accurately assess severe outcomes (i.e. violence). We would recommend this practice be applied, although such efforts may not be feasible with the proprietary restrictions of the LS/CMI provider.

Further, when agencies identify risk assessment underperformance, it is not uncommon to consider switching to another tool. Common tools in use include the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and the Ohio Risk Assessment System (ORAS). Like the LS/CMI, these proprietary tools were developed in one location and would likely provide similar issues if items and weights were not optimized to Minnesota's population and agency needs.

However, more recently developed tools, such as the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) and the Static Risk Offender Needs Guide - Revised (STRONG-R) provide a greater number of items/responses and adjust response weights to both extend the range and accuracy of the tool. Further, assessments of bias can identify items that can

be removed from tool scoring, while retaining predictive accuracy. Finally, these tools have been developed to provide gender-response and violence prediction models, with the ability to remove/reduce the use of trailer tools used across Minnesota's jurisdictions. Therefore, if Minnesota is inclined to replace and adopt a new tool, we recommend adopting one with features that allow for greater accuracy, less bias, and the ability to reduce reliance on trailer tools that are not designed for the local population.

Finally, rather than adopting an existing tool, many agencies have developed local versions that are optimized to meet their needs. Given the existing software, training, and resources of the state, this may be a feasible solution to both improve accuracy and equity of risk and need assessment. Using existing assessment and routinely collected administrative records, a tool could be designed specifically for the Minnesota probation population. Given Minnesota's excellent reliability results, the LS/CMI is unlikely to provide better performance than the minimally acceptable results observed. Instead, improvements in performance are less likely to come from standardizing the LS/CMI, and more likely to result from adopting one of the recommended options and considering an assessment approach that is better aligned with Minnesota's needs.

Conclusion

The current study was designed to provide a comprehensive evaluation of how the LS/CMI functions for Minnesota's probation population. The qualitative findings identified several usability and implementation issues, particularly around standardization. Participants described ongoing confusion, differences in local scoring guidance, variation in the use of trailer assessments, and inconsistency in how counties translate scores into supervision decisions. At the same time, the quantitative findings show that the tool currently meets national standards for predictive validity, although at a borderline level, which is notable given the complexity of Minnesota's decentralized probation system. With this said, the tool struggles to predict reconvictions for violent offenses, and the bias and calibration analyses indicate that important limitations remain. The study also found that standard LS/CMI cut points do not appear to provide the best fit for Minnesota's probation population, while the recalibrated thresholds and probabilities, examined here, suggest that more locally optimized classification approaches could improve performance.

Considering these findings, we provided several recommendations. Under current national benchmarks, the LS/CMI demonstrates sufficient predictive accuracy to justify continued use if the state chooses to retain it. However, the broader findings suggest that maintaining the status quo may leave important gains in accuracy, equity, and practical utility unrealized. Alternative tools may offer improvements, particularly those with greater flexibility to predict both general and violent recidivism and to better align with Minnesota's population and supervision context. A locally developed or calibrated tool may offer the greatest long-term potential, as it could be designed around Minnesota's existing administrative data, operational needs, and resource environment. Such an approach could improve prediction, reduce overclassification, strengthen equity, and potentially reduce the costs associated with proprietary assessments, freeing resources for ongoing training, quality assurance, and implementation support.

REFERENCES

Andrews, D., Bonta, J., & Wormith, J. S. (No year specified.). *Level of Service/Case Management Inventory (LS/CMI™)* [Database record]. APA PsycTests.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct*. Routledge.

Barnoski, R., & Aos, S. (2003). Washington's Offender Accountability Act: An analysis of the Department of Corrections' risk assessment. *Olympia: Washington State Institute for Public Policy*.

Bonta, J., & Wormith, J. S. (2018). Adult offender assessment and classification in custodial settings. *The Oxford Handbook of Prisons and Imprisonment*, 397.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Council of State Governments Justice Center. (2023). *Justice reinvestment initiative in Minnesota: Improving supervision investments and outcomes*. https://csgjusticecenter.org

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2018). Performance of recidivism risk assessment instruments in US correctional settings. *Handbook of recidivism risk/needs assessment tools*, 1-29.

