Predicting Sex Offense Recidivism: The Perils of Professional Judgment and the Home-Field Advantage

Authors
Grant Duwe, Ph.D.
Research Director
1450 Energy Park Drive, Suite 200
St. Paul, MN 55108-5219
Email: grant.duwe@state.mn.us

Michael Rocque, Ph.D.
Bates College
Department of Sociology
265 Pettengill Hall
Lewiston, Maine 04240
Email: mrocque@bates.edu

This information will be made available in alternative format upon request.
Printed on recycled paper with at least 10 percent post-consumer waste
Research Summary

When sex offenders in Minnesota are assigned risk levels prior to their release from prison, correctional staff frequently exercise professional judgment by overriding the presumptive risk level per an offender’s score on the MnSOST-3, a sexual recidivism risk assessment instrument. These overrides enabled us to evaluate whether the use of professional judgment resulted in better predictive performance than reliance on “actuarial” judgment (MnSOST-3). Using multiple metrics, we also compared the performance of a home-grown instrument (the MnSOST-3) with a global assessment (the Static-99R) in predicting sexual recidivism for 650 sex offenders released from Minnesota prisons in 2012. The results showed that use of professional judgment led to a significant degradation in predictive performance. Likewise, the MnSOST-3 outperformed the Static-99R for both sexual recidivism measures (rearrest and reconviction) across most of the performance metrics we used. These results imply that actuarial tools and home-grown tools are preferred relative to those that include professional judgement and those developed on different populations.
Introduction

Meta-analyses have indicated the average recidivism rate for sex offenders tends to be around 13% within four to five years (Hanson & Bussiere, 1998; Hanson & Morton-Bourgon, 2005), which is lower than estimates made by the public (Helmus et al., 2012a). This does not mean, of course, that sex offenders necessarily represent a low threat to public safety as sexual offending is often seen as more dangerous and potentially damaging than other types of criminal acts. Yet, not all sex offenders are created equally, for some are more at risk for sexual recidivism than others.

Because sex offenders do not represent a monolithic class of high-risk offenders, but rather vary tremendously with respect to recidivism risk, assessing their sexual recidivism risk is important for guiding treatment strategies and improving public safety. Given that research has demonstrated clinical judgment does a poor job predicting recidivism (Hanson and Bussiere (1998) reported a correlation of .10 between clinical judgment and sexual recidivism), a number of actuarial risk assessment tools have been created specifically to classify sex offenders. These include the Sex Offender Risk Appraisal Guide (SORAG) (Quinsey et al., 2006), the Rapid Risk Assessment for Sex Offense Recidivism (RRASOR) (Hanson, 1997), the Minnesota Sex Offender Screening Tool (MnSOST) (Duwe & Freske, 2012; Epperson, Kaul, Huot, Goldman, & Alexander, 2003), the Structured Anchored Clinical Judgment (SACJ) (Thornton, 1997), and a combination of the latter two instruments, the Static-99 (Hanson & Thornton, 2000).

While risk assessment tools have long been utilized, the ongoing revisions to the primary tools available suggest they are works in progress (Harris & Hanson, 2010). Among the unresolved issues within the sex offender risk assessment literature, there are two, in
particular, that have received relatively little empirical scrutiny to date. First, even though it is now generally accepted that actuarial instruments outperform clinical judgment in predicting recidivism, the question of whether clinical judgment is a useful supplement to actuarial tools remains open. Some argue that clinical judgment is subjective and not empirical and thus should not be used (Abbott, 2011), while others claim it can provide useful information alongside actuarial tools (Harris, 2006; Sreenivasan et al., 2010). Second, it is unclear whether tools developed and validated specifically for one population are appropriate or as effective for other populations. Although it may be reasonable to assume a “home-grown” tool should outperform a non-native tool when applied to the population for which it was created (Miller & Lin, 2007), existing research has yet to evaluate whether local instruments truly have a “home-field” advantage in sex offender risk assessment.

