Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) 3.0 Revalidation

Final Report

Submitted to

The Minnesota Department of Corrections (MnDOC)

On:

2025-08-31

Completed by:

Zach Hamilton Consulting, LLC.

EXECUTIVE SUMMARY

For Departments of Correction, risk and needs assessments (RNAs) are commonly used to identify appropriate supervision intensity and determine correctional programming eligibility. While many states adopt proprietary tools, developed elsewhere, in 2013 the MnDOC developed and implemented the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) assessment. Now in its third iteration, the MnSTARR was designed as a fully automated, gender-specific instrument, that is designed to predict both violent and felony reoffending. Prior to implementation, it is best practice to evaluate a RNA's performance. Further, the Minnesota Rehabilitation and Reinvestment Act (MRRA) was recently passed, where MnSTARR results are outlined to be used to be used to identify individuals eligible for early release and supervision abatement. The current study provided a validation of the MnSTARR 3.0 and examined the potential impact of the MRRA on public safety outcomes (i.e., recidivism).

Using a robust evaluation design, assessment, programming, and recidivism data were gathered from the MnDOC. A sample of 102,562 individuals who were once incarcerated and released to the community was gathered and divided into training and testing sets. The training set was used to construct four risk scores – Male Felony, Male Violent, Female Felony, and Female Violent – and the training set was used to evaluate the MnSTARR 3.0's predictive performance. Risk levels categories (RLCs) were also created, setting cut points within the Violent and Felony scores to identify Low, Medium, High, and Very High-risk levels categories. Specifically, the training sample consisted of those released between 2006 and 2016 (n=72,421; 70.6%) and the test sample was composed of 2017 through 2021 releasees (n=30,141; 29.4%). To evaluate MnSTARR 3.0, risk scores, RLCs, and potential sources of gender and race/ethnicity disparity were assessed. Finally, the impact of risk score and level changes and RLC adjustments were evaluated.

Findings revealed consistent and positive effects of the MnSTARR 3.0. Across dozens of predictive performance tests, the evaluation found:

- The four risk scores identified moderate-to-strong effects that far exceed national standards for post-conviction assessment tools (Desmaris et a., 2022).
- Risk levels demonstrate a "stairstep" effect, where each progressive risk level identifies substantially greater rates of felony and violent recidivism.
- Regarding disparity, across several examinations, there was minimal bias detected.
  - Gender differences were identified for the violent risk score but are accounted for through gender-specific risk scoring and risk level cut points.
  - Minimal effects were identified for test of racial/ethnic disparity.
  - Regarding risk levels, the MnSTARR 3.0 demonstrates relatively equal rates of felony and violent recidivism for race/ethnicity categories.
- MnSTARR 3.0's dynamic items allow for score reductions/increases, where consistent and substantial impacts were identified.
  - On average, individuals decrease their risk score from intake to release.

    - Those that decreased risk scores substantially reduced their recidivism likelihood post-release.
    - Individuals with medium-to-long sentence lengths possessed a greater opportunity to reduce their risk score.
    - Those that decreased a full risk level demonstrated moderate-to-large recidivism reduction effects.

In addition to study findings, recommendations are provided, including:

- An alternate method of creating MnSTARR risk levels,
- Further expanding the collection of dynamic items, and
- Develop a study plan to examine the effects of the MRRA initiative

Overall, the review of the MnSTARR 3.0 indicated a tool that is designed to predict post-release recidivism outcomes for individuals supervised by MnDOC. Using criminal history, correctional events, and programming information, the tool relies on weighted item responses to optimize recidivism prediction for its local, Minnesota population. Findings indicate that the MnSTARR is one of the most effective tools utilized by a state department of corrections (Singh et al., 2018). Moreover, the totality of findings indicate that the MnSTARR 3.0 is an evidence-based assessment that is ideal for the MnDOC population. With the multiple versions developed to date, it is likely that the tool will continue to improve and adapt to fit agency and individuals' supervision and programming needs.

Table of Contents

1.0 INTRODUCTION

For over four decades, the use of risk and needs assessments (RNAs) have become common tools for providing supervision classification and assisting with correctional programming recommendations. Recognized as an evidence-based practice (EBP) (Taxman, 2018), these tools provide a mechanism for standardization, reducing biases and, in turn, more efficient uses of correctional resources (Lowenkamp & Latessa, 2004). As a result of their proven effects, most state departments of correction use some form of RNA to predict recidivism upon reentry from prison and while on community supervision (Desmaris & Singh, 2013). Further, many states have mandated the use of RNAs to guide supervision intensity and program prioritization (Mackey et al., 2022).

Regardless of an agency's motivation, RNAs are either adopted or created locally. While adopting a tool can be quicker and perceived to consume less labor and resources to start, recent research demonstrates that tools created locally provide greater accuracy. Specifically, RNAs work best for a population in which they were originally designed to predict and, when created locally, the items and responses can be tailored to the needs of the agency (Duwe, 2024).

However, not all RNAs are effective, or accurate predictors of recidivism. Ineffective tools do not contain sufficient information and may miscalculate risk. Furthermore, many RNAs are built with predominantly White-male subjects, where prediction of risk may be biased for females and persons of color (POCs). Recent standards have been developed (Desmaris et al., 2022) that outline the need to validate assessments prior to, and following, deployment. Completed routinely, evaluations of RNAs effectiveness assure that tools meet national standards for continued use.

In 2013, the MnDOC developed the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) assessment (Duwe & Rocque, 2019). The tool has been updated twice, with Version 2.0 deployed in 2016, and the current 3.0 version was implemented in 2025. While previous versions of the MnSTARR were internally validated, the MnDOC sought an external validation to provide added objectivity of the 3.0's effects. Hence, in November of 2024 a Request for Funding Proposal (RFP) was posted and awarded to Zach Hamilton Consulting, LLC. The scope of work includes two project phases – 1) validation of 3.0 and 2) dynamic impacts of the tool and recommendations for future use. The current report provides the findings from Phases 1 and 2.

1.1 Background

Modern correctional risk assessments were developed over the course of the last 40 years and perceived to span several generations (Andrews et al., 2006). First-generation assessment involved professionals using their experience to make decisions regarding which offenders were more likely to recidivate. Second-generation assessments introduced actuarial tools, using quantifiable factors to assess risk to improve validity and reliability of prediction. Actuarial tools included risk factors that could, ideally, be scored objectively, with risk factors added together to arrive at an overall score, which have consistently shown improved accuracy compared to professional judgement (Bonta & Andrews, 2007; Duwe & Rocque, 2018). Third-generation assessments were built upon the success of second-generation tools to include dynamic, or changing, factors that point to areas of possible intervention. Finally, fourth-generation tools expanded the scope of third generation assessments to include case management guidance (Andrews et al., 2006).

Post-conviction risk and needs assessments (RNAs) aim to serve two important functions. First, they provide a method of standardizing the classification/ranking of individuals' risk of recidivism, relative to the correctional population supervised. Second, the results of assessment tools are then used to outline supervision (i.e., early release and supervision intensity) and prioritize programming (treatment and services), which are then utilized to target needs and reduce risk of recidivism over time. Yet, despite their prominent use, the methods of developing RNAs are rarely described to users and administrators. In the next section I describe how assessments are made to provide a basis of understanding regarding the need for, and components of, the MnSTARR 3.0 validation.

1.2 How assessments are made

All assessments are composed of a series of items/questions, in which responses are gathered from an individual's criminal history record and/or via a semi-structured interview. Typically, an assessment is created in the following seven stages:

1. Gather a pool of potential items that predict recidivism
2. Administer to justice involved individuals
3. Track who reoffends
4. Select/Weight items that predict reoffending
5. Set cut points to create risk levels
6. Deploy assessment
7. Evaluate tool – predictive validity and bias

Typically, an RNA developer will consult with a group of Subject Matter Experts (SMEs) and review prior criminological literature to identify potential items and responses (Stage 1). The items are then administered to a sample of subjects from the agency's target population, which is commonly referred to as the RNA's *development sample* (Stage 2). Using the development sample, individuals are tracked using the developer's, or an agency's, definition of recidivism (i.e. new felony convictions within three years) to see who reoffends (Stage 3).

In Stage 4, assessment responses and recidivism data are analyzed, where RNA developers use statistical analyses to select items that predict recidivism, and provide greater weights/values to responses that are stronger predictors. Items are provided values, or scores, that when summed, create a composite risk score. Larger risk scores indicate a greater likelihood of recidivism, and the collection of a development samples scores creates a *risk score distribution*. Using this continuum of risk scores, in Stage 5 the development sample is then divided using thresholds, or cut points, that identify risk levels (i.e., high, moderate, and low risk). These risk levels are commonly used by correctional agencies to determine supervision intensity and program eligibility, where greater intensity and programming are reserved for higher risk individuals.

Following the creation of risk levels, staff are trained to administer the assessment, and the tool is deployed (Stage 6). The RNA is then administered for three to five years, dependent on the agency's recidivism definition. After enough subjects have been scored on the assessment and possessed sufficient time in the community to assess recidivism, the tool can be evaluated for predictive validity and bias (Stage 7).

It is important to note that research has found that tools are more efficient when risk scores can *discriminate* between low and high risk. Meaning, larger risk scores and higher risk levels are associated with greater rates of recidivism and smaller risk scores, and lower levels indicate with reduced rates of recidivism. Further, effective tools predict equally for males and females and all races/ethnicities.

1.3 The MnSTARR

The MnSTARR is a post-conviction risk tool and was originally developed with a sample of males and females released from prison to the community (Duwe, 2014). It is intended to assess risk of recidivism, scoring and classifying individuals into risk levels, which are then used to help guide decisions relating to institutional programming and post-release supervision. Notably, the MnSTARR was developed with a relatively large sample (N=12,475) of individuals released from prison to the community between 2003 and 2006 and predicts recidivism using the MnDOC's definition – new convictions within three years of release. Logistic regression models were used to create a multi-band assessment tool, predicting five different offense types – nonviolent, felony, nonsexual violence, first-time sex offending, and repeat sex offending. Models were computed separately for samples of males and females, using multiple logistic regression models to select and weight each assessment item. Bootstrap resampling methods were then used to evaluate the tool's predictive validity, with findings exceeding industry standards for post-conviction risk assessment tools (Desamaris et al., 2022; Duwe 2014).

The MnSTARR was manually scored until 2016, at which time a new assessment version was implemented as a fully automated tool (MnSTARR 2.0). Automating the MnSTARR required a process of selecting and weighting objective items that were obtained electronically from administrative databases. To confirm the accuracy of the updated tool, the predictive performance, as well as gender and race/ethnic disparities were evaluated among 8,997 individuals assessed just prior to prison release. Again, predictive validity estimates exceeded industry standards (Duwe & Rocque, 2021), demonstrating "excellent"- predictive performance (see Rice & Harris, 2005).

More recently, the MnSTARR 3.0 was developed and validated with more than 100,000 individuals released from Minnesota prisons between 2006 and 2021. As with the prior version, the 3.0 is a fully automated, gender-specific instrument, and designed to predict violent and felony recidivism of individuals released from prison to the community. In Phase 1 of the project, Version 3.0 was evaluated. In Phase 2 the tool's dynamic effects were studied. Specifically, the evaluation identified the 1) amount of change in risk scores from intake to release and 2) the reduction in recidivism as result of positive behavior and program participation. In the next section the study methodology is provided, describing the evaluation procedures.

2.0 METHODOLOGY

2.1 Data preparation

The study data was gathered and prepared for analysis. A video conference was convened to discuss the structure of the data. Retrospective MnSTARR assessment and recidivism data was provided by the MnDOC. Data were structured in a rectangular database, where each row represented an MnSTARR assessment. Specifically, a unique identifier was provided that represented a person and their prison release date, along with columns representing each tool item.

Recidivism was defined as a new felony conviction within three years of release. In addition, a second outcome, new violent offense conviction three years following release was measured. All subjects were released between 2006 and 2021. Additionally, subjects sex (male or female) and primary race/ethnicity (White, Black, American Indian, Hispanic, or Asian) measures were provided. MnDOC subject matter experts (SMEs) were consulted to obtain data documentation and background publications outlining the development of MnSTARR 3.0 and prior versions. Next, descriptive statistics (means & frequencies) of each measure were computed.

2.2 MnSTARR design details

MnSTARR 1.0 was developed in 2013 as a *multi-band* tool. Meaning that a general pool of items is used to select and weight multiple models, including nonviolent, violent, felony, and nonsexual violent recidivism. Further, gender-specific, or separate models, were computed for males and females, for a total of eight bands. While the 1.0 version demonstrated strong predictive performance (Duwe & Rocque, 2019), developers perceived greater predictive accuracy, increased assessment capacity, as well as reduced bias and labor requirements with an automated tool. In 2016 the MnSTARR 2.0 was created with a larger development sample, a more sophisticated statistical procedure to select and weight items, the first-time sex offense band was removed, and an additional 50 items were added to the now automated scoring schematic. As with the previous version, dynamic items were included regarding program participation, visitations, Security Threat Group (STG) involvement, suicidal tendencies, and prison misconduct.

As described in technical documentation (Duwe, 2025), the MnSTARR 3.0 is a multi-band, gender-specific, fully automated tool, gathering item response data from the state's criminal history repository to populate offense history items. The Correctional Operations Management System (COMS) – the MnDOC's centralized database was also used to gather measures on study subjects' characteristics (e.g., gender, age, and marital status), institutional behavior (e.g., misconduct and gang affiliation), and participation in programming (e.g., post-secondary education, substance use disorder treatment, and cognitive-behavioral therapy). However, the 3.0 version was designed to have more transparent scoring, which was perceived to increase understanding of the tool's scoring for both residents and staff, as compared to the previous version. Second, this updated version used a very large development sample, including over 100,000 subjects. Finally, only the violent and felony models were retained, and when completed for both males and females comprised four total bands[1].