Duwe, G., & Rocque, M. (2016). A jack of all trades but a master of none? Evaluating the performance of the Level of Service Inventory–Revised (LSI-R) in the assessment of risk and need. *Corrections*, *1*(2), 81-106.

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.

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.

Kitzmiller, M. K., Paruk, J. K., & Cavanagh, C. (2022). Criminogenic risk score trajectories of justice-involved youth: An investigation across race/ethnicity. *Criminal Justice and Behavior*, *49*(9), 1342-1358.

Landis, J. R., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data*. Biometrics, 33(1), 159–174.

Mei, X., Hamilton, Z., Kigerl, A., Krushas, A., & Taxman, F. S. (2025). Best Practices for Neglected Assumptions: Multi-Site Confirmation of the MPACT-6. *Crime & Delinquency*, *71*(6-7), 2032-2060.

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.

Palmer, E. J., & Hollin, C. R. (2007). The Level of Service Inventory—Revised with English women prisoners: A needs and reconviction analysis. *Criminal Justice and Behavior*, *34*(8), 971-984.

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.

Schmidt, F., Hoge, R. D., & Gomes, L. (2005). Reliability and validity analyses of the youth level of service/case management inventory. *Criminal Justice and Behavior*, *32*(3), 329-344.

Schmidt, N., Lien, E., Vaughan, M., & Huss, M. T. (2017). An examination of individual differences and factor structure on the LS/CMI: does this popular risk assessment tool measure up?. *Deviant behavior*, *38*(3), 306-317.

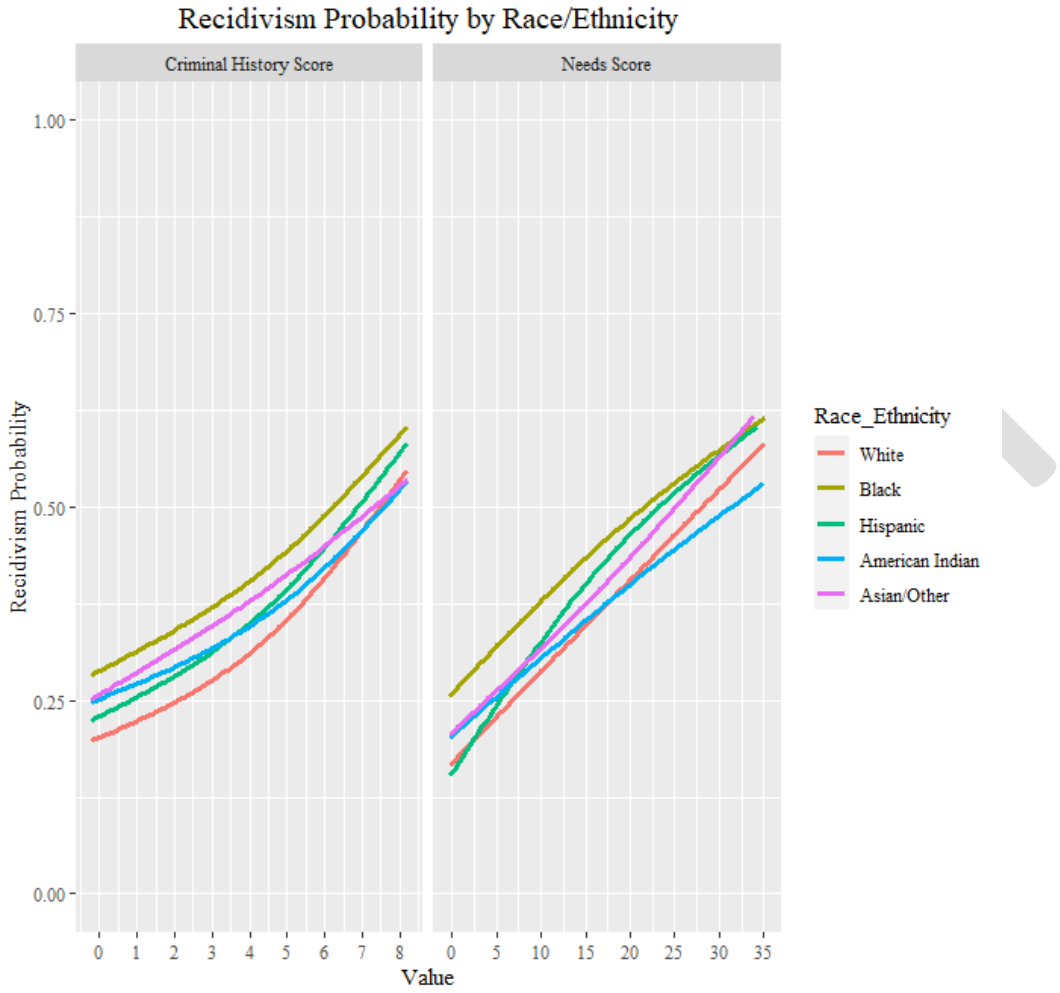
Singh, J. P., Kroner, D. G., Wormith, J. S., Desmarais, S. L., & Hamilton, Z. (Eds.). (2018). *Handbook of recidivism risk/needs assessment tools*. John Wiley & Sons.