To address these questions, we analyze sexual recidivism outcomes over a four-year follow-up period for 650 sex offenders who had been scored on both the Static-99R and the MnSOST-3 prior to their release from Minnesota prisons in 2012. Since 1997, the Minnesota Department of Corrections (MnDOC) has assigned risk levels to sex offenders released from prison. Although most of the sex offenders in our sample received a presumptive risk level according to their MnSOST-3 score, MnDOC staff can override the MnSOST-3 and assign a different risk level based on their professional judgment. The presence of these overrides enables us to assess whether the use of professional judgment, in addition to actuarial tools, increases the accuracy of classification decisions. Moreover, because the 650 offenders were each assessed on the Static-99R and the MnSOST-3, we compare the predictive performance of these two instruments to determine whether there is a home-field advantage in sex offender risk assessment. Finally, we carry out a comprehensive assessment of predictive
performance by using six different metrics. In addition to relying on the widely-used area under the curve (AUC) to evaluate predictive performance, we introduce a number of metrics that are new to the sex offender risk assessment literature—accuracy (ACC), the precision-recall curve (PRC), Hand’s H-measure, root mean squared error (RMSE) and SHARP (squared error, Hand’s H-measure, AUC, RMSE, and PRC).

**Professional Versus Actuarial Judgment**

Research has shown that clinical observations are relatively ineffective in discriminating between those who present higher from lower risk of reoffending (Lindsay & Beail, 2004). Studies evaluating the performance of actuarial tools and unguided clinical observation have tended to indicate clinical observation degrades predictive ability. For example, Bengtson and Langstrom (2007) examined whether the Static-99 and Static-2002 outperformed unstructured clinical observation in predicting recidivism. Using the AUC to measure performance, Bengtson and Langstrom found that clinical observation did not predict recidivism whereas the Static-99 and Static-2002 did.

In analyses of whether professional overrides improve predictive performance, research also suggests actuarial tools work best without such changes. McCafferty (2017) found that risk assessment scores before and after clinical override were related to recidivism, but the scores without override were superior (though generally not significantly so). A study using the Level of Service/Case Management Inventory (LS/CMI) similarly found that clinical overrides, which were more likely to be used to increase risk levels, reduced predictive performance (Wormith, Hogg, & Guzzo, 2012).

Although actuarial instruments generally outperform clinical judgment, their overall performance in predicting recidivism has varied widely across validation studies. Therefore,
the question remains as to whether clinical judgment remains a useful tool for practitioners in the face of uncertainty or when information not taken into account by actuarial instruments is available. For example, writing specifically about the Static-99 tools, Sreenivasan and colleagues (2010) argue the risk estimates are not only unstable due to limitations with the tool’s development samples but they also do not apply to persons with difficult to quantify risk factors. Retaining some degree of clinical judgment, they argue, is essential to take these idiosyncratic factors into account.

Some suggest that due to the highly political nature of sex offender management, as well as the highly variable nature of the population, some degree of professional judgment (or a “middle ground” between actuarial and professional assessment (Harris, 2006, p. 36)) is needed. Others, however, suggest that risk assessment approaches using actuarial tools often fail to translate to risk reduction. For instance, Abbott (2011) argues that “Combining actuarial science with clinical judgment obscures the transparency, accountability, and consistency necessary to establish the reliability and validity of the risk prediction” (Abbott, 2011, pp. 226-227). Although Abbott (2011) argues there is currently no research available to support clinical judgment as a supplement to actuarial tools, Phenix and Epperson (2016) reviewed research that shows this method can “add incremental validity to the Static-99R (p. 445). Thus, whether some degree of “judgment” is necessary or even practical as a supplement to actuarial tools has not been determined.

**Risk Level Assignment and End-of-Confinement Review Committees in Minnesota**

Prior to their release from prison, sex offenders in Minnesota are assigned risk levels, which, in turn, determine the extent to which the community will be notified. Prisoners subject to predatory offender registration are assigned a risk level prior to their release from
prison by an End of Confinement Review Committee (ECRC), which is composed of the prison warden or treatment facility head where the offender is confined, a law enforcement officer, a sex offender treatment professional, a caseworker experienced in supervising sex offenders, and a victim services professional. Following the ECRC meetings, sex offenders are assigned a Level 1 (lower risk), Level 2 (moderate risk), or Level 3 (higher risk). Historically, about 55% receive a Level 1 assignment, 30% a Level 2 assignment, and 15% a Level 3 assignment.

Before receiving a risk level assignment from ECRCs, offenders are assessed for sexual recidivism risk by MnDOC staff from the Risk Assessment/Community Notification (RACN) unit. Since the enactment of the Community Notification Act in 1997, the main instruments the MnDOC has used to assess sexual recidivism risk are the Minnesota Sex Offender Screening Tool (MnSOST), the Minnesota Sex Offender Screening Tool-Revised (MnSOST-R) (from 1999 to 2011), and the MnSOST-3 (since 2012). Based on the scores from these instruments, offenders have received presumptive risk levels. During the late 2000s, the MnDOC also began administering the Static-99 (and later the Static-99R) to predatory offenders in prison, although scores from this instrument have not been used to place inmates in presumptive risk level categories.