As described, when developing an RNA, items that predict recidivism are selected and weighted to produce the composite risk score. However, to ensure that the tool is effective, the risk score must be tested on a separate sample of individuals not used at the selection and weighting stage. This process is termed *split-sample cross-validation*. The sample in which the items are selected and weighted are referred to as the *training sample* and those in which the risk score is validated, is termed the *test sample*. To ensure that the model developed on the training sample is effective in future samples, it is ideal to reserve a sample of the most recent releases for the test sample. Further, the training sample requires a greater number of cases to ensure sufficient response variation and predictive pattern recognition in the item selection and weighting stage. Thus, the training sample was composed of subjects released between 2006 and 2016 (n=72,421; 70.6%) and the test sample was composed of those released from 2017 through 2021 (n=30,141; 29.4%).

For each of the four bands, logistic regression models were computed to select and weight items in the training sample. As described in the development stages, a large pool of potential predictors is first gathered and then items that do not predict recidivism are removed from said pool. The MnSTARR developers used a sophisticated statistical procedure, bootstrap variable selection, to remove non-predictive measures. Using the training sample, this procedure draws 1,000 samples, with replacement, computing a logistic regression model for each sample draw. Items are only retained if identified as predictive in at least 50% of the 1,000 draws. This process ensures that only consistently predictive items are retained in the final model. Across the four models, 23 items were selected for the male violent, 27 for male felony, 19 for female violent, and 20 items were selected for female felony recidivism model.

Four final logistic regression models were computed for each of the bands, which produced item coefficients. These coefficients represent the log-odds of recidivism and were ultimately converted into odds ratios, where a 1-point increase in the response weight was associated with 5 percent increase in the odds of recidivism[2]. After assigning response scores for each item, the training sample participants scores on all items and bands were summed to produce total felony and violent risk scores.

Using the scoring schematic established in the training sample, the four risk scores are computed using the test sample. Thresholds, or cut points, were then established to create risk level categories. The developers opted to set cut points based on risk scores predicted recidivism probability. Notably, cut points are established separately for males and females using both the violent and felony risk scores. Four risk levels were established, where violent or felony risk scores associated with the top 15 percent, or 85 to 100 percent of recidivism likelihood were classified as *Very High*,

[1] It is noted that the sex offense recidivism score for men is computed through a separate tool (MnSOST-4), which is still used as part of risk classification. For women, the ADVISOR is used for sex offense risk classification.

[2] The MnSTARR 3.0 documentation also indicates that additional point value adjustments were also completed to minimize racial biases observed with in the outlined scoring procedures.

those in the upper 79 to 60 percent of recidivism likelihood were classified as *High*, those in the next 59 to 31 percent of recidivism likelihood were identified as *Medium*, and those in the lower 30 percent recidivism likelihood were classified as *Low-Risk*.

2.3 Predictive validity of the MnSTARR 3.0

To evaluate MnSTARR 3.0's predictive validity, risk scores and levels were evaluated. individuals' computed scores at release were provided, which were used to assess their risk upon release to the community. Recidivism events were identified if a felony or violent felony conviction were recorded within three years of release.

However, validity analyses cannot be completed on the full sample, as cases used to select and weight items for the MnSTARR (i.e., training sample) should not be used to validate the tool. If training cases are used in the validation process, findings may demonstrate "overfitting" or the predictive validity statistics are observed to be artificially high. Therefore, predictive validity analyses were completed twice, once with the MnSTARR 3.0 development/training sample (2006-2016 releases), and again using subjects assessed and released between 2017 and 2021, or the test set. Analysis conducted on the training set are completed to confirm internal validation findings of MnDOC researchers. However, predictive validity is only concerned with the metrics completed on the test set.

To assess predictive validity, several statical metrics were evaluated. Specifically, to evaluate the four risk scores of the MnSTARR 3.0, the 1) accuracy, 2) calibration and 3) discrimination and 4) a combined metric were computed. *Discrimination* is an assessment's ability to rank individuals or groups with different scores/levels of risk. The industry standard metric for assessing discrimination is the Area Under the Curve (AUC), which balances the errors of false positive and false negative with true positives and true negatives of a scale's prediction of recidivism. The AUC ranges from 0.5 to 1.0. The magnitude of predictive validity is assessed using an industry standard effect size scale, where values of 0.50-0.55 are 'negligible', 0.56-0.63 is 'small', 0.64-0.70 is 'moderate', and 0.71 and above is 'strong' (Rice & Harris, 2005). Industry standards have identified that tools with an AUCs above 0.64 are acceptable for use in post-conviction correctional populations (Desamaris et al., 2022).

Accuracy describes the overall correctness of predictions, or how closely an assessment's prediction aligns with the actual outcomes, in this case recidivism. Accuracy (ACC) statistics are computed by calculating the number of correct predictions divided by the total number of predictions. *Calibration* measures the agreement between predicted scores (or probabilities) and observed outcomes. A well-calibrated model provides probabilities that reflect the true likelihood of outcomes. The Root Mean Square Error (RMSE) is a widely used metric for evaluating calibration performance. The RMSE measures the average magnitude of the error between predicted and observed values. To place calibration values in the same range as the ACC and AUC the value is subtracted from 1.

A combined metric, the SAR (squared error, accuracy, ROC [receiver operating characteristic]) is a combined measure of discrimination, accuracy and calibration, and its formula is: (AUC + ACC + (1

– RMSE))/3 (Caruana, et al., 2004). In previous correctional research that has used the SAR, values ranged from a low of 0.62 to a high of 0.90. Ultimately, the SAR provides a comprehensive evaluation of the assessment's discrimination, accuracy and calibration.

Notably, the Council of State Governments Justice Center has worked with a group of researchers, practitioners, and policymakers to develop practical guidelines that advise criminal justice agencies on the use of risk and needs assessments (Desamaris et al., 2022). This report outlines the need for test sample AUCs to exceed a value of 0.64, deeming a tool 'appropriate' for further use with post-release populations.

2.4 Examine risk levels & cut points

To assess the validity of MnSTARR, the risk levels must also be evaluated for their predictive performance. There are multiple scores used to predict general and violent recidivism, and each are computed separately for males and females. Cut points, or thresholds, were then set, dividing the scoring continuum into categories – Low, Medium, High, and Very High-Risk. Using the test set, risk category proportions were examined, and recidivism of each category were computed. While the underlying risk score provides an indication of discrimination via the AUC, risk level categories are further examined. Specifically, the *stairstep effect*, is identified when each successive risk level category indicates a greater rate of recidivism. To measure the magnitude of predictive discrimination, odds ratios (ORs) were computed, comparing individuals classified as Low-Risk to higher risk categories. OR ranges have been established as 1.0 no effect, 1.44 to 2.46 "small", 2.47 to 4.24 "moderate", and 4.25 or above is "large" (Chen, et al., 2010).

2.5 Evaluation of predictive disparities by sex and race/ethnicity

Further, composite scores and risk levels may vary by sex and race/ethnicity, where disparities in the tool's prediction must also be examined. Demographic indicators were used to divide the population into sub-samples based on their identified sex or race/ethnicity. Disparity analyses were computed based on the guidelines provided in *The Standards* for psychometric testing (AERA, APA, & NCME, 2014).

First, accuracy, calibration, and discrimination metrics were compared across sub-groups. Scatter plots were computed to examine trends across the four risk score distributions. Next, intercept and slope bias were measured via a logistic regression, with three coefficients – a) risk score, b) group (e.g., White/Non-White), and an c) interaction of risk score and group. With the outcome of recidivism, it is anticipated that the risk score will demonstrate a significant prediction ($p<.05$). However, if intercept bias is present, the group coefficient will also be significant. Further, a significant interaction term demonstrates slope bias. It should be noted; it is not uncommon for risk models to demonstrate intercept and/or slope bias.

Furthermore, the concern of disparity often centers on overclassification, where a particular group is identified to recidivate at a lower rate, despite possessing the same risk score. Researchers evaluate algorithmic fairness by examining *error rate balance* (Hamilton, 2019). Overclassification is a particular concern for the ‘higher risk’ portion of the distribution, where the False Positive Rate (FPR) is used to measure the proportion of non-recidivists’ cases that are incorrectly classified as high risk. In contrast, the Positive Predictive Value (PPV) measures the proportion of recidivist predictions that are correctly identified as high risk. While industry standard criteria for FPR and PPV rates are not yet developed, prior studies have identified that when comparing groups, differences greater than 5 percent are concerning, 10 percent are worrisome, and 20 percent are troublesome (Hamilton et al., 2024a; 2024b).

To examine race/ethnicity and gender bias, scatterplots were computed, which provide a visual representation of intercept and slope bias, and the range of scores in which overclassification reaches concerning levels. Next, separate logistic regression models were computed for males and females, with further breakdowns by race/ethnicity groups (i.e., White, Black, Hispanic, Native American, Asian). In addition, both FPR and PPV rates were computed by gender and race/ethnicity subgroup.

2.6 Examine the design and use of the MnSTARR 3.0 within the RNR assessment system

The RNR assessment system is designed to use both static and dynamic items to produce a composite that determines likelihood of reoffending. The MnSTARR is unique, in that scoring includes misconduct, idle time, and program participation. Therefore, individuals are assessed at prison intake and can both increase and decrease their risk score prior to release. The dynamic nature of the tool, and its potential to motivate individuals to participate in programming and display positive behavior is important and will be reflective in policies used to support the Minnesota Rehabilitation and Reinvestment Act (MRRA).

Briefly, the MRRA was signed into law in 2023, where act provisions are anticipated to be implemented with the MnSTARR 3.0 roll out. The MRRA had three key provisions. First, based on MnSTARR results, individuals will work with their case manager and a multidisciplinary team to identify programming that will effectively reduce their risk of future reoffending. Using the results of these tools, the IRP outlines programming in which, if the individual participates, it is anticipated that said programming will decrease their risk and needs upon reassessment. Under the MRRA, assessment scores are used to identify lower risk subjects for *early release* and those classified in higher risk categories are prioritized for programming and intensive supervision upon release. At prison intake, an individual’s MnSTARR score is used to identify their level of risk. The early release provisions of MRRA are anticipated to be tied to MnSTARR 3.0 risk levels.

As part of Phase 2, risk reduction and scoring were examined. First, individuals’ risk levels at intake were calculated for items that could be changed by positive behavior (i.e., reduced of idle time, lack of misconduct, and increased structured activity days), programming participation (i.e. education, Moving On, Substance Use Disorder Treatment, EMPOLY, Prison Fellowship Academy [PFA], Challenge Incarcerated Program [CIP], and work release), and prison visitations[3]. Comparisons were made between MnSTARR 3.0 and calculated intake scores, examining differences in scores' means.

Next, *change scores* were calculated by subtracting MnSTARR 3.0 and calculated intake scores, which were further grouped into *change levels* by classifying individuals that increased, decreased, or had "no change" in their scores from intake to release. AUCs were calculated for intake and change scores to identify potential reductions in predictive accuracy and impact of risk reduction on recidivism outcomes. Significance tests and odds ratios were also computed to examine change scores and levels' ability to identify the amount and magnitude of risk reduction effects on recidivism post-release.

To examine the potential impact as it relates to MRRA, the proportion of subjects that reduced their MnSTARR risk levels were assessed. Further, change levels were compared by time spent in prison to identify the group of individuals most likely to benefit from MRRAs early release, or those that have sufficient time to complete programming and reduce their risk level. Finally, based on these findings, potential adjustments to cut points are provided to better incorporate the MnSTARR risk level changes and early release provisions of MRRA.

3.0 RESULTS

In this section study findings are presented. Using the data provided by the MnDOC we began by examining sample descriptives. Next, we describe findings from the predictive validity analyses. We then describe disparity analyses, examining predictive validity distinctions between gender and race/ethnicity sub-groups by risk score and level. Finally, the dynamic risk score and level changes between intake and release are examined.

3.1 Sample descriptives

First, MnSTARR 3.0 sample descriptives were computed, which were broken down by test and training samples. Findings are provided in Table 1. Ideally means and frequencies are relatively equal across training and test samples, indicating population stability and relative consistency when using a score developed in one sample and applied to the other.

When reviewing the MnDOC prison population, roughly 10% are female. Regarding race/ethnicity, roughly 50% are White, slightly over 30% are Black, around 11% are American Indian, nearly 5% are Hispanic, and roughly 2% are Asian. Across these two demographic indicators, there were no substantial differences between test and training samples.

Regarding recidivism, roughly 37% of the sample was reconvicted of a new felony within three years of release. When considering violent recidivism only 17% of the test and 18% of the training sample

[3] It should be noted that Prison Time Served is also measured and is dynamic, theoretically, this item should not change because of positive behavior or program participation and hence, was not included in the intake calculation.

were identified to be reconvicted of this type of offense. Again, distinctions between the test and training sample are not substantial.

The composite MnSTARR 3.0 felony and violent risk scores were also compared. Interestingly, the training sample's mean risk scores were roughly 3 points lower than the test sample, on average, for male felony (Mean = 56.9 vs. 59.3), female felony (Mean = 43.9 vs. 47.4), male violent (Mean = 37.0 vs. 40.2), and female violent (Mean = 9.3 vs. 12.9). While a three-point difference is not terribly concerning, it is interesting that descriptives findings provide a consistent indication of the latter, testing samples demonstrating a higher average rating of risk on the MnSTARR 3.0.

Further, MnSTARR risk levels were provided for the test sample. Risk level categorization indicates descending proportions, where Low-Risk was found to be the largest category (36.1%), followed by Medium (33.0%), High (21.4%), and Very-High (9.5%). Notably, the sample provided subject's risk based on the assessed score prior to release; therefore, risk categories may represent greater proportions of higher risk individuals for assessments collected at prison admission.

When examining the number of prior convictions, again, the test sample identified a larger average than the training sample (14.9 vs. 13.4). While distinctions were small, typically less than 1%, the test sample possessed a larger average rate of all offense types and number of prison admissions. Regarding index offenses, the test sample subjects indicated a greater proportion of all types, apart from property offenses. Collectively, these findings indicate that the more recently released test sample possesses a greater number and severity of offenses than the test sample.

Regarding prison time, the test sample possessed two extra months, on average, compared to the training sample (16.4 vs. 14.1). When examining misconduct there is a substantial difference between samples, where nearly two-thirds of the training sample (66.3%) had no prior misconduct, compared to just over half of the test sample (55.3%). Further, eight percent fewer testing sample subjects received prison visitations (32.9% vs. 40.5%).