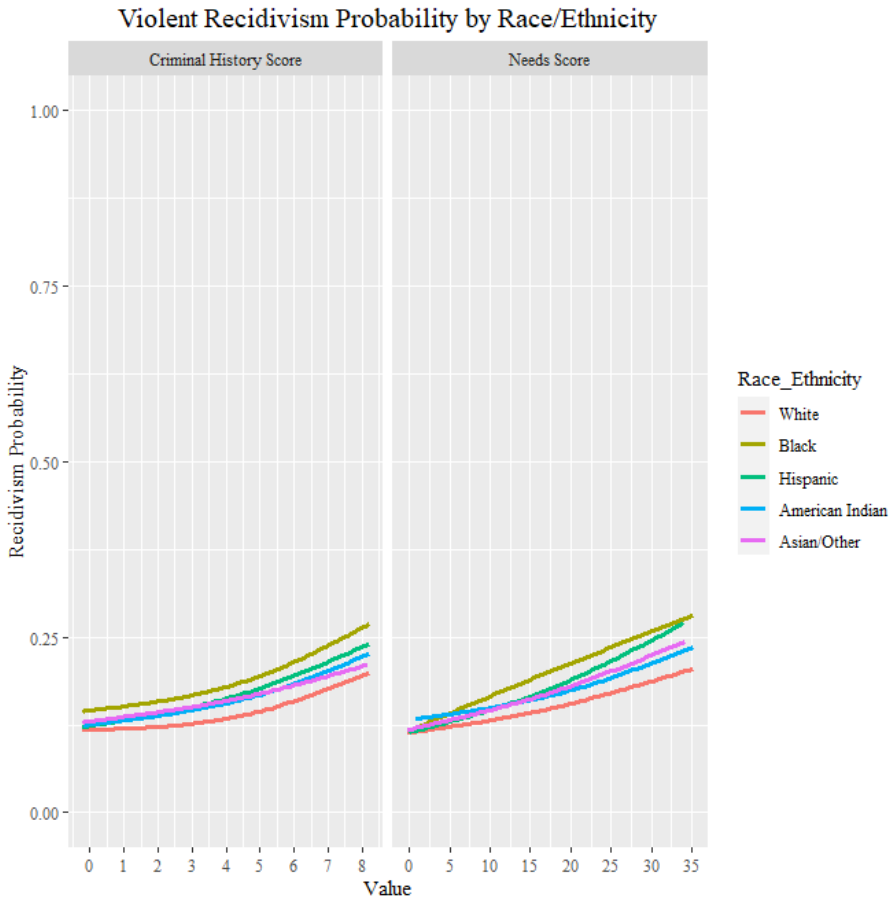
Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*(4), 680-712.