In assigning risk levels, ECRCs consider scores from actuarial risk assessment tools as well as additional factors that ostensibly increase or decrease the risk of reoffense (e.g., an offender’s stated intention to reoffend following release or a debilitating illness or physical condition). As a result, ECRCs may override the risk level suggested by the risk assessment tool. For example, if an offender has a MnSOST-3 score of 10%, which would place him in
the presumptive level 3 category, an ECRC can assign this offender a risk level of 1 or 2 if it believes a lower risk level is warranted.

As we see later, ECRCs overrode the MnSOST-3’s presumptive risk level in roughly half the cases involving offenders released from prison in 2012. These overrides provide an opportunity to compare the performance of the presumptive risk levels (i.e., actuarial judgment) versus the assigned risk levels (i.e., professional judgment or clinical override) in predicting sexual recidivism. In doing so, we evaluate whether the use of professional judgment results in more accurate classification decisions for offenders.

**Local Versus Global Risk Assessment**

Actuarial tools, which draw upon a combination of empirically-informed measures to create an overall risk score, can provide both absolute and relative risk assessments of offenders. Relative risk assessment simply provides information concerning whether an individual is more or less likely to reoffend than others. Absolute risk assessment, on the other hand, provides an estimate of how likely it is that the individual will reoffend within a specific period of time. Assessing absolute risk is particularly germane to sex offender civil commitment, for one of the common commitment criteria in the 20 states that operate these programs is that the offender’s likelihood of sexual recidivism over the rest of their lifetime is “more likely than not”, which generally translates to a probability of greater than 50% (Duwe, 2014). Estimates of absolute recidivism risk, however, are influenced by the base rate observed within the offender sample used to develop an instrument. For example, when applied to a sample with a 5% base rate, a tool developed on a sample with a 20% base rate may overestimate the absolute risk of recidivism.
In addition to the base rate, other differences between a tool’s development sample and the population on which the instrument is administered could potentially affect predictive validity. For example, differences between the development sample and the assessment population among the items on the instrument could compromise predictive performance. Yet, even without such differences, the items found to predict recidivism for the development sample may no longer be predictive when applied to another population. It is imperative, therefore, to ensure tools are effective in populations outside of those in which they were developed (Lowenkamp & Bechtel, 2007). As Austin (2006, p. 59) argues, “Generally, if a risk assessment instrument has not been tested on multiple populations under varying conditions, it will not work well on populations it has not been tested on.” Some research supports this statement, with locally-grown risk assessment tools performing better than those developed and validated elsewhere (Drake, 2014). Miller and Lin (2007) suggested that local instruments include factors relevant to local contexts (here, New York City) that “generic” ones did not.

Research on the Level of Service Inventory-Revised (LSI-R), a popular general recidivism risk assessment tool, has indicated that for certain subgroups, the LSI-R performs less well than initial studies showed (Hsu, Caputi, & Byrne, 2009; Schlager & Simourd, 2007). Olver, Stockdale, and Wormith (2014) found the LSI-R, which was developed on a sample of Canadian offenders, performs significantly worse for offender populations from the United States. It is, however, unclear how much of these variations are caused by population differences versus differences in implementation.

Differences across populations initially used to norm sex offender risk assessment tools have resulted in changes to those tools. The Static-99R and Static-2002, for example,
were developed to take age into account in terms of recidivism predictions. This indicates that base rates, and thus classification, are sensitive to the sample used for the analyses. Helmus and colleagues (2012a) examined the stability of absolute risk of recidivism predictions across 23 samples using the Static-99R and the Static-2002R and found “substantial variation” that may lead to “meaningfully different conclusions concerning an offender’s likelihood of recidivism” (p. 1164).

**Static-99R**

Using data from samples of sex offenders from Canada and the United Kingdom (UK), the Static-99 is a “global” risk assessment instrument that is the most widely-used around the world, often on sex offenders from other countries besides Canada or the UK (Helmus et al., 2012a). The instrument contains 10 items, all of which are static. These items are summed to produce a total score that ranges from 0 to 12. The Static-99 includes a revised version (Static-99R) and a variant, the Static-2002. The Static-2002 has 14 items, which produce a score from 0 to 14. Both are used to group offenders into four risk groups from low to high and also includes a revised version, the Static-2002R. The original versions are identical to their ‘‘R’’ versions, with the exception of the cut-points and weights accorded to age (Harris & Hanson, 2010, p. 299).