Reviewing subjects' mental health issues, a two-percent greater rate of self-injury concerns (11.7 vs. 9.1) and a 6% difference for suicidal concerns (23.5 vs. 17.3) when comparing the test to the training sample. Again, the test sample demonstrates a slightly elevated risk around mental health concern.

Regarding programming, the testing sample possessed a higher rate of participation. Roughly 6% more testing sample subjects earned a post-secondary degree in prison compared to the training sample (17.7% vs. 11.5), 5% more participated in the EMPLOY program (7.9% vs. 2.4), 3% more participated in the Prison Fellowship Academy (5.2 vs. 2.2), 7% more completed substance use disorder treatment (20.3% vs. 12.9%), and additional 30 hours, on average, of structured activity time (344.6 vs. 304.6).

Table 1. MnSTARR 3.0 Sample Descriptives by Test & Training Samples (N = 102,562)

Item	Test (n= 30,141) Mean(SD)/%	Training (n=72,421) Mean(SD)/%
Gender		
Female	10.9	10.1

Male	89.1	89.9
Race/Ethnicity		
White	49.7	50.4
Black	30.7	32.3
American Indian	12.5	10.5
Hispanic	4.7	4.9
Asian	2.3	1.8
Recidivated Felony (3-year)	37.8	36.9
Recidivated Violent (3-year)	17.0	18.0
Felony Score		
Male	59.3 (21.4)	56.9(21.0)
Female	47.4 (19.6)	43.9 (20.5)
Violent Score		
Male	40.2 (23.9)	37.0 (21.6)
Female	12.9 (21.8)	9.3 (20.1)
Risk Levels		
Low	36.1	
Medium	33.0	
High	21.4	
Very High	9.5	
Convictions		
Total	14.9 (10.7)	13.4 (11.7)
Felony	5.0 (3.7)	3.1 (3.0)
Violent	1.96 (2.2)	1.4 (1.9)
Drug	1.3 (1.8)	0.8 (1.3)
Violations of Order for Protection (VOFP)	0.5 (1.2)	0.2 (0.7)
Driving While Intoxicated (DWI)	0.6 (1.1)	0.6 (1.1)
Failure to Register (FTR)	0.1 (0.5)	0.1 (0.5)
Prison Admission	3.44 (3.1)	3.0 (2.6)
Index Offense		
Person	29.4	18.8
Sex	11.0	8.3
Drug	25.4	19.9
Property	13.8	17.0
Driving While Intoxicated (DWI)	6.5	5.9
Other	14.0	10.8
Unauthorized/Idle Assignments	1.1 (2.1)	0.8 (1.7)
Prison Time (in months)	16.4 (27.5)	14.1 (21.4)
Security Threat Group (STG)	24.9	26.9
Major Mental Illness	3.4	4.0
Self-Injury Concern	11.7	9.1
Mental Health Criteria	0.6 (0.9)	0.5 (0.9)
Suicidal Concern	23.5	17.3
Infraction Behavior		
Discipline Convictions	3.8 (11.3)	3.4 (11.2)
Total Segregation Misconducts	2.8 (8.6)	2.4 (7.3)
Total Violent Misconducts	0.2 (0.9)	0.2 (0.8)
Any Violent Misconducts	0.1 (0.2)	0.1 (0.2)

Serious and Frequent Misconduct	9.7	9.6
No Misconduct	55.8	66.3
Release Age (in years)	35.8 (10.1)	34.2 (10.1)
Intake/Discharge Type		
Released to Supervision	94.7	87.3
New Commitment	35.8	39.8
Parole Violation	31.3	27.8
Release Violator	32.9	32.4
Education Level		
Less than Secondary Degree	33.7	37.0
Secondary Degree	60.8	59.2
Post-Secondary Degree	5.6	3.8
Education in Prison		
Earned Secondary Degree in Prison	14.3	14.2
Earned Post-Secondary Degree in Prison	17.7	11.5
Earned Education Degree in prison	32.1	25.7
Education Classes in Prison	3.5 (6.7)	3.2 (4.8)
Visitations		
In-Person Visits	32.6	40.4
Video Visits	2.2	0.3
Any Visits	32.9	40.5
Programming		
Work Release	7.8	7.9
EMPLOY	7.9	2.4
Prison Fellowship Academy	5.2	2.2
Challenge Incarceration Program	7.1	5.2
Completed Substance Use Disorder Treatment	20.3	12.9
Moving On	0.3	0.3
Structured Activity in Hours	344.6(671.8)	304.6(534.7)

Next, the four risk models' scoring items were compared across the training and test samples. The Male Felony and Violent model comparisons are provided in Table 2. A total of 23 items were selected for the Male Violent and 27 for Male Felony model.

The male test and training samples were similar, yet consistent findings were indicated. The training sample identified fewer high-risk items, as indicated by the greater proportion of the training sample indicating 'Less than 15' total convictions, '0' violent convictions, '1' felony convictions, '0' VOFPs, '1' prison admission, yet fewer misconducts, self-injury concerns, person and drug offenses, and program participation (e.g., CIP, EMPLY, SUD Treatment, and PFA).

Table 2. Assessment Descriptives – Male (N=91,942)

Item	Male Violent Test (n=26,854)	Male Violent Training (n=65,088)	Male Felony Test (n=268,54)	Male Felony Training (n=65,088)
Total number of convictions				

Less than 15	56.5	64.3	5.5	10.1
15-24	27.5	22.0	11.7	16.2
25-29	6.4	5.2	14.2	15.2
30-34	4.1	3.1	29.9	26.7
35 or more	5.4	5.4	24.7	19.7
Violent Convictions				
0	26.5	40.3		
1	23.3	23.6		
2	17.6	14.1		
3	12.0	8.6		
4	7.9	5.4		
5 or more	12.6	8.0		
Felony Convictions				
1			11.6	39.0
2-3			29.1	31.4
4-6			34.0	18.8
7-10			18.3	7.6
11-20			6.5	3.0
21 or more			0.5	0.2
VOFP				
0	76.7	87.2	78.3	88.0
1	16.6	10.4	10.3	7.0
2 or more	6.8	2.4	11.4	5.0
FTR Convictions				
0			91.3	91.5
1			5.2	5.3
2 or more			3.4	3.2
DWI Conviction	33.3	32.0		
0			66.7	68.0
1			18.1	17.0
2			7.7	8.1
3			7.4	6.9
Prison Admissions				
1	32.2	35.2		
2-3	30.5	35.2		
4-5	16.3	15.5		
6-7	9.7	7.3		
8-9	5.4	3.5		
10 or more	5.9	3.2		
Prison Admissions				
1			34.6	36.8
2-4			39.4	43.8
5.7			15.7	13.0
8-10			6.4	4.3
11+			3.9	2.1
Unassigned/Unauthorized idle				

0 assignment	60.2	65.7		
1 idle assignment	17.8	16.6		
2-3 idle assignment	12.5	10.9		
4 or more idle assignments	9.6	6.8		
Serious & frequent misconduct	90.3	90.4	9.7	9.6
Misconduct			44.2	33.7
Active Security Threat Group	27.4	29.4		
Self-Injury Concern			11.7	9.1
Mental Health Criteria				
None	67.2	68.6		
1	18.5	17.5		
2 or more	14.3	13.9		
Index Offense	31.1	19.6		
Person	31.1	19.6	31.1	19.6
Drug			22.0	18.2
Property	13.2	16.1		
Other	15.0	11.5		
Education level				
Post-Secondary	5.6	3.8		
Secondary	60.8	59.2		
Less than secondary	33.7	37.0		
Earned education degree in prison	34.1	26.4	85.7	85.8
Visitation	33.3	42.9		
CIP	6.8	4.9		
Work release	7.0	7.4		
EMPLOY	7.9	2.3		
SUD treatment	20.3	12.9	20.3	12.9
PFA	5.3	2.2	94.7	97.8
Age at release				
65 or older	0.8	0.5	0.7	0.5
55-64	5.4	3.2	5.1	3.1
10	13.5	14.1	13.5	14.2
35-44	28.9	24.6	29.0	25.1
30-34	20.6	17.5	20.8	17.6
25-29	19.5	21.5	19.8	21.4
21-24	9.8	15.8	9.7	15.5
Less than 21	1.5	2.8	1.4	2.6
Structured activity days				
More than 1460	4.1	3.1		
1096-1460	2.0	2.1		
731-1095	5.7	5.3		
366-730	14.2	14.2		
270-365	8.9	8.5		
180-269	10.6	9.3		

90-179	14.5	16.9		
Less than 90 days	40.1	40.7		
Length of stay (in months)				
60 or greater	4.4	3.0		
40-59	3.9	3.6		
26-39	9.0	7.9		
20-25	6.0	5.9		
13-19	13.5	13.7		
9-12	13.7	12.1		
5-8	16.4	20.8		
3-4	14.5	17.1		
2	10.5	8.2		
1	5.2	4.6		
Less than 1	2.7	3.1		
Length of stay (in months)				
90 or more			4.4	3.0
61-89			3.9	3.6
37-60			9.0	7.9
61-89			6.0	5.9
37-60			13.5	13.7
25-36			13.7	12.1
13-24			16.4	20.8
9-12			14.5	17.1
5-8			10.5	8.2
3-4			5.2	4.6
Less than 3			2.7	3.1
Unsupervised release			5.5	12.8

Similar trends were observed when comparing training and test samples for both Female Felony and Violent models. Specifically, the training sample subjects demonstrated greater rates of responses considered lower risk, including, ‘Less than 15’ total convictions, ‘0’ violent convictions, ‘1’ felony convictions, ‘0’ VOFPs, ‘0’ DWI Convictions, ‘1’ prison admission, ‘0’ assignments (idle), greater lengths of stay (in months), yet fewer person and drug offenses, visitations, and program participation (e.g., CIP, EMPLY, SUD Treatment, and PFA). Female risk score comparisons are provided in Table 3.

A total of 19 items were selected for Female Violent and 20 for female felony model. Notably, few items are selected for female compared to the male models. This reduced number is likely the result of a smaller female sample (n=7,333), compared to males (n=65,088), where a larger sample size aids in the identification of recidivism prediction patters among the pool of potential items. Further, many of the MnSTARR items measure individuals’ criminal history or correctional involvement, which are less frequently identified among female samples (see Hamilton et al., 2023; Van Voorhis et al., 2010).

Table 3. Assessment Descriptives – Female (N=10,620)

Item	Female Violent Test (n=3,287)	Female Violent Training (n=7,333)	Female Felony Test (n=3,287)	Female Felony Training (n=7,333)
Total Number of Convictions				
Less than 15	57.9	65.3		
15-24	26.7	21.5		
25-29	6.3	5.0		
30-34	4.0	3.0		
35 or more	5.1	5.1		
Total Number of Convictions				
Less than 6			17.2	26.4
6-10			23.7	24.1
11-15			20.5	17.8
16-25			24.7	19.7
Violent Convictions				
0	60.6	67.0		
1	19.8	17.6		
2	8.2	7.1		
3	2.6	3.3		
4	2.7	1.7		
5 or more	4.2	3.3		
Felony Convictions				
1			20.6	49.2
2			22.2	20.2
3			16.8	12.0
4-5			20.4	10.9
6 or more			19.9	7.7
VOFP				
0	91.5	95.5		
1	4.8	3.0		
2 or more	3.7	1.5		
DWI Conviction				
0	66.3	68.8	84.9	85.4
1 or more	33.7	31.2	15.1	14.6
DWI Conviction				
< 2			84.9	85.4
2+			15.1	14.6
Prison Admissions				
1			34.6	36.8
6 or more			39.4	43.8
8			26.0	19.4
Unassigned/Unauthorized Idle				
0	64.0	84.3		
1-3	28.4	12.5		
4 or more	7.6	3.2		
Violent Misconduct			5.5	5.8

New Commit/Violator Admission			68.7	72.2
Probation Violator Admission			31.3	27.8
Mental Health Criteria				
0	96.6	96.0		
1	3.4	4.0		
2 or more				
Major Mental Illness			3.4	4.0
Index Offense				
Person	15.7	11.7	15.7	11.7
Drug			52.7	35.0
Property			18.4	25.5
Other	5.2	5.1	5.2	5.1
Education level				
Post-secondary degree intake	7.9	6.6		
Secondary degree intake	61.9	53.2		
Less than secondary	30.2	40.2		
Earned secondary prison	7.3	11.9	82.2	88.5
Earned post-secondary prison	7.8	7.6		
Education Classes				
11+	9.1	12.1		
6-10	13.5	15.3		
3-5	19.7	21.0	42.3	48.4
1-2	22.1	24.6	22.1	24.6
0			35.6	27.0
Visitations			29.8	19.3
CIP	9.6	7.7	9.6	7.7
Work release	14.9	11.7	7.8	7.9
EMPLOY	8.1	3.9	8.1	3.9
SUD treatment	33.6	10.2	20.3	12.9
PFA	5.2	2.2		
Moving On	0.3	0.3		
Age at release				
55 or older	5.9	3.5		
45-54	13.5	14.2	16.3	16.8
35-44	29.0	25.1	10.6	13.4
30-34	20.8	17.6	37.1	30.4
25-29	19.8	21.4	19.0	16.3
21-24	9.7	15.5	14.6	19.1
Less than 21	1.4	2.6	2.4	4.1
Length of Stay (months)				
37+			9.8	7.8
25-36			8.4	7.4
18-24			32.4	30.9
9-17			16.4	20.8

Overall, the differences between the training and test samples are slight but consistent, with the testing sample demonstrating slightly elevated risk across many items. This distinction demonstrates a common shift observed in correctional populations over time. While the root cause of the shift is beyond the scope of this study, these observed changes are a primary reason for updating and recalibrating a tool's items and weights over time.

3.2 Phase 1: MnSTARR 3.0 Predictive validity

Following the examination of sample descriptives and MnSTARR 3.0's predictive performance was examined. To provide an indication of stability, the full sample, as well as training and test samples were examined for predictive performance across all four models, for a total of twelve analysis sets. We examined each model's predictive performance using accuracy (ACC), calibration (1-RMSE), discrimination (AUC), and a combined metric (SAR). Predictive performance findings for MnSTARR 3.0's predictive validity are provided in Table 4.