Wildermuth, J., Nonemaker, D., & Carlson, J. (2021, April). *Validation of the Level of Service Case Management Inventory (LSCMI) for assessing risk of re offense in Hennepin County*. Hennepin County Department of Community Corrections and Rehabilitation, Office of Strategy, Planning, and Evaluation.

APPENDIX

Appendix A: LS/CMI Performance by Race/Ethnicity (3-Year Outcomes)





Appendix B: Risk Level Categories Before and After Re-Calibration

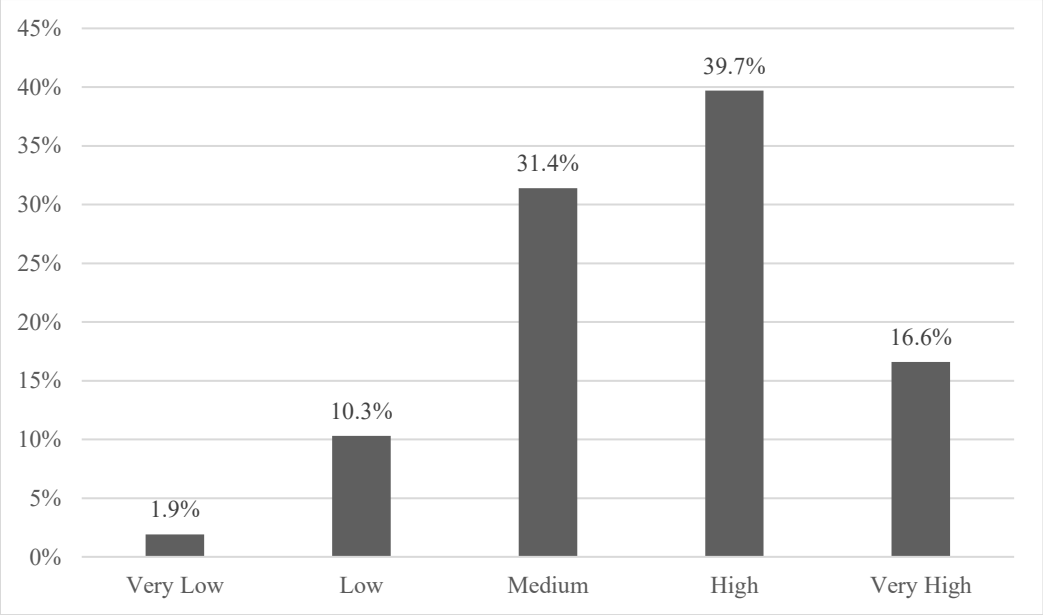
We demonstrate how recalibration affects risk level categories by displaying the distribution and odds ratios of risk-level categories using the standard LS/CMI classification scheme. We then do the same after recalibration, where *the expected rate of reoffending within each risk level classification remains the same*, but the cut-points are changed based on our recalibrated probabilities.

Since Minnesota does not have a standardized classification scheme, Figure 16 displays the distribution of classifications using the cut-points recommended by LS/CMI providers (MHS, 2006).[14] As shown in the figure, roughly 56% of probationers are classified as High Risk or Very High Risk. Moreover, less than 2% are classified as Very Low risk, and 10% classified as Low

[14] Scores of 0-4 are 'Very Low', 5-10 are 'Low', 11-19 are 'Medium', 20-29 are 'High', and 30+ are 'Very High'.

Risk. This distribution reflects the overclassification observed in the first calibration plot (Figure 14).

Figure 1B. Risk-Classification Distribution Using Standard LS/CMI Scheme



After recalibration, the distribution of risk classifications is closer to normal and representative of a probation sample (see Figure 2B). Notably, the proportion of cases that were classified as either High or Very High Risk falls from 56% to roughly 26%; a 30% reduction in the highest supervision levels. Moreover, a much higher proportion of cases are categorized as Very Low or Low-Risk. This recalibrated scheme would reduce supervision burdens experienced by Minnesota's probation agencies and is also a more accurate representation of the actual risk Minnesota's probationers represent to the community in terms of general reconviction risk.