While the Static-99R is relatively new, some research has found it is adequate in predicting recidivism. For example, a meta-analysis (Helmus et al., 2012a) examining 23 samples (over 8,000 offenders) from across the world evaluated recidivism predictions for both the Static-99 and Static-99R (average follow-up period was 8.2 years). The authors found the Static-99R had an AUC of 0.705 in the combined sample and a weighted average
AUC of 0.693, which suggests its performance in discriminating recidivists from non-recidivists is fair.

The vast majority of studies on the Static-99 have been based on assessments that were scored for research purposes. To date, there have been only three published studies that evaluated assessments administered by field staff. In their examination of 1,928 sex offenders screened for civil commitment in Texas, Boccaccini, Murrie, Caperton, and Hawes (2009) reported the Static-99 had relatively weak predictive discrimination. In contrast, Static-99R field assessments had much stronger predictive performance in the two most recent studies on sex offenders from Canada (Hanson, Helmus, & Harris, 2015) and California (Hanson, Lunetta, Phenix, Neeley, & Epperson, 2014).

**MnSOST-3**

In 2012, Duwe and Freske (2012) significantly revised the MnSOST-R with their development of the MnSOST-3. The sample used to develop the MnSOST-3 consisted of 2,535 sex offenders released from Minnesota prisons. The 2,535 offenders were drawn from two separate samples: the MnSOST-R cross-validation sample and a contemporary sample of released sex offenders. The MnSOST-R cross-validation sample contained 220 offenders released from Minnesota prisons during the early 1990s, whereas the contemporary sample included 2,315 sex offenders released from Minnesota prisons between 2003 and 2006. Relying on sex offense reconviction within four years as the outcome measure, Duwe and Freske (2012) used multiple logistic regression to create the instrument. Moreover, they used bootstrap resampling to not only select the items included in the instrument, but also to internally validate the model.
The MnSOST-3 contains 11 predictors—nine main effects and two interaction effects. Of the nine main effects, only three were items derived from the MnSOST-R (public place, completion of chemical dependency and sex offender treatment, and age at release). The other items are: male victims, predatory offenses, felony offenses, violations of orders for protection/stalking/harassment, disorderly conduct (last three years), and unsupervised release. Thus, while some items are similar between the Static and MnSOST tools (age, male victims), others differ (violations of orders for protection for the MnSOST and relationship to victim for the Static). Using the area under the curve (AUC) as their predictive performance metric, Duwe and Freske (2012) reported the MnSOST-3 had an AUC of 0.796 indicating adequate to good performance. Shortly after implementing the MnSOST-3 in January 2012, the MnDOC began using the MnSOST-3.1 the following month. The MnSOST-3.1 differs from the MnSOST-3 in that it excludes the two interaction effects.

**Data & Method**

**Participants**

Our overall sample consists of 650 sex offenders released from Minnesota prisons in 2012 who had been scored on both the MnSOST-3 and STATIC-99R. MnDOC staff from the RACN unit scored both assessments on all 650 offenders, and RACN staff had received training in how to score both tools. In the instances where offenders had multiple assessments from either instrument during their confinement, we used the most recent one prior to their release. The assessment data for the 650 offenders contain the overall score along with the values for the individual items on both of the instruments.

In comparing professional judgment with actuarial assessments in predicting recidivism, we used a sub-sample of 441 cases from the overall sample of 650 offenders.
Although the vast majority of these offenders were assigned a risk level prior to their release from prison, the MnSOST-3 was the risk assessment tool used to help guide the risk level decision for 441 of the 650 offenders. Among the 209 cases where the MnSOST-3 was not used, the ECRC used the MnSOST-R, the predecessor to the MnSOST-3. As shown later, the risk level data contain the presumptive risk level, per the MnSOST-3, and the actual risk level assigned by the ECRC.

The predicted outcome in this study is sex offense recidivism, which we measured as a rearrest and reconviction. Consistent with the development of the MnSOST-3, we measured recidivism over a four-year follow-up period from the date of the offender’s release from prison in 2012. Recidivism data were collected on offenders through December 31, 2016. Data on arrests and convictions were obtained electronically from the Minnesota Bureau of Criminal Apprehension. As with official crime data in general, these data will not capture sex offenses that were either not reported to the police or, if they were reported, did not, at the very least, result in an arrest. Moreover, these data measure only arrests and convictions that took place in Minnesota. The MnDOC reviewed and approved the research conducted for this study.