Performance metrics range from 0.5 to 1, where generally, values lower than 0.6 are considered weak, those larger than 0.6 moderate, and values greater than 0.7 are considered strong. Regarding the AUC, Rice and Harris (2005) translated AUC values into common effect size indicators, where values below 0.55 are considered negligible, 0.56 to 0.63 are small/weak, 0.64 to 0.70 are moderate, and 0.71 and above are large/strong.

Regarding ACC, all models demonstrated moderate or acceptable levels of accuracy, with values ranging from 0.64 to 0.74. When examining calibration (1-RMSE), similar performance was identified, where all values were moderate/acceptable and ranged from 0.61 to 0.86. Interestingly, the discrimination measure, AUC, is considered most relevant within the field and is identified to have higher levels of performance and values range from moderate (AUC = 0.70), to strong (AUC = 0.80). Most important, the Male Violent (AUC = 0.74) and Felony (AUC = 0.71) and Female Violent (AUC = 0.80) all exceeded 0.71, or strong performance. While still acceptable by national standards (see Desmaris et al., 2022) the Female Felony model was identified to be a moderate effect (AUC = 0.70). Finally, when examining the combined metric (SAR) values ranged from moderate (SAR = 0.67) to large (SAR = 0.74). Generally, findings indicate the four MnSTARR 3.0 models are valid predictors of recidivism, which meet, and often exceed, industry predictive performance standards.

Table 4. Predictive performance metrics by risk model

Metric	ACC	1-RMSE	AUC	SAR
Male				
Violent All	.65	.70	.71	.69
Violent Training	.66	.71	.70	.69
Violent Test	.65	.67	.74	.69
Felony All	.65	.84	.72	.74
Felony Training	.66	.86	.72	.75
Felony Test	.65	.81	.71	.72
Female				
Violent All	.74	.81	.77	.71
Violent Training	.65	.61	.77	.68
Violent Test	.76	.62	.80	.73
Felony All	.65	.73	.73	.70
Felony Training	.64	.72	.74	.70
Felony Test	.67	.73	.70	.70

3.3 Phase 1: MnSTARR 3.0 Predictive disparity

Next, MnSTARR 3.0 risk models were examined for predictive disparity. Specifically, analyses sought to examine if the tool's four risk models predicted equally for males and females and across race/ethnicity sub-groups. An evaluation of tool bias is multi-faceted. This process begins with a visual inspection of group differences, using scatter plots to track sub-group trends. Next, we examine predictive performance metrics, comparing model performance by sub-group. Finally, logistic regression models were used to assess the significance and magnitude of predictive disparity.

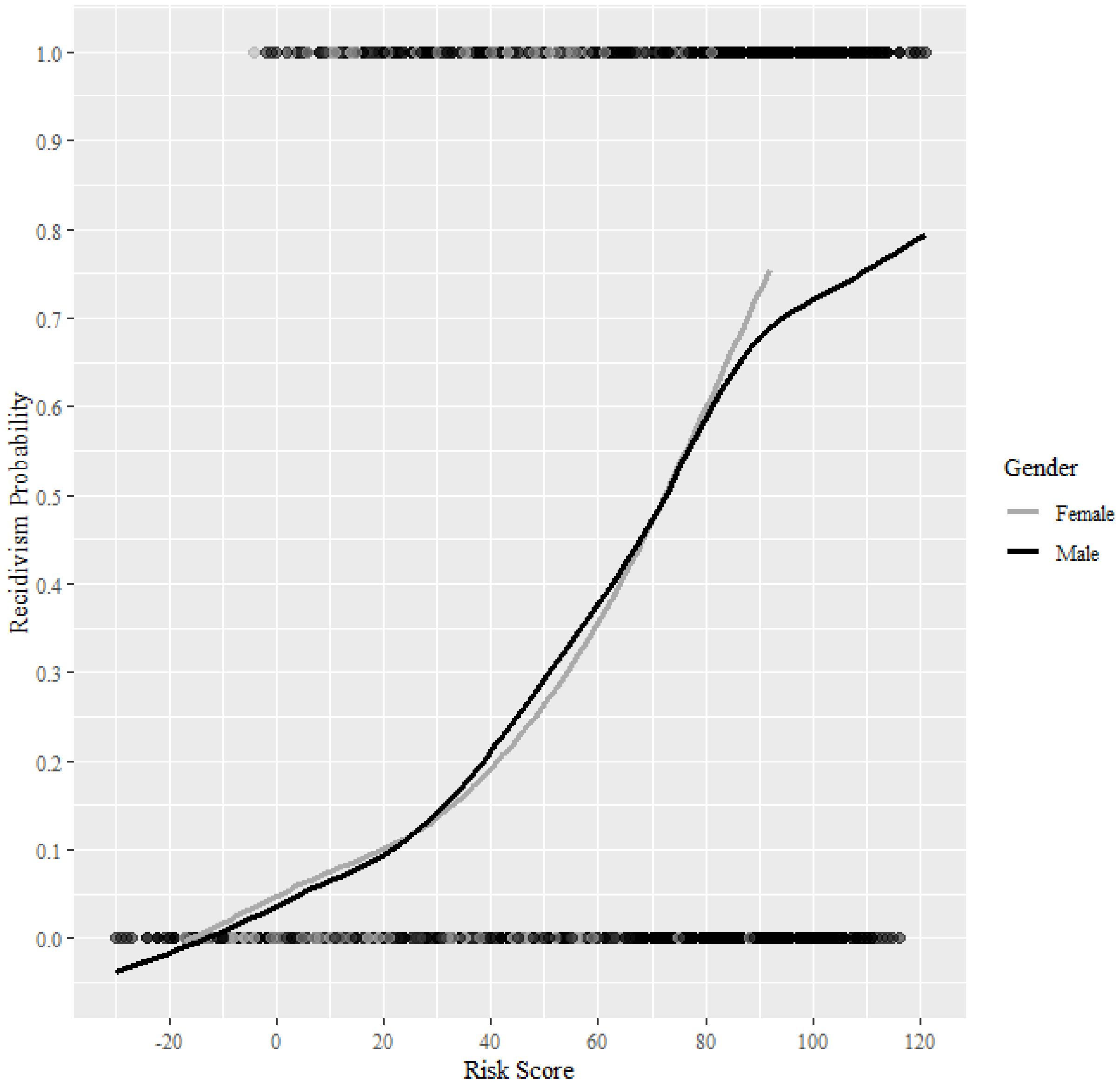
Visual Inspection of Bias

Scatter plots were first created to provide a visual examination of potential sources of gender and race/ethnicity bias within the MnSTARR 3.0 scoring. For all scatter plots, the risk score is displayed on the horizontal axis and recidivism probability is displayed on the vertical axis. The trend lines indicate the coordinates the average recidivism probability associated with each risk score. As described previously, gaps between the trendlines indicate prediction disparity, and, as a point of reference, each one of the gray squares represents 10 points on the MnSTARR and 5 percent recidivism probability.

The MnSTARR felony risk score scatter plot, comparing males and females, is presented in Figure 1. Male scores range from -30 through 120, while female scores range from -20 to just over 90. Notably, the trend becomes steeper between the values 30 to 90, indicating a stronger prediction between that range.

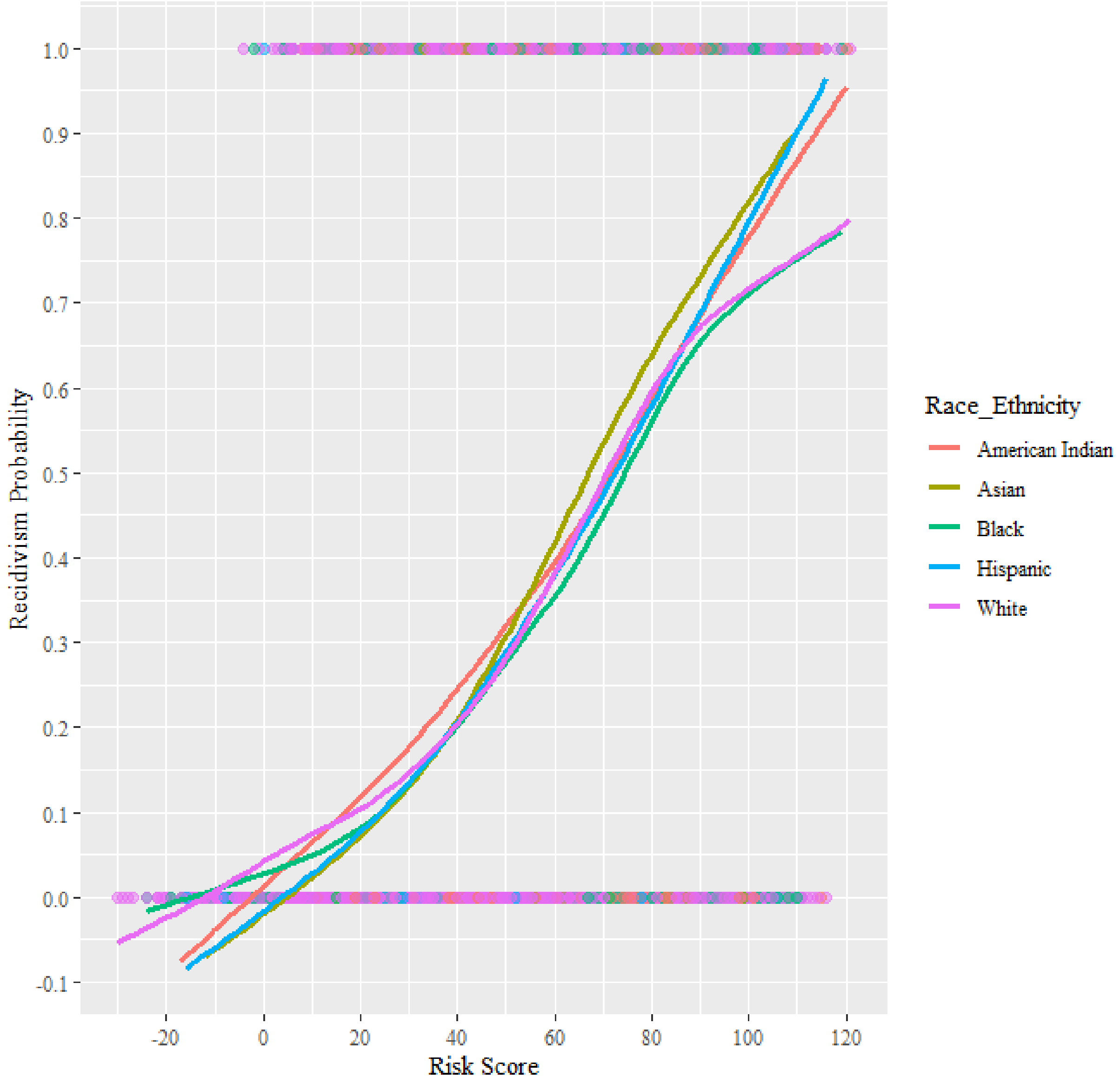
The key takeaway from Figure 1 is the notable absence of bias, or disparity of prediction when observing the trend lines. The only evidence of overclassification for females appear between values 40 and 60, where a slight (roughly 1%) gap in prediction appears. Further the trendlines separate, which is likely attributable to the limited number of subjects scoring over 80 on the MnSTARR Felony model.

Figure 1. Felony risk score and recidivism by gender



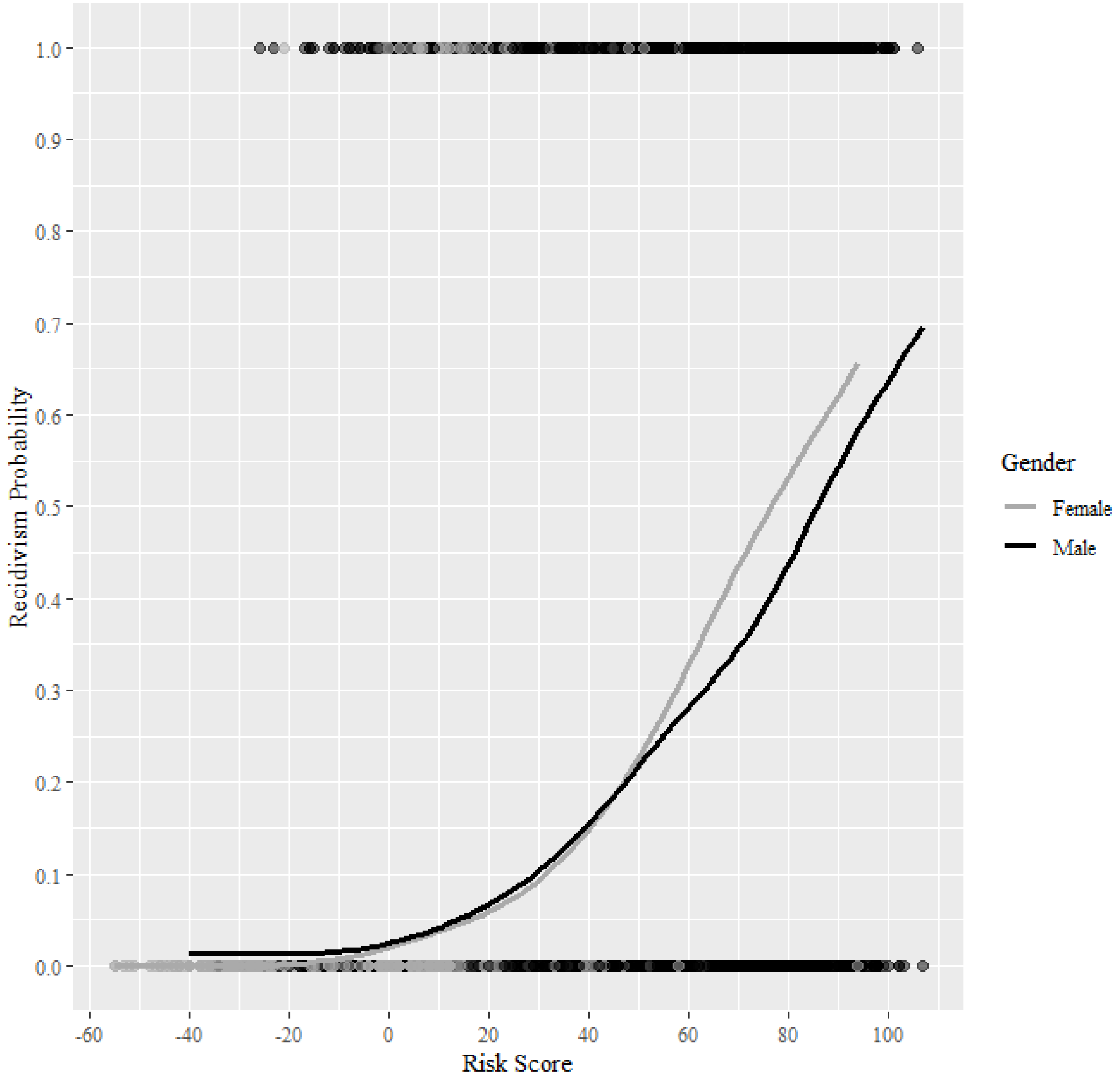
Next, race/ethnicity prediction disparities were examined. Again, scores range from -30 through 120, where the White trendline spans the entire range. Similar to Figure 1, White and Black trends appear flatter, beginning at a score of 90, where the cut point for Very High-Risk begins at a score of 100. Again, this deviation is likely the result of having few cases with risk scores above 90 creating a less stable prediction in the tail of the distribution. Aside from this High-Risk deviation, the race/ethnicity trends are similar, with gaps between trendlines of roughly 5% or less, and therefore, do not display substantial disparity in prediction across the risk score.