Figure 2B. Case Distribution After Re-Calibrating Scoring Probabilities & Thresholds

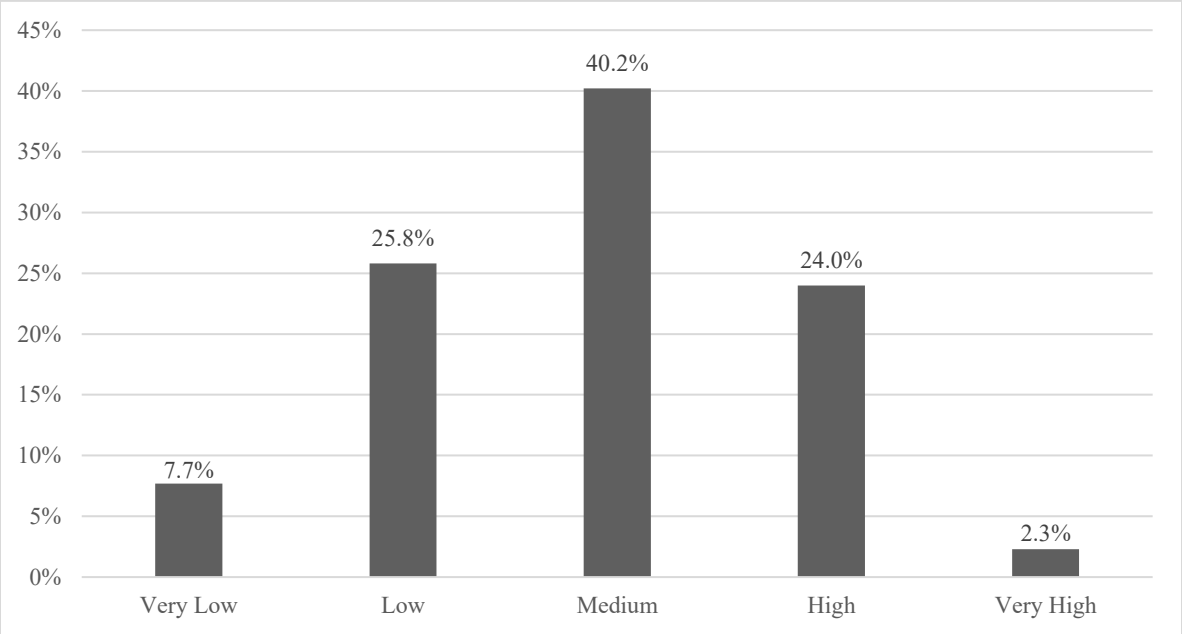


Table 1B displays the cut-points and odds ratios (with Very Low as the reference) before and after re-calibration. The revised cut-points expand the Very Low risk category from scores ranging from 0 to 4 to 0 to 8. However, looking at the ORs, the revised category provides better distinction between Low and Very Low in terms of reoffending probability. Moreover, the expected reoffending rate for the revised categories increases more with each subsequent risk level category. At the Very High-risk end (demarcated by a score of 30+ on the LS/CMI) it is expected that 50.8% of individuals will reoffend. After recalibration, however, this increases slightly to 54%. Thus, fewer individuals require high supervision intensity to maintain similar levels of reoffending.

Table 1B. Thresholds, Odds Ratios, & Expected Reoffending Before & After Re-Calibration

	Original			Re-calibrated		
RLC	Scores	OR	Reoffending Rate	Scores	OR	Reoffending Rate
Very Low	0-4	Ref.	7.9%	0-8	Ref.	11.0%
Low	5-10	2.0	15.1%	9-17	2.2	22.5%
Medium	11-19	4.1	26.1%	18-27	4.0	35.6%
High	20-29	7.4	39.0%	28-35	6.7	48.6%
Very High	30+	11.7	50.8%	36+	8.5	54.0%