**Materials and Procedure: Predictive Performance Metrics**

Consistent with recent research that has used multiple metrics to evaluate the predictive performance of risk assessment instruments (Duwe & Kim, 2016; Duwe & Rocque, 2017; Hamilton, Neuilly, Lee, & Barnoski, 2015; Tollenaar & van der Heijden, 2013), we relied on six different statistics to assess predictive validity. There are three main dimensions of predictive validity: 1) accuracy, 2) discrimination, and 3) calibration. Predictive accuracy assesses how well a model makes correct classification decisions. One of
the more commonly-used metrics is accuracy (ACC), a threshold-based measure that looks at the extent to which an assessment correctly classifies offenders as recidivists or non-recidivists. For example, if a recidivist had a predicted recidivism probability less than 50%, then this offender would be incorrectly classified as a non-recidivist (i.e., false negative). Conversely, if this offender had not recidivated, then s/he would be accurately classified (i.e., true negative). The ACC value ranges from 0 to 100%, and higher ACC values reflect greater accuracy in making correct classification decisions.

The main limitation with using a threshold-based predictive accuracy metric is that it can be less informative and meaningful for highly imbalanced datasets. As we show later, our dataset is highly imbalanced due to the fact that our predicted outcome—sexual recidivism—was very rare. Indeed, only four % were rearrested for a new sex offense within four years, and only two % were reconvicted. Moreover, all but a few of the offenders in our dataset had predicted probabilities—from either the Static-99R or the MnSOST-3—that were less than 50 %. To help illustrate the limitation of applying threshold-based accuracy metrics to imbalanced datasets, we included the ACC in our evaluation of predictive performance.

The second dimension of predictive validity, discrimination, measures the degree to which an assessment separates—in this instance—the recidivists from the non-recidivists. We used three metrics to assess predictive discrimination—the AUC, the H measure developed by Hand (2009), and the precision-recall curve (PRC). The AUC is relatively robust across different recidivism base rates and selection ratios (Smith, 1996), and it has arguably been the most widely-used metric to assess recidivism prediction performance. With values that range from 0 to 1, the AUC statistic is interpreted as the probability that a randomly selected recidivist has a higher score on a risk assessment instrument than a
randomly selected non-recidivist. Values at either end of the spectrum (0 or 1) reflect perfect prediction, whereas a value of 0.50 indicates the prediction tool does no better than chance. According to the literature, an AUC between 0.90 and 1.00 is considered excellent, between 0.80 and 0.89 is good, between 0.70 and 0.79 is fair, between 0.60 and 0.69 is poor, and between 0.50 and 0.59 represents a failure to achieve predictive discrimination (Baird et al., 2013; Thornton & Laws, 2009).

Davis and Goadrich (2006) point out, however, that the AUC can provide an overly optimistic estimate of predictive discrimination for imbalanced datasets. Moreover, as Hand (2009) demonstrates, the AUC uses different misclassification cost distributions for dissimilar classifiers and can provide misleading results if receiver operating characteristic (ROC) curves cross. Given that the AUC evaluates different classifiers using different metrics, Hand (2009) developed a predictive discrimination metric, the H measure, that uses a common cost distribution for all classifiers, with higher values indicating better performance (Hand, 2009). The H measure has seldom been used in existing research on recidivism prediction, although previous studies have reported H values that ranged from 0.02 to 0.40 (Duwe, 2017; Duwe & Rocque, 2017; Hamilton et al., 2015). Another alternative to the AUC is the precision-recall curve (PRC), which uses the precision and recall values to assess predictive discrimination. Compared to the AUC, the PRC has been found to be a better metric for highly imbalanced datasets (Davis & Goadrich, 2006). Like the H measure, PRC values range from 0 to 1, with higher values denoting better performance. In the only recidivism study that has, to our knowledge, used the PRC, the values ranged from 0.05 to 0.24 (Duwe, 2017).
Calibration measures how well the predicted probabilities from a model correspond with the observed outcome being predicted. Whereas predictive discrimination assesses relative risk, calibration taps into absolute risk. In order for a prediction instrument to make accurate absolute assessments of risk, the model’s predicted probabilities must be calibrated with the observed recidivism outcomes. For example, let us assume we have 100 offenders who have recidivism probabilities between 45 and 55 %, which amounts to an average of 50 %. If our model is well-calibrated, we would expect to see 50 recidivists out of the 100 offenders (i.e., the recidivism rate of 50 % equals the average of the predicted probabilities). If, however, the observed recidivism rate was, say, 30 % for the 100 offenders, then it would be a poorly calibrated model that overestimates absolute risk. Conversely, if the observed recidivism rate was 70 %, then the model would underestimate absolute risk. We used root mean square error (RMSE) as our calibration metric. With values that range from 0 to 1, RMSE measures the squared root of the average squared difference between observed recidivism and predicted probabilities. The closer the RMSE value is to zero, the better the calibration.