Figure 2. Felony risk score and recidivism by race/ethnicity



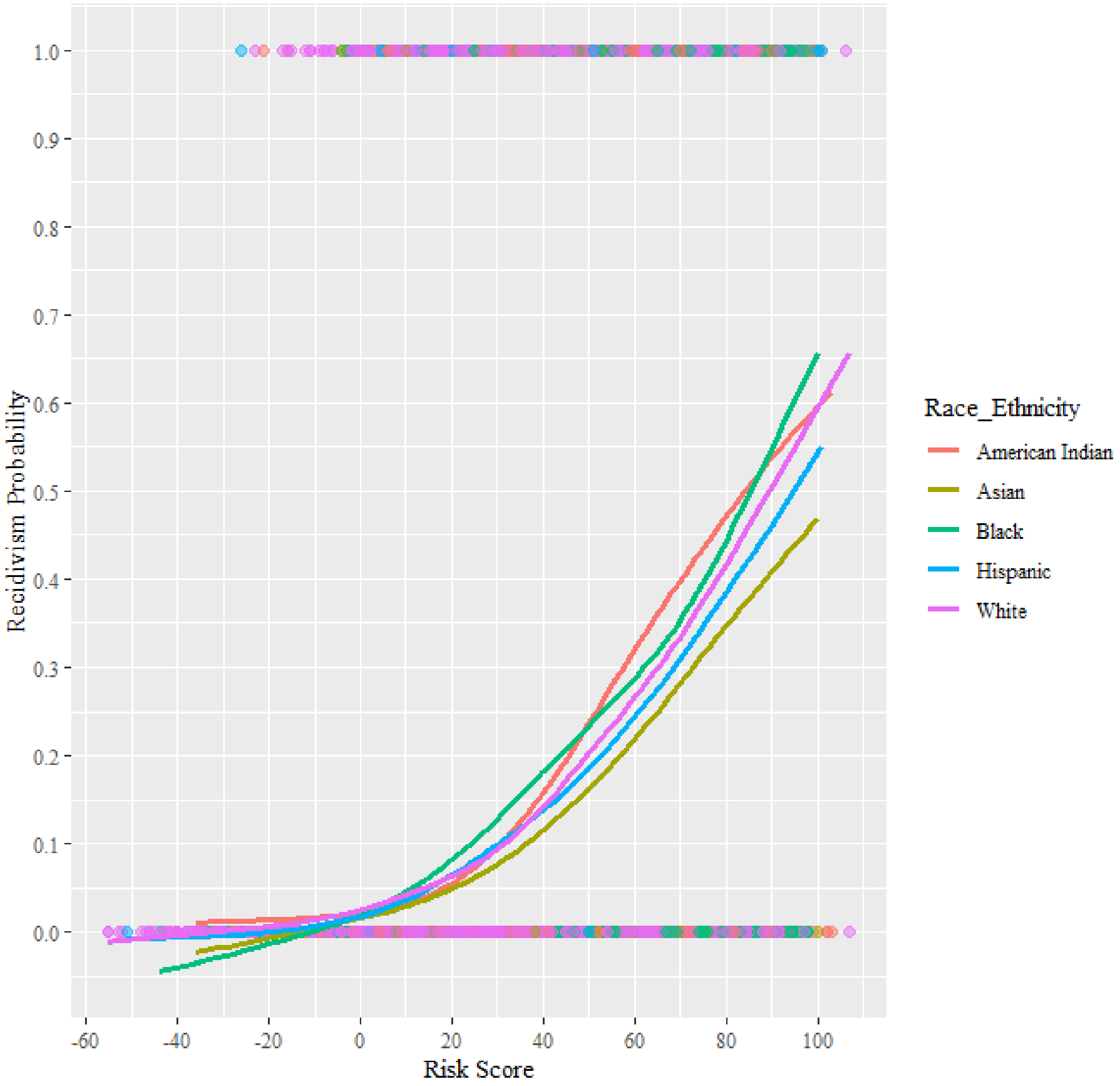
Next, gender comparisons on the MnSTARR violence risk score were visually inspected. The scatter plot is provided in Figure 3. In this visual a more pronounced gap is observed between Males and Females. Beginning at a score of 60, representing the Very High Risk cut point for females, a gap is observed, indicating a small, yet notable, amount of disparity between groups. Specifically, after the score of 60 the MnSTARR violence risk score demonstrates intercept bias and overclassification of males. With this said, it should be noted that the Male and Female scores were not deigned to be equivalent, where cut points for risk scores are used to shore up differences between groups. Yet, prior to the cut point adjustment process it is encouraging to see only minimal differences between groups.

Figure 3. MnSTARR Violent risk score and recidivism by gender



Our final scatter plot compares race/ethnicity scores across the MnSTARR violence risk score. Compared to the felony risk score evaluation in Figure 2, a wider dispersion of trendlines is observed by race/ethnicity. Generally, the MnSTARR violence score demonstrates a steeper/stronger prediction for Black and American Indian, followed by White, Hispanic, and Asian groups. The gap between the highest (Black) and lowest (Asian) is roughly 10% at a score of 50, which is the Moderate Risk cut point for females, and expanding to nearly 15% at a score of 75, which is the Very High-Risk cut point for males.

Figure 4. MnSTARR Violent Recidivism Probability by Race/Ethnicity



Overall, the visual inspection of the risk score disparity indicates minimal-to-small levels of gender and race/ethnicity disparity for the MnSTARR Felony Risk Score. However, greater rates of disparity are observed for the MnSTARR Violent Risk Score. For gender, intercept bias is observed, however, it is counter intuitive as males appear to be overclassified by comparison to females. It should be noted that the Female Violent Risk Score cut point is 60, indicating a recidivism rate of roughly 33%, while the Male Violent Risk Score cut point is 75, indicating a recidivism rate of roughly 38%. Therefore, based on cut point adjustments, Very High-Risk females appear to be overclassified by comparison to males by roughly 5% on the Violent Risk Score.

Regarding race/ethnicity, the Violent Risk Score demonstrates a wider disparity gap, beginning at a score of 20 and expanding though the end of the scoring range. However, with five race/ethnicity categories, it is more likely to observe trendline variations. Further, the most common concern for race/ethnicity bias is the comparison of White subjects compared to persons of color (or all other groups). However, the White trend line appears in the middle of the other race/ethnicity groups, indicating that White subjects are overclassified by comparison to Black and American Indian, and Hispanic and Asian subjects are overclassified by comparison to White subjects. Therefore, this visual inspection is limited, and disparity requires further testing.

Race/Ethnicity Risk Score Disparity

Based on the visual examination, additional examinations were completed for both Felony and Violent Risk scores. Because MnSTARR provides gender-specific scoring, performance metric comparisons were completed separately by gender. Further, the evaluation of disparity is focused on the finished product that is deployed, rather than the development of the tool, and therefore, these analyses were completed on the test sets.

Table 5 provides the study predictive performance metrics by race/ethnicity for males. When examining the test set Accuracy (ACC) Felony Risk Scores are all moderate, ranging from 0.63 to 0.69, while Violent Model scores were small-to-moderate, ranging from 0.58 to 0.67. Regarding calibration (1-RMSE), Violent Risk Score values were all moderate, ranging from 0.64 to 0.70, while Felony Risk Score models were strong, ranging from 0.81 to 0.84. Regarding predictive discrimination (AUC), Violent Risk Scores were all strong, and ranged from 0.71 to 0.74, while the felony risk scores were moderate-to-small, ranging from 0.68 to 0.75. The SAR statistics identified moderate strength for Violent Risk score models, ranging from 0.64 to .69, while the Felony Risk Score models rated as strong, with values ranging from 0.72 to 0.76.

Generally, the performance metric comparison demonstrated moderate-to-strong effects, with only two ACC exceptions (Violent Risk Score White = 0.60 & Asian = 0.58). Again, based on prior examinations of risk assessment scoring, when comparing across races/ethnicities (i.e. highest to lowest statistical value), a less than 5-point difference is considered good, 6-to-10 acceptable, greater than 10-points is worrisome, and a 20- point difference is troublesome. The findings indicated that all races/ethnicities demonstrated good-to-acceptable predictive performance for males.

Given the visual results displayed in Figure 4, displaying greater disparity among higher risk scores by race/ethnicity, two additional performance metrics were provided. The last two statistics are concerned with overclassification, where FPRs identify the rate that higher risk subjects *do not* recidivate (false positives), while the PPV identifies the rate of higher risk individuals that *do* recidivate (true positives). For the Felony Risk Scores, the FPR ranged from 0.30 to 0.39 and the Violent Risk Score ranged from 0.39 to 0.33. Regarding the PPV, Felony Risk Scores ranged from 0.50 to 0.61 and the Violent Risk Score ranged from 0.23 to 0.34. Again, these ranges are considered acceptable-to-good, indicating that higher risk scores are comparable in both error (FPRs) and accuracy rates (PPVs). Further, the PPV test for the Felony Risk scores were all above 0.5, were correctly identified to recidivate with greater than 50% accuracy.

Table 5. Male predictive performance by race/ethnicity

Metric	ACC	1-RMSE	AUC	SAR	FPR	PPV
All Male						
Violent Test	.65	.67	.73	.68	.33	.30
Felony Test	.65	.81	.71	.72	.36	.54
White						
Violent Test	.59	.67	.71	.66	.32	.25
Felony Test	.65	.81	.69	.72	.36	.51
Black						
Violent Test	.66	.69	.73	.68	.38	.34
Felony Test	.63	.81	.73	.72	.39	.54
American Indian						
Violent Test	.63	.67	.73	.68	.37	.39
Felony Test	.62	.86	.68	.72	.31	.60
Hispanic						
Violent Test	.67	.66	.74	.69	.38	.23
Felony Test	.63	.84	.74	.73	.29	.50
Asian						
Violent Test	.57	.67	.73	.64	.27	.22
Felony Test	.69	.84	.74	.76	.35	.57

Table 6 provides the study predictive performance metrics by race/ethnicity for females. When examining the test set, Accuracy (ACC) Felony Risk Scores are all moderate, ranging from 0.63 to 0.75, while Violent Risk Model scores were small-to-moderate, ranging from 0.66 to 0.76. Regarding calibration (1-RMSE), Violent Risk Score values were all moderate, ranging from 0.66 to 0.69, while Felony Risk Score models were moderate-to-strong, ranging from 0.63 to 0.78. For predictive discrimination (AUC), Violent Risk Scores were moderate-to-strong, and ranged from 0.67 to 0.80, while the Felony Risk Scores were moderate-to-strong, ranging from 0.68 to 0.72. The SAR statistics identified small-to-moderate strength for Violent Risk Score models, ranging from 0.59 to 0.73, while the Felony Risk Score models rated as moderate-to-strong, with values ranging from 0.65 to 0.71.

Overall, female performance metrics demonstrated a greater range of variation that the Male models, with several metrics exceeding a 10-point range, including felony accuracy, violent and felony calibration, violent and felony discrimination. Further, several metrics indicated values that were weak/small, such as, Hispanic and Asian calibration (1-RMSE = 0.54 & 0.44, respectively) and the combine metric for Asian subjects (SAR = 0.59). These findings indicate that, while most predictive performance metrics indicate acceptable performance, there is greater instability in the Female models, where both Violent and Felony Risk Models demonstrate inconsistent prediction and may reflect that the items are not optimally capturing the recidivism prediction pattern (i.e., underfitting). This may be as a result of fewer females committing felony and violent recidivism (i.e., lower base rates) across all race/ethnicity groups, and Asian and Hispanic females in particular.

In Table 6, FPRs and PPVs were again examined. For the Felony Risk Scores, the FPR ranged from 0.29 to 0.49 and the Violent Risk Score ranged from 0.19 to 0.49. Regarding the PPV, Felony Risk

Scores ranged from 0.34 to 0.46 and the Violent Risk Score ranged from 0.02 to 0.11. Again, Female demonstrate a wider range of variation as compared to Males. While findings indicated relatively low error rates for those classified as higher risk (FPRs), the low accuracy rates for higher risk individuals (PPVs) indicate two issues. First, higher risk Females do not recidivate at the same rate as Males and rates, and this finding is not consistently observed across races/ethnicities. Second, while the MnSTARR 3.0 is performing well for Females generally, the items are better at identifying those lower risk Females that will not recidivate (true negatives) than higher risk Female that recidivate (true positives).