In addition to these metrics, we used the SHARP (Squared error, H-measure, ACC, Receiver Operating Characteristic, and PRC) statistic, which combines accuracy (ACC) and calibration (RMSE) with the three predictive discrimination measures (Duwe, 2017). The formula for SHARP is: \( \frac{(H\text{-measure} + AUC + PRC + ACC + (1 - \text{RMSE}))}{5} \). Designed specifically for assessing overall predictive performance within highly imbalanced datasets, SHARP is similar to the SAR statistic developed by Caruana, Niculescu-Mizil, Crew, and Ksikes (2004) except that it weights predictive discrimination more heavily by including the H and PRC statistics. The values for SHARP range from 0 to 1, with higher values signifying
better predictive performance. In the only study that has used the SHARP metric, the values ranged from 0.50 to 0.63 (Duwe, 2017).

To estimate the ACC, RMSE, AUC, H-measure, and PRC statistics, we used the “Metrics”, “hmeasure”, and “PRROC” packages in R. For the SHARP statistic, we entered the formula described above into an Excel spreadsheet. We then calculated the SHARP values by entering the values for the ACC, RMSE, AUC, H-measure, and PRC statistics in the Excel spreadsheet.

Results

Among the 650 sex offenders in this study, 26 (4.0 %) were rearrested for a new sex offense within four years of their release from prison in 2012. Of the 26 who were rearrested, 13 (2.0 % of the 650) were reconvicted. Static-99R scores ranged from a low of -2 to a high of 10, with the average being 3.41. Using the sexual recidivism probabilities provided by Phenix, Helmus, and Hanson (2015), the average Static-99R predicted probability was 10.66 %. The average predicted probability for the MnSOST-3.1, which is the “score” for this instrument, was 5.95 %. Therefore, the average predicted probabilities for both instruments are higher than the base rate for this sample, regardless of which sexual recidivism measure is being used.

The Static-99R vs. the MnSOST-3

In Table 1, we show descriptive statistics for the ten Static-99R items and the nine MnSOST-3 items. Moreover, to provide an indication of how well each item predicted sexual recidivism, we include AUC values for both measures of sexual reoffending. The results for the Static-99R show the prior sentencing dates item had the highest AUC for rearrest, whereas prior non-sexual violence had the highest AUC for reconviction. The ever lived with
a lover item, on the other hand, had the lowest AUC for both recidivism measures. Likewise, for the MnSOST-3, the predatory offenses and male victims items had the lowest AUC for rearrest and reconviction, respectively. Felony offenses had the highest AUC for rearrest, while VOFP’s had the highest AUC for reconviction.

Table 1. Descriptive Statistics for MnSOST-3 and STATIC-99R

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Description</th>
<th>Mean/%</th>
<th>SD</th>
<th>Rearrest</th>
<th>Reconvict</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STATIC-99R Items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Age at release</td>
<td>0.368</td>
<td>0.677</td>
<td>0.500</td>
<td>0.465</td>
</tr>
<tr>
<td>Ever lived with lover</td>
<td>Ever lived with lover (two years)</td>
<td>42.8%</td>
<td>0.495</td>
<td>0.458</td>
<td>0.360</td>
</tr>
<tr>
<td>Index Non-Sexual Violence</td>
<td>Any convictions for index, non-sexual violence</td>
<td>10.8%</td>
<td>0.310</td>
<td>0.564</td>
<td>0.484</td>
</tr>
<tr>
<td>Prior Non-Sexual Violence</td>
<td>Any convictions for prior non-sexual violence</td>
<td>48.2%</td>
<td>0.500</td>
<td>0.630</td>
<td>0.647</td>
</tr>
<tr>
<td>Prior Sex Offense</td>
<td>Number of prior sex offenses (0-3)</td>
<td>0.340</td>
<td>0.680</td>
<td>0.555</td>
<td>0.507</td>
</tr>
<tr>
<td>Prior Sentencing Dates</td>
<td>Prior sentencing dates, excluding index</td>
<td>50.0%</td>
<td>0.500</td>
<td>0.660</td>
<td>0.637</td>
</tr>
<tr>
<td>Non-Contact Sex Offenses</td>
<td>Convictions for non-contact sex offenses</td>
<td>5.1%</td>
<td>0.220</td>
<td>0.534</td>
<td>0.553</td>
</tr>
<tr>
<td>Unrelated Victims</td>
<td>Any unrelated victims</td>
<td>84.9%</td>
<td>0.358</td>
<td>0.538</td>
<td>0.498</td>
</tr>
<tr>
<td>Stranger Victims</td>
<td>Any stranger victims</td>
<td>17.2%</td>
<td>0.378</td>
<td>0.591</td>
<td>0.608</td>
</tr>
<tr>
<td>Male Victims</td>
<td>Any male victims</td>
<td>10.6%</td>
<td>0.308</td>
<td>0.465</td>
<td>0.485</td>
</tr>
<tr>
<td><strong>MnSOST-3 Items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predatory Offenses</td>
<td>Number of predatory offense sentences</td>
<td>1.365</td>
<td>0.983</td>
<td>0.442</td>
<td>0.509</td>
</tr>
<tr>
<td>Felony Offenses</td>
<td>Number of felony offense sentences</td>
<td>4.288</td>
<td>3.234</td>
<td>0.712</td>
<td>0.694</td>
</tr>
<tr>
<td>Violations of Orders for Protection</td>
<td>Number of violations of orders for protections</td>
<td>0.435</td>
<td>1.082</td>
<td>0.630</td>
<td>0.716</td>
</tr>
<tr>
<td>Recent Disorderly Conduct</td>
<td>Number of disorderly conduct sentences</td>
<td>0.052</td>
<td>0.255</td>
<td>0.576</td>
<td>0.516</td>
</tr>
<tr>
<td>Age at Release</td>
<td>Age at release (years)</td>
<td>35.294</td>
<td>10.974</td>
<td>0.544</td>
<td>0.607</td>
</tr>
<tr>
<td>Unsupervised Release</td>
<td>Discharged or released to no supervision</td>
<td>8.0%</td>
<td>0.272</td>
<td>0.619</td>
<td>0.577</td>
</tr>
<tr>
<td>Complete SO/CD treatment</td>
<td>Completed sex off. &amp; chem. dep. treatment in prison</td>
<td>4.8%</td>
<td>0.213</td>
<td>0.495</td>
<td>0.515</td>
</tr>
<tr>
<td>Male Victims</td>
<td>Number of predatory offenses with male victims</td>
<td>0.094</td>
<td>0.371</td>
<td>0.482</td>
<td>0.503</td>
</tr>
<tr>
<td>Public Place</td>
<td>Has committed sex offense in a public location</td>
<td>14.9%</td>
<td>0.357</td>
<td>0.583</td>
<td>0.581</td>
</tr>
</tbody>
</table>

In Table 2, we present a confusion matrix that helps illustrate the accuracy of the Static-99R and MnSOST-3 in predicting sexual recidivism. Because the results were similar for both measures of sexual recidivism, we focus on those for sex offense rearrest. Of the 650 offenders, only one had a Static-99R probability of 50 % or higher while five had a MnSOST-3 probability of 50 % or higher. Given the low base rate for this sample (2 % for reconviction and 4 % for rearrest), the vast majority were correctly classified as “true negatives” (i.e., non-recidivists with predicted probabilities less than 50 percent). For the Static-99R, the lone offender with a probability of 50 % or higher was a “false positive” insofar as he did not recidivate. Of the five positives predicted by the MnSOST-3, one
recidivated (“true positive”) while the other four did not (“false positive”). The MnSOST-3 thus had one fewer “false negative” (i.e., recidivists who had predicted probabilities less than 50 percent) than the Static-99R.

**Table 2. Confusion Matrix for Static-99R and MnSOST-3: Sex Offense Rearrest**

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Rearrest</th>
<th>False Negatives</th>
<th>True Negatives</th>
<th>False Positives</th>
<th>True Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative (Non-Recidivists)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Static-99R = 26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MnSOST-3 = 25</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive (Recidivists)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Static-99R = 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MnSOST-3 = 1</td>
</tr>
</tbody>
</table>

The results in Table 3 indicate the MnSOST-3 achieved better predictive performance than the Static-99R for both measures of sexual recidivism for all but one of the six metrics. Because the MnSOST-3 had three more false positives (and three fewer true negatives), the Static-99R had a slightly higher ACC (by 0.003) for both measures of sexual recidivism. On the other hand, the MnSOST-3 had marginally better calibration, as reflected by the slightly lower RMSE values.