Table 6. Female predictive performance by race/ethnicity

Metric	ACC	1-RMSE	AUC	SAR	FPR	PPV
All Females						
Violent Test	.76	.62	.80	.71	.35	.14
Felony Test	.67	.73	.70	.70	.40	.37
White						
Violent Test	.76	.57	.80	.73	.29	.11
Felony Test	.71	.69	.72	.71	.40	.35
Black						
Violent Test	.68	.60	.78	.69	.33	.19
Felony Test	.66	.70	.68	.65	.36	.34
American Indian						
Violent Test	.70	.66	.79	.71	.42	.20
Felony Test	.63	.78	.69	.70	.49	.42
Hispanic						
Violent Test	.74	.54	.71	.66	.19	.13
Felony Test	.75	.63	.73	.70	.32	.40
Asian						
Violent Test	.66	.44	.67	.59	.49	.02
Felony Test	.65	.63	.70	.66	.46	.46

Bias Test – The Cleary Method

The prior examinations of disparity explored different ways in which bias may be present, for which groups, and the impact on predictive performance. The final assessment is the *Cleary Method*, which is used to test if bias exists at a statistically significant level. This approach is applied to determine if the MnSTARR Felony and Violent Risk models are "fair" by examining whether there are significant differences in the relationship between risk scores and recidivism across gender and race/ethnicity groups. Four logistic regressions were computed for comparing Males and Females across 1) Felony and 2) Violent Risk Scores, and race/ethnicity across 3) Felony and 4) Violent Risk Scores. Aside from the model intercept, three coefficients are included – risk score, group indicator, and an interaction of the group indicator and the risk score.

The first model examined differences among Males and Females on the Felony Risk Score, and findings are provided in Table 7. The model Chi-Square was identified to be significant (χ^2= 428.414, p<.001) and the R^2 identifies that 8% of the variance in felony recidivism is explained by the variables in the model. As anticipated, Felony Risk Score was statistically significant (p<.001), indicating that it is a strong predictor of felony recidivism. However, the gender coefficient (Male) and the interaction effect (Felony Score*Male) were non-significant, indicating a lack of intercept and slope bias. These findings confirm the visual inspection observed via the scatter plot in Figure 1.

Table 7. Logistic Regression Felony Score – Gender Slope & Intercept Bias

Coefficient	**Logit**	**S.E.**	**p value**	**OR**
Felony Score	.043	.003	**<.001**	1.044
Male	.153	.155	.323	1.166
Felony Score * Male	-.002	.003	.550	.998
Intercept	-3.146	.147	<.001	.043
Model Fit	**Estimate**			
Model Chi-Square	4283.414	--	<.001	--
Pseudo R-Squared	.180			

The second model examined differences among race/ethnicity groups on the Felony Risk Score, and findings are provided in Table 8. The model Chi-Square was identified to be significant (χ^2=4312.437, p<.001) and the R^2 identifies that 18.2 percent of the variance in felony recidivism is explained by the variables in the model. Again, the Felony Risk Score was statistically significant (p<.001), indicating that it is a strong predictor of felony recidivism. However, the race/ethnicity coefficient and the interaction effects were non-significant, indicating a lack of intercept and slope bias. These findings confirm the visual inspection observed via the scatter plot in Figure 2.

Table 8. Logistic Regression Felony Score – Race/Ethnicity Slope & Intercept Bias

Coefficient	**Logit**	**S.E.**	**p value**	**OR**
Felony Score	.042	.001	**<.001**	1.043
Race/Ethnicity	7.152	--	.128	--
White (ref)	--	--	--	--
Black	-.028	.110	.795	.972
American Indian	.256	.155	.098	1.292
Hispanic	-.285	.221	.197	.752
Asian	-.449	.329	.172	.638
Race/Ethnicity	7.759	--	.101	--
*Felony Score * White (ref)*	--	--	--	--
*Felony Score * Black*	-.001	.002	.470	.999
*Felony Score * American Indian*	-.003	.002	.195	.997
*Felony Score * Hispanic*	.004	.003	.224	1.004
*Felony Score * Asian*	.009	.005	.069	1.009
Intercept	-3.020	.064	<.001	.049
Model Fit	**Estimate**			
Model Chi-Square	4312.437	--	<.001	--
Pseudo R-Squared	.182	--	--	--

The third model examined differences among males and females on the Violent Risk Score and findings are provided in Table 9. The model Chi-Square was identified to be significant (χ^2= 23938.254, p<.001) and the R^2 identifies that 18.4% of the variance in felony recidivism is explained by the variables in the model. As anticipated, Violent Risk Score was statistically significant ($p<.001$), indicating that it is a strong predictor of subject's violent recidivism. Specifically, male subjects have 68.8% greater odds of recidivating than female subjects (OR=1.463) indicating intercept bias. Further, the gender coefficient (Male) and the interaction effect (Felony Score*Male) were significant ($p<.001$). Also, the odds ratio of 0.987 indicates that the male odds of recidivism decrease by 1.3% for every additional Violent Risk Score point, which is indicative of slope bias. These findings confirm the visual inspection observed via the scatter plot in Figure 3. However, again, it should be noted that male and female scores were created with gender-specific weighting, where equivalence was not anticipated and differences between scores were adjusted via risk level cut points.

Table 9. Logistic Regression Violent Score – Gender Slope & Intercept Bias

Coefficient	Logit	S.E.	p value	OR
Violent Score	.053	.003	**<.001**	1.054
Male	.524	.142	**<.001**	1.688
Violent Score * Male	-.013	.003	**<.001**	.987
Intercept	-3.886	.134	<.001	.021
Model Fit	**Estimate**			
Model Chi-Square	23938.254	--	<.001	--
Pseudo R-Squared	.184	--	--	--

The fourth model examined differences among race/ethnicity groups on the Violent Risk Score and findings are provided in Table 10. The model Chi-Square was identified to be significant ($\chi^2=23910.494$, $p<.001$) and the R^2 identifies that 18.6 percent of the variance in violent recidivism is explained by the variables in the model. Again, the Violent Risk Score was statistically significant ($p<.001$), indicating that it is a strong predictor of subject's violent recidivism. Further, the race/ethnicity coefficient was significant ($p<.001$), and Black individuals possess significantly greater rates of recidivism than White subjects ($p<.001$). Specifically, Black individuals have 46.3% greater odds of recidivating than White subjects (OR=1.463), indicating intercept bias in the Violent Risk Score across race/ethnicity groups. Meaning that, compared to White individuals, the MnSTARR 3.0 predicts greater rates of recidivism for Black individuals across all risk scores. For example, Black individuals with a score of 0, 50, 75, or even 100 would be predicted to recidivate at a significantly greater rate than White individuals with the same score. Further, the interaction for race/ethnicity and Violent Risk Score was non-significant ($p=0.54$), indicating a lack of slope bias.

Table 10. Logistic Regression Violent Score – Race/Ethnicity Slope & Intercept Bias

Coefficient	Logit	S.E.	p value	OR
Violent Score	.041	.001	**<.001**	1.042
Race/Ethnicity	18.473	--	**<.001**	--
White (ref)	--	--	--	--
Black	3.81	.098	**<.001**	1.463
American Indian	.020	.139	.887	1.020
Hispanic	-.004	.211	.983	.996
Asian	-334	.332	.313	.716
Race/Ethnicity	9.281	--	.054	--
*Violent Score * White (ref)*	--	--	--	--
*Violent Score * Black*	-.004	.002	.021	.996
*Violent Score * American Indian*	.003	.002	.293	1.003
*Violent Score * Hispanic*	-.002	.004	.654	.998
*Violent Score * Asian*	.001	.006	.820	1.001
Intercept	-3.506	.061	<.001	.030
Model Fit	**Estimate**			
Model Chi-Square	23910.494	--	.111	--
Pseudo R-Squared	.186	--	--	--

Overall, the examination of bias demonstrated consistent findings. The felony models, for both males and females and across race/ethnicity subgroups demonstrated minimal-to-no disparities and a lack of intercept and slope bias. However, the violent models indicated disparities in prediction for both gender and race/ethnicity comparisons. Specifically intercept bias was identified for both genders and race/ethnicity comparisons and slope bias was also found when comparing Males and Females on the Violent Risk Score. Finally, the male models' predictive performance metrics (see Table 4) was not as strong on the Violent Risk Score, demonstrating less consistent performance across race/ethnicity groups.

3.4 Phase 1: MnSTARR 3.0 risk level categories

A final aspect of the evaluation is an examination of the MnSTARR risk levels. It is important to understand that assessment developers attempt to place risk levels to achieve two primary outcomes. First, developers will attempt to align risk levels with policy and resource considerations, where programming is reserved for higher risk individuals. Therefore, higher risk levels are set to capture sufficient higher risk individuals for programming resources. Second, risk levels provide a grouping of the continuous risk scores, where each successive category should provide an increased rate of recidivism. Again, when examining recidivism rates by risk level, this concept is often referred to as the "stairstep effect".

The MnSTARR outlines four levels set separately for males and females. The Very High-Risk level is set at a Violent Risk Score of 60 for Females and 75 for Males and a Felony Risk Score of 100 for both Females and Males. The High-Risk score is set at a Violent Risk Score of 50 for Females and 60 for Males, and a Felony Risk Score 80 for Females and Males. The Medium-Risk level is asset at a

Violent Risk Score of 33 for Females and 40 for Males and a Felony Risk Score 60 for Females and Males. Finally, the Low-Risk score identifies all those not yet classified.

Table 11 provides percentages MnSTARR risk levels percentages using the test set. A breakdown of risk level percentages is provided by gender and race/ethnicity groupings. The risk levels are roughly divided into thirds, with 36% identified as Low, 33% as Medium, and a combined 31% identified as High (21.4%) and Very High-Risk (9.5%). Given that Males represent a much larger proportion of the population than Females, their risk level percentages are similar to that of the Total, however, Females indicate few Very High (3.4%) and High-Risk (4.6) individuals but possess a similar proportion of Medium (31.3%), and a substantially greater proportion of Low Risk (60.7%) by comparison to Males. Compared to the Total sample, a greater proportion of White individuals are classified as Low-Risk (428%), while fewer Black (28.0%) and American Indian individuals (27.2%) are identified as Low. All other risk levels by race/ethnicity group are within 5% of the Total percentage.

Table 11. MnSTARR 3.0 Risk Levels by Gender & Race/Ethnicity

RLCs	Total%	Males%	Females%	White%	Black%	American Indian%	Hispanic%	Asian%
Low	36.1	33.0	60.7	42.8	28.0	27.2	39.6	39.7
Medium	33.0	33.2	31.3	33.0	32.0	34.9	32.5	37.4
High	21.4	23.4	4.6	18.1	25.8	24.7	20.3	17.7
Very High	9.5	10.3	3.4	6.1	14.2	13.2	7.6	5.1

Next, recidivism rates were compared across risk levels. To examine the magnitude of risk level's discrimination, or the MnSTARR's ability to discern recidivism rate differences between risk level categories (i.e., the stairstep effect), odds ratios (ORs) are provided. Again, odds ratios represent and effect size estimate where values between 1.01 and 1.43 are negligible, 1.44 to 2.46 are small, and 2.47 to 4.24 are medium, and 4.25 or greater are large.

Risk level by felony recidivism rates and gender are provided in Table 12. When examining the Total sample, a stair step is identified, albeit with non-equivalent steps. Specifically, the Low-Risk group recidivated at a 21.6% rate, followed by nearly a doubling of the rate for Medium (40.5%), a roughly 10% increase for High (51.9%), and Very High-Risk (58.0%). The corresponding odds ratios comparing Low-Risk to the three higher risk categories indicate a medium effect for Medium (OR=2.5) and High (OR=3.9), and a large effect for the Very High-Risk group (OR=5.0).

When comparing recidivism rates by gendered risk levels, rates progressively increase, and odds ratios indicate similar effects by group. This is a positive finding, reflective of the stairstep effect. However, for each risk level, compared to Males, the Female recidivism rate is 2 to 5% lower, reflective of a minor rate of overclassification of females across risk levels.

Table 12. Risk level felony recidivism performance by gender

RLCs	Total%	OR	Males%	OR	Females%	OR
Low	21.6	--	22.0	--	19.4	--
Medium	40.5	2.5	40.7	2.4	38.8	2.6
High	51.9	3.9	52.0	3.8	47.7	3.8
Very High	58.0	5.0	58.4	5.0	49.1	4.0

Next, the rates of felony recidivism by risk level and race/ethnicity were examined. As the majority of the MnDOC population is White, their rates of recidivism and odds ratios roughly mirror the Total sample. The recidivism rates of Black individuals risk level groups are within 5% and odds ratios are identical of the Total, indicating a negligible disparity and good category discrimination. When examining American Indian felony recidivism rates, each risk level identifies a greater rate of recidivism, yet lesser odds ratios by comparison to the Total. These distinctions for American Indian individuals suggest the tool is under-classifying this group and that their recidivism likelihood is greater than that of other groups, on average. Regarding Hispanic individuals, felony recidivism rates were all within 5% and odds ratios are substantially stronger for High (OR=5.0) and Very High-Risk (OR=6.8) groups compared to the Total. When examining Asian individuals, a lower recidivism rate is observed for the Low-Risk group (19.4%), yet greater rates for Medium (46.2%), High (58.9%), and Very High-Risk (77.8%), which is also reflected this group's larger odds ratios when compared to the Total. Generally, the MnSTARR risk levels demonstrate relatively equal rates of prediction for White and Black individuals, while risk levels demonstrate slight weaker rates of prediction for American Indian and stronger discrimination for Hispanic and Asian individuals.

Table 13. Risk level felony recidivism performance by race/ethnicity

RLCs	Total %	OR	White %	OR	Black %	OR	American Indian%	OR	Hispanic %	OR	Asian %	OR
Low	21.6	--	21.4	--	20.3	--	27.5	--	18.9	--	19.4	--
Medium	40.5	2.5	41.4	2.6	37.8	2.5	42.7	2.0	39.0	2.7	46.2	3.6
High	51.9	3.9	53.1	4.1	48.6	3.9	55.3	3.2	53.8	5.0	58.9	5.9
Very High	58.0	5.0	56.3	4.7	57.1	5.0	61.4	4.2	61.5	6.8	77.8	14.5

While the MnSTARR risk level uses both the Felony and Violent Risk Models to set risk level cut points, it is important to examine how effective the risk levels discriminate both felony and violent recidivism outcomes. Table 14 provides violent recidivism rates and odd ratios for the test sample and broken down by gender. The stairstep effect for violence recidivism is optimal for the MnSTARR risk levels, indicating a near 10% increase from Low (5.8%) to Medium (15.4%), a 13% increase from Medium to High (27.1%), and a 15% increase from High to Very High (42.0%). The optimal discrimination is reflected in the odds ratios, where a medium effect is observed when comparing Low to Medium (OR=3.0), a doubling and a strong effect for High (OR=6.0), and a near doubling of an effect size of a very strong effect for Very High Risk (OR=11.8).

Examining violent recidivism, Male rates and odds ratios were similar to the Total. However, when examining the Female violent recidivism rate, Male rates and odds ratios were similar to the Total sample. Further, when examining the Female violent recidivism rate, the Low-Risk group possessed roughly 3% lower (3.0%), Medium 7% lower (8.2%), and High 10% lower (18.5%) violent recidivism rates than males. Odds ratios for Female levels were similar, indicating a medium effect for Medium (OR=2.9), and large for High-Risk (OR=7.3) and Very High Risk (OR=25.0). Generally, these findings indicate the MnSTARR risk levels are providing good-to-excellent discrimination, with each progressively higher risk level identifying a greater rate of violent recidivism. The increasing rate of violent recidivism demonstrates a proportionate increasing pattern, where the stairstep effect is present, and effect sizes roughly double (or more) with each risk level increase.

Table 14. Risk level violent recidivism performance by gender

RLCs	Total%	OR	Males%	OR	Females%	OR
Low	5.8	--	6.4	--	3.0	--
Medium	15.4	3.0	16.3	2.8	8.2	2.9
High	27.1	6.0	27.3	5.5	18.5	7.3
Very High	42.0	11.8	41.9	10.6	43.8	25.0

Finally, violent recidivism by risk level and race/ethnicity were examined. White and Black individuals' rates of violent recidivism and odds ratios roughly mirror the Total sample. However, White individuals possessed slightly reduced and Black individuals possessed slightly increased violent recidivism rates compared to the Total sample at each risk level, yet all White and Black violent recidivism proportions are within 5% of the Total and indicated good-to-excellent discrimination. When examining American Indian individuals' violent recidivism rates, recidivism proportions are within 5% with similar effect sizes as the Total sample.

Regarding Hispanic individuals, violent recidivism rates were all within 5% and three of the four effects were similar to the Total test set. However, the Very High-Risk effect is larger (OR=15.8), by comparison to the Total. When examining Asian individuals, a lower recidivism rate is observed for all risk levels, and while odds ratios are similar to the Total, the rate of violent recidivism for High (15.3%) and Very High-Risk (38.9%) is much lower than the Total sample.

Generally, the MnSTARR risk levels demonstrate relatively equal rates of recidivism for White, Black, American Indian, and Hispanic individuals, while risk levels demonstrate slightly lower rates of violent recidivism for Asian individuals, which provides some evidence of overclassification for this group.

Table 15. Risk level violent recidivism performance by race/ethnicity

RLCs	Total %	OR	White %	OR	Black %	OR	American Indian%	OR	Hispanic %	OR	Asian %	OR
Low	5.8	--	5.2	--	7.3	--	6.5	--	4.2	--	4.3	--
Medium	15.4	3.0	14.0	3.0	18.7	3.0	14.4	2.4	15.3	4.0	11.5	2.9
High	27.1	6.0	23.4	6.0	30.7	6.0	31.4	6.6	22.4	6.5	15.3	4.0
Very High	42.0	11.8	37.1	11.8	44.1	11.8	45.6	12.1	41.3	15.8	38.9	14.1

3.5 Phase 2: MnSTARR 3.0 Dynamic Scoring Impact

In Phase 2, changes in risk scores from intake to release were computed and compared. Specifically, MnDOC provided individuals' MnSTARR 3.0 scores at release, which included points that reflected negative (i.e. misconduct, idle time) and positive (i.e. visits, program participation) behavior during incarceration. To assess their progress, these items were recoded to "0", to reflect individuals' scores at intake. Intake and release scores were compared for mean differences, AUC performance, and the direction of change. It should be noted that all comparisons were completed with the test set. This section provides the results from the comparisons described.

Initially, mean differences (SD) from intake to release on the four MnSTARR 3.0 scores were compared. To identify both significance and magnitude of the differences, t-values and their associated probability levels, as well as Cohen's d values are provided[4]. When examining comparisons across the four models, several consistent findings are identified. First, across all four models, MnSTARR 3.0 scores at release were significantly lower than at intake (p<.001), with mean difference ranging from 3.11 to 8.28 points. Cohen's d effect sizes were small-to-moderate for Male Violent (d=0.40) and Female Violent (d=0.56), and large for Male Felony (d=0.95) and Female Felony (d=0.85) differences. These findings indicate that, on average, individuals reduce their risk scores, with the potential to change their risk level and be eligible for early release under MRRA.

Table 16. Mean Differences Comparing Intake & Release Scores

Metric	Male Violent	Male Felony	Female Violent	Female Felony
Intake Mean (SD)	43.31 (20.13)	62.46 (20.44)	18.58 (17.61)	55.71 (14.29)
Release Mean (SD)	40.19 (23.97)	59.30 (21.40)	12.98 (21.84)	47.42 (19.66)
Score Difference Mean (SD)	3.11 (7.80)	3.16 (3.33)	5.60 (10.07)	8.28 (9.68)
t-value	65.38***	155.00***	31.85***	49.05***
Cohen's d	0.40	0.95	0.56	0.85

*p<.05; **p<.01; ***p<.001

Next, risk score discrimination using the AUC performance metric was examined. Specifically, the intake and release risk score's ability to predict recidivism were compared, identifying potential performance improvement after accounting for positive and negative behavior patterns during incarceration. Further, AUCs of the Score Difference was assessed for predictive performance to isolate the impact of incarceration on recidivism. To provide a measure of effect magnitude, odds ratios (OR) of the Score Difference, predicting recidivism, were also computed. Findings are presented in Table 17.

Results, again, provide consistent findings across the risk models. For all four models Release AUCs are larger than Intake AUC, however, the improvement is minimal and only represents a 1% to 2% improvement. However, AUCs were relatively large to begin with, where substantial increases in performance based on dynamic measures were not anticipated.

[4] Cohen's d values provide an evaluation of effect size magnitude, where values from 0.01 to 0.19 are considered negligible, 0.20 to 0.49 small, 0.50 to 0.79 medium, 0.80 to 1.19 large, and 1.20 or greater are very large.

With that said, the impact of incarceration experiences is still substantial. When examining just the Score Difference, AUCs ranged from 0.57 to 0.64, which indicates small-to-moderate effects. Therefore, the improvement/worsening of risk scores as a result of positive and negative behaviors occurring during incarceration represented a notable impact, collectively. Further, when examining the ORs, each point added (or reduced) between intake and release and represented a 6% to 8% increase in the odds of recidivism. To put this in context, for Males, committing a serious misconduct infraction adds 2-points to their Felony Risk Score, and represents a 16% increase in the odds for recidivism post release. In contrast, for Females, earning a post-secondary degree while incarcerated is worth negative 6-points on the Felony Risk Score, and reduces their odds of recidivism by 36%

Table 17. Predictive Performance Differences Comparing Intake & Release Scores

Metric	Male Violent	Male Felony	Female Violent	Female Felony
Intake AUC	0.73	0.70	0.78	0.68
Release AUC	0.74	0.71	0.80	0.70
Score Difference AUC	0.62	0.57	0.64	0.63
Score Difference OR	1.08	1.08	1.07	1.06

*p<.05; **p<.01; ***p<.001

Next, the direction of risk score change was examined to identify the predictive effects of positive and negative behavior patterns during incarceration. Specifically, three groups were established, those that 1) increased scores, 2) remained the same, or 3) reduced scores from intake to release. The proportion of each category and their associated rate of recidivism was examined for each of the four risk scores. Again, ORs were computed to compare those that increased scores and stayed the same to those that decreased their risk scores. Study findings are provided in Table 18.

Again, findings demonstrate consistent and positive effects of the MnSTARR 3.0. Across the four risk scores, more individuals reduced than remain the same or increased their risk score, with Felony scores demonstrating greater reductions than Violent models. Regarding recidivism, the stairstep effect is observed, where each progressive category demonstrates a greater rate of recidivism. The difference in the rate of recidivism by category is not only significant for each of the four risk scores (p<.001), but the odds ratios also increase progressively and range from moderate to small effects.

Overall, the findings demonstrate the notable effects of programming and other positive behavior patterns on recidivism. Regarding MRRA, these findings should instill confidence that individuals' demonstrating positive behavior are identified to have significantly reduced their odds of recidivism. Further, while these analyses do not represent a traditional program evaluation, the findings indicate that collectively, the MnDOC is providing interventions that are effective and demonstrate evidenced-based recidivism reduction properties.

Table 18. Scoring Change Categories Comparing Intake & Release Scores

Metric	Male Violent	Male Felony	Female Violent	Female Felony
Reduced Score Pop.%	52.3	73.6	61.4	73.0
Remained the Same Pop.%	22.9	17.8	18.5	24.7
Increased Risk Score Pop.%	24.8	8.7	20.1	2.3
Reduced Score Recid.%	14.2	37.0	4.9	23.9
Remained the Same Recid.%	19.3	44.5	7.4	36.6
Increased Risk Score Recid.%	25.7	46.0	11.8	55.8
Reduced Score (ref.)	--***	--***	--***	--***
Remained the Same OR	1.5	1.5	1.6	1.8
Increased Risk Score OR	2.1	1.6	2.6	4.0

*p<.05; **p<.01; ***p<.001

3.6 Phase 2: MnSTARR 3.0 Dynamic Risk Levels

Finally, the MRRA is designed to provide early release for those that demonstrate positive behavior and participate in programming. However, some individuals may not have sufficient time in prison to reduce their scores and achieve early release credits. In a final set of analyses, prison durations were converted to a set of ordinal categories and cross-tabulations were computed to identify the proportion that reduced, remained the same, or decreased their risk scores. Significance tests were computed for each risk score comparison and findings are provided in Table 19.

First, it should be noted that all cross-tabulations demonstrated significant differences across groups and prison times for all four risk scores. Further, across all three prison durations and all four risk scores, the largest group in each comparison was those that reduced their risk score at release. However, with the exception of the Male Felony model, the greatest proportion of those that "Remained the Same" were found in the "<6 Month" release category.

As one examines the "6-23" and "24+" groups there are fewer that "remain the same", which is reflective of the time required to complete programming, receive visits, commit infractions and demonstrate other forms of positive and negative behaviors. These findings have implications for MRRA, where those with longer durations have the greatest opportunity to benefit and likely should be prioritized when there are limited programming slots.

Table 19. Scoring Change Categories Comparing Intake & Release Scores by Prison Time

Metric	< 6 mon.	6-23 mon.	24+ mon.
Male Violent Score***			
Reduced Score Pop.%	41.7	58.3	60.7
Remained the Same Pop.%	42.2	12.6	6.1
Increased Risk Score Pop.%	16.1	29.1	33.2
Male Felony Score***			
Reduced Score Pop.%	83.7	68.4	64.2
Remained the Same Pop.%	14.3	23.7	12.2
Increased Risk Score Pop.%	2.0	7.9	23.6
Female Violent Score***			
Reduced Score Pop.%	51.6	74.2	67.8
Remained the Same Pop.%	30.1	4.9	18.5
Increased Risk Score Pop.%	18.3	4.4	20.1
Female Felony Score***			
Reduced Score Pop.%	57.7	90.6	91.9
Remained the Same Pop.%	40.9	5.4	6.4
Increased Risk Score Pop.%	1.4	4.0	1.7

*p<.05; **p<.01; ***p<.001

Next, analyses were computed to assess the impact of risk level changes. As it pertains to MRRA, an individual may be assessed at intake to be Medium or High-Risk and, upon reassessment, reduce their risk prior to release. Those that move a full level down are the focus of the MRRA's provision, earning early release as a result of substantial positive behavior change. To examine the potential effect on recidivism for those that move down risk levels, categories were reorganized to identify those that reduced a level and those that increased or stayed at their current level from intake to release. Assuming that MRRA credits would not be provided to the High and Very High-Risk levels, these categories were combined for the analysis purposes. With the new organization of risk categories, population descriptives and violent and felony recidivism rates were examined. Further, ORs were provided as a magnitude of each level's effect compared to the lowest (reference) level – Remained Low.

Risk level change findings, from intake to release, are provided in Table 20. Two categories indicate reductions in risk level – "Medium to Low" and "Very High/High to Medium/Low". Collectively, these groups represent 10.8%, indicating the proportion of individuals expected to reduce their risk and potentially become eligible for early release via MRRA.

When examining the rates of recidivism, increasing proportions of both violent and felony recidivism are observed from the "Remain Low" through the "Increased to/Remained Very High/High" category. Like the examination of the MnSTARR 3.0 risk levels (see Tables 11 through 15), the observed "stairstep" effect is encouraging. However, these findings indicate that reductions in risk levels translate to substantial recidivism decreases.

Further, the differences between categories are significant (p<.001), indicating that individuals who reduce their risk level are less likely to recidivate when compared to those that remained or increased their scores to higher risk levels. ORs also demonstrate substantial differences, with small-to-large effects (ORs=1.81 to 8.63) when comparing individuals in the "Remained Low" group to those in higher risk categories

Table 20. Risk Level Change Comparing Intake & Release on recidivism

Level Change	Pop.%	Violent Recid.%	Felony Recid.%
Remained Low	29.4	5.1	19.2
Medium to Low	6.6	8.9	32.0
Increased to/Remained Medium	28.9	14.8	39.2
Very High/High to Medium/Low	4.2	19.7	49.7
Increased to/Remained Very High/High	30.9	31.7	53.8
OR Comparison		Violent Recid. OR	Felony Recid. OR
Remained Low (ref.)	--	--***	--***
Medium to Low	--	1.81	1.98
Increased to/Remained Medium	--	3.23	2.71
Very High/High to Medium/Low	--	4.58	4.16
Increased to/Remained Very High/High	--	8.63	4.90

*p<.05; **p<.01; ***p<.001

3.7 Phase 2:. Recommendations to improve the MnSTARR 3.0.

Finally, following the completion of the revalidation evaluation, additional analyses were completed to help support the development of the MnSTARR 3.0. Specifically, during the evaluation there were notable distinction between Male and Female when comparing recidivism rates across the MnSTARR 3.0 risk levels. As indicated in Table 14, Female rates of violent recidivism are slightly larger for the Very High-Risk group and lower rates for High, Medium, and Low-Risk Females. Regarding felony recidivism (see Table 12), the Female rates are slightly lower for Low and Medium-Risk, and substantially lower for High and Very High-Risk Females.

One reason for the distinction in recidivism rates is likely found within the design of the cut points. The MnSTARR 3.0 creates risk levels by combining Violent and Felony Risk Scores. The intent of using both a general felony and a violent model is to ensure that both the likelihood and seriousness of the recidivism event is accounted for in the classification. Specifically, violent offenses are of greater importance to public safety, and the highest risk level is commonly reserved for those with the highest risk of violence. A violence risk assessment model is termed a "narrow band" model, as it only predicts violent reoffending. An alternate method of using the Felony and Violent models is to create a *hierarchical risk level design*, where the highest risk category is used as a "flag" to indicate a High Violent-Risk level. Using this method, male and female cut points can be adjusted, so that Male and Female recidivism rates are roughly equal across risk levels.

To demonstrate the hierarchical method, the Violent models were used to identify cut points for both Male and Female Violent Scores. Cut points were selected separately for each group, roughly equating the rate of violent recidivism for the Very High-Risk group. Odds ratios were also computed and a comparison of the original and alternate risk level categories (RLCs). Findings are provided in Table 21. One can observe that the violent recidivism rate is roughly equal for the Very High-Risk group for both Males (44.6%) and Females (44.8%). Note for the remaining three groups, the violent recidivism rate is not considered and thus, the violent recidivism rates still differ. Further, while the OR effect sizes for the Very High-Risk groups are both large, the female ORs decreases from 25.0 to 18.8. meaning, that while this hierarchical method increases equity between male and female risk levels, the magnitude of the discrimination effect is reduced.

Table 21. Risk level violent recidivism performance by gender

RLCs	Total%	OR	Males%	OR	Females%	OR
Low	5.8	--	6.4	--	3.0	--
Medium	15.4	3.0	16.3	2.8	8.2	2.9
High	27.1	6.0	27.3	5.5	18.5	7.3
Very High	42.0	11.8	41.9	10.6	43.8	25.0
Alt. RLCs	Total%	OR	Males%	OR	Females%	OR
Low	6.7	--	7.3	--	4.0	--
Medium	15.3	2.5	16.5	2.5	7.0	1.8
High	22.7	4.1	23.0	3.8	11.3	3.1
Very High	44.6	11.8	**44.6**	10.2	**44.8**	18.8

Next, this process was repeated for the felony model, equating Male and Female recidivism rates for Low (18.9% vs. 18.2%), Medium (36.6% vs. 37.4%), and High-Risk (54.8% vs. 54.6%) groups, respectively. Findings are provided in Table 22. Further, ORs increase from moderate to large for both Males (3.8 to 5.2) and Females (3.8 to 5.4). Note, felony rates for Males and Females in the Very High-Risk group still differ.

Table 22. Risk level felony recidivism performance by gender

RLCs	Total%	OR	Males%	OR	Females%	OR
Low	21.6	--	22.0	--	19.4	--
Medium	40.5	2.5	40.7	2.4	38.8	2.6
High	51.9	3.9	52.0	3.8	47.7	3.8
Very High	58.0	5.0	58.4	5.0	49.1	4.0
Alt. RLCs	Total%	OR	Males%	OR	Females%	OR
Low	18.7	--	**18.9**	--	**18.2**	--
Medium	36.7	2.5	**36.6**	2.5	**37.4**	2.7
High	54.8	5.3	**54.8**	5.2	**54.6**	5.4
Very High	59.2	6.3	59.6	6.4	49.5	4.4

Finally, the population percentages were computed, comparing the original and alternate risk levels. Findings are presented in Table 23. The substantive differences between the two RLCs are found in the Medium and High-Risk groups, where fewer Medium (33.0% vs. 25.1%) and a greater proportion of High-Risk individuals (21.4% vs., 31.0%) are identified for the Alternate versus the Original RLCs, respectively.

Table 23. MnSTARR 3.0 Risk Levels by Gender

RLCs	Total%	Males%	Females%
Low	36.1	33.0	60.7
Medium	33.0	33.2	31.3
High	21.4	23.4	4.6
Very High	9.5	10.3	3.4
Alt. RLCs	Total%	Males%	Females%
Low	35.6	32.6	60.2
Medium	25.1	24.7	27.9
High	31.0	33.7	8.6
Very High	8.3	9.0	3.2

It should be noted that the Alternate RLCs are one possible cut point formulation, this version provides a slightly different aim to risk level classification. Specifically, the Very High-Risk designation is reserved for those with the highest propensity for violent recidivism. While there is considerable overlap between predictors of both general felony and violent recidivism, the Alternate RLCs better isolate the violent risk. Given the overlap between those that recidivate generally and those that recidivate violently, there is likely a substantial proportion of individuals in the Very High-Risk groups in both the Original and Alternate RLCs. However, the Alternate RLCs attempt to remove those individuals that may be High-Risk for reoffending generally, but not violent recidivism specifically. This type of hierarchical design is used in other states to drive supervision standards, programming eligibility, and diversion programming.

4.0 CONCLUSION

The use of risk and needs assessments has become a common and evidence-based practice within state correctional systems. However, not all assessment tools are created equal. Tools must account for local variations, as well as gender and race/ethnicity disparity. Further, best practice indicates the need to assess any tool's validity following deployment and routinely thereafter (i.e., every 3 to 5 years).

MnSTARR was developed in 2013, and updates have attempted to improve the tool's efficiency and performance. The current version, MnSTARR 3.0, was designed to have automated scoring and improve predictive accuracy through updated statistical modeling. Responding to the RFP in 2024, the current report provided a validity assessment of the MnSTARR 3.0.

Data from the MnDOC provided MnSTARR items/responses, Felony and Violent Risk Scores, and risk level categories. Recidivism measures were also included, measuring felony and violent felony convictions within three years of release from an MnDOC facility. As part of the MnSTARR development, the sample was divided into training and test sets, totaling over 102,562 subjects.

Using this very large sample, validation analysis findings provide a thorough and robust assessment of predictive performance. When comparing training and test samples, while demonstrating relatively similar descriptive findings, the test sample indicated greater rates of prior and more serious convictions (i.e. violent), prison misconduct, yet more program completions. Given that the test sample was designed to represent more current cases (2017 through 2021), it is likely that changes in statutes, policy, and resources have led to distinctions in the types of individuals and the availability of programming. However, this finding highlights the need to conduct validation assessments routinely, as the populations and the propensity of item responses are demonstrated to change over time, potentially shrinking tool accuracy.

When examining predictive performance, study analyses assessed the full sample, and break downs by training, testing, violent, felony, as well as male and female models were provided. Analyses demonstrated consistent model findings. Importantly, all validity metrics exceeded acceptable performance standards, and the key indicator of discrimination (AUC) often demonstrated strong effect size estimates. Overall, the MnSTARR 3.0 rates as a good-to-exceptional prediction tool and, by comparison to other nationally recognized tools (Singh et al., 2018), demonstrates exceptional predictive performance.

An examination of disparity was also conducted, assessing potential sources of predictive bias across gender and race/ethnicity subgroups. Findings revealed minimal-to-no disparity within the MnSTARR Felony Score, yet some indications of disparity in the Violent Risk score for gender and race/ethnicity were observed. Regarding gender, risk scores were computed separately for males and females, and it was anticipated that differences may occur when comparing the two raw scores Further, some of the identified issues may be the result of smaller sample sizes for sub-groups (e.g., Asian individuals). Additional examination and testing are needed to uncover response weights, items, or cut points that may play a role in the disparity findings uncovered in this report.

Next, MnSTARR risk cut points were examined for both predictive discrimination and disparity. The tool provides four risk levels using cut points, or risk point thresholds, for the Felony and Violent Risk Scores. Findings indicated the progressive prediction of each risk level and the magnitude of effects. Notably a stair step effect was demonstrated, where risk levels increase their prediction of both felony and violent recidivism. With that said, the risk levels demonstrate a greater discrimination effect for violent than felony recidivism and some areas of disparity were identified that require further investigation.

Following the revalidation analyses, additional findings were presented to describe the changes in risk scores and levels that will likely be addressed in MRRA early release consideration. Differences were computed between intake and release scores, indicating significant reductions, across all four risk scores. Substantial differences were identified, where risk point changes were found to provide reduced odds of felony and violent recidivism post-release. Further, those that reduced their risk score identified greater recidivism reduction, where those that reduced their score enough to sufficiently change levels indicated substantial reductions in recidivism odds compared to those that remained or increased risk levels. Finally, greater changes in risk scores were observed for those with longer prison stay, indicating that greater time provides more opportunity to reduce (or increase) risk scores.

These findings provide evidence that support the legislative changes outlined via MRRA. Specifically, the MRRA is designed to provide early release to those individuals that reduce their MnSTARR risk level. Moreover, findings indicate that those individuals that refrain from misconduct and idle time, and participate in programming reduce their odds of recidivism, which will likely motivate those currently incarcerated to refrain from negative behaviors and participate in prosocial activities.

Finally, after an examination of risk levels, there were some notable distinctions between the recidivism rates of males and females. An alternate, hierarchical method of setting cut points was provided, where the Violent models were used to set the Very High-Risk cut point and the Felony model used to set High, Medium, and Low-Risk. This method provides improvements in equity between male and female prediction and offers a potential strategy to determine eligibility for specialized programming and supervision.

5.0 RECOMMENDATIONS

The findings of the revalidation analysis identify that the MnSTARR tool exceeds all predictive standards and possesses minimal levels of disparity. The dual risk scores of both Felony and Violent recidivism predictions retain the MnSTARR's multi-band prediction and the gender-specific modeling captures unique aspect of each gender's predictive items and responses. Further the localized weighting schematic ensures that the tool will provide a strong prediction for the MnDOC population for years to come. The current MnSTARR 3.0 stands as one of the *most effective tools* developed and used by a state DOC.

Given the tool's dynamic scoring, decreases in risk levels demonstrate substantial reductions in recidivism. Tying these reductions to early release will provide a motivator for reluctant individuals to participate in programming. Further, prior studies have indicated that the inclusion of a larger proportion of dynamic items has the potential to further reduce assessment bias (Butler et al., 2022). We recommend that the MnDOC consider expanding their assessment of needs to provide greater opportunities to measure individuals' progress while incarcerated. As additions of quality items is a progress requiring substantial effort, we recommend that additional items be developed following the MnSTARR 3.0's implementation, where additional dynamic items can be tested before consideration for inclusion in future MnSTARR versions.

Finally, as the MnDOC begins to implement MRRA's initiatives it is important to have a plan in place to study its impact. The early release and supervision abatement practices outlined in the Act have the potential to substantially reduce the prison and community supervision population. While there have been similar decarceartion initiatives implemented in other states (Martin, 2016; Pettus-Davis & Epperson, 2015; Schrantz et al., 2018), retrospective examinations are common and make it difficult to identify effective elements. It is recommended the MnDOC create a prospective evaluation design to study the Act's effects and navigate hurdles as they arise. When paired with a robust evaluation, these multi-pronged initiatives can properly parse the "wheat from the chaff with the potential to expand and be replicated elsewhere.

6.0 REFERENCES

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & delinquency*, *52*(1), 7-27.

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation*, *6*(1), 1-22.

Butler, L. C., Hamilton, Z., Krushas, A. E., Kigerl, A., & Kowalski, M. (2022). Racial bias and amelioration strategies for juvenile risk assessment. *Handbook on inequalities in sentencing and corrections among marginalized populations*, 70-118.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004, July). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning* (p. 18).

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation®*, *39*(4), 860-864.

Desmarais, S. L., D'Amora, D. A., & Tavárez, L. P. (2022). *Advancing fairness and transparency: National guidelines for post-conviction risks and needs assessment*. US Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

Desmarais, S., & Singh, J. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*.

Duwe, G. (2014). A randomized experiment of a prisoner reentry program: Updated results from an evaluation of the Minnesota Comprehensive Offender Reentry Plan (MCORP). *Criminal Justice Studies*, *27*(2), 172-190.

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.

Duwe, G. (2025). *Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) 3.0. Minnesota Department of Corrections*. St Paul, MN.

Duwe, G., & Rocque, M. (2018). The home-field advantage and the perils of professional judgment: Evaluating the performance of the Static-99R and the MnSOST-3 in predicting sexual recidivism. *Law and Human Behavior*, *42*(3), 269.

Duwe, G., & Rocque, M. (2019). *The predictive performance of the Minnesota screening tool assessing recidivism risk (MnSTARR): An external validation*. St Paul, MN.

Hamilton, M. (2019). The sexist algorithm. *Behavioral sciences & the law*, *37*(2), 145-157.

Hamilton, Z., Kowalski, M., & Kigerl, A. (2024a). Comparing Meters to Yards: A Nationally Representative. *Justice Quarterly*, *41*(6), 845-869.

Hamilton, Z., Kigerl, A., Allen, B., Ursino, J., & Krushas, A. (2024b). Never Going to Let You Down: Preventing Predictive Shrinkage via the STRONG-R Assessment Method. *Justice Quarterly*, 1-21.

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

Mackey, B. J., Appleton, C. J., Lee, J. S., Skidmore, S., & Taxman, F. S. (2022). At the intersection of research and practice: Constructing guidelines for a hybrid model of community supervision. *Aggression and Violent Behavior*, *63*, 101689.

Martin, W. G. (2016). Decarceration and justice disinvestment: evidence from New York state. *Punishment & Society*, *18*(4), 479-504.

Pettus-Davis, C., & Epperson, M. W. (2015). From mass incarceration to smart decarceration.

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.

Schrantz, D., DeBor, S., & Mauer, M. (2018). *Decarceration Strategies*: *How 5 States Achieved Substantial Prison Population Reductions* (Research and Advocacy for Reform, Issue.

Singh, J. P., D. Kroner, S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.). 2018. *Handbook of Recidivism Risk Assessment*. New York: Wiley.

Taxman, F. S. (2018). Risk assessment: Where do we go from here?. *Handbook of recidivism risk/needs assessment tools*, 269-284.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, *37*(3), 261-288.