When we focus on the three predictive discrimination metrics, the results suggest the MnSOST-3 was better at distinguishing recidivists from desistors. The AUC value was higher for the MnSOST-3 (lower confidence interval = 0.662; upper confidence interval = 0.851 for rearrest; lower confidence interval = 0.583; upper confidence interval = 0.849 for reconviction) compared to the Static-99R (lower confidence interval = 0.569; upper confidence interval = 0.794 for rearrest; lower confidence interval = 0.446; upper confidence interval = 0.766 for reconviction) for both sexual recidivism measures, although neither ROC
curve comparison was significantly different at the $p<.05$ level ($p = 0.215; Z = 1.239$ for rearrest; $p = 0.314; Z = 1.007$ for reconviction) (DeLong, DeLong, and Clarke-Pearson, 1988). The values for the other two predictive discrimination metrics, Hand's H-measure and the PRC, were also higher for the MnSOST-3. As a result of the differences observed for the predictive discrimination metrics, the MnSOST-3 had higher SHARP values than the Static-99R for both sexual recidivism measures.

Table 3. Predictive Performance Results

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Discrimination</th>
<th>Calibration</th>
<th>Consolidated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex Offense Rearrest</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MnSOST-3</td>
<td>ACC 0.955</td>
<td>PRC 0.135</td>
<td>H 0.098</td>
<td>AUC 0.756</td>
</tr>
<tr>
<td>Static-99R</td>
<td>ACC 0.958</td>
<td>PRC 0.084</td>
<td>H 0.036</td>
<td>AUC 0.682</td>
</tr>
<tr>
<td>Presumptive Risk Level</td>
<td>ACC 0.961</td>
<td>PRC 0.071</td>
<td>H 0.034</td>
<td>AUC 0.723</td>
</tr>
<tr>
<td>Assigned Risk Level</td>
<td>ACC 0.961</td>
<td>PRC 0.040</td>
<td>H 0.002</td>
<td>AUC 0.541</td>
</tr>
<tr>
<td><strong>Sex Offense Reconviction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MnSOST-3</td>
<td>ACC 0.975</td>
<td>PRC 0.118</td>
<td>H 0.096</td>
<td>AUC 0.716</td>
</tr>
<tr>
<td>Static-99R</td>
<td>ACC 0.978</td>
<td>PRC 0.039</td>
<td>H 0.018</td>
<td>AUC 0.606</td>
</tr>
<tr>
<td>Presumptive Risk Level</td>
<td>ACC 0.977</td>
<td>PRC 0.037</td>
<td>H 0.016</td>
<td>AUC 0.701</td>
</tr>
<tr>
<td>Assigned Risk Level</td>
<td>ACC 0.977</td>
<td>PRC 0.022</td>
<td>H 0.000</td>
<td>AUC 0.515</td>
</tr>
</tbody>
</table>

ACC = Accuracy; higher values denote better performance  
PRC = Precision-Recall Curve; higher values denote better performance  
H = Hand's H measure; higher values denote better performance  
AUC = Area under the Curve; higher values denote better performance  
RMSE = Root Mean Squared Error; lower values denote better performance  
SHARP = Squared error, H-Measure, Accuracy, Receiver Operating Characteristic, and PRC; higher values denote better performance

To place these predictive performance results in a more practical perspective, let us assume that, consistent with the risk principle (Bonta & Andrews, 2007), we used the Static-99R and MnSOST-3 to prioritize higher-risk inmates for sex offender treatment. Further, given resource limitations, we can only deliver treatment to about one-fourth of all sex offenders, which is generally the case within the MnDOC. The data show that 170 offenders
(26 % of the total sample) had a Static-99R score of 5 or higher. Likewise, there were 170 offenders with a MnSOST-3 score of 5.8 % or higher.

When we use this cut point for the Static-99R, we see that it captured 14 sex offense rearrests (54 % of all rearrests) and 5 sex offense convictions (38 % of all reconvictions). In comparison, when we use the 5.8 % cut point for the MnSOST-3, we see that it netted 17 sex offense rearrests (65 % of all rearrests) and 7 sex offense convictions (54 % of all reconvictions). Therefore, under this scenario, the MnSOST-3 captured three more offenders who were rearrested, two of whom were reconvicted, for a new sex offense.

Professional vs. Actuarial Judgment

Table 3 also presents the results for the MnSOST-3 presumptive risk levels and the ECRC assigned risk levels, which represent a comparison between actuarial risk assessment and professional judgment. Because the presumptive and assigned risk levels represented the “scores” for this comparison, we generated predicted probabilities for the sub-sample of 441 offenders. Similar to the approach used by Duwe and Rocque (2016) in which they converted LSI-R scores into predicted probabilities, we created predicted probabilities by regressing the risk level values on the two sexual recidivism measures (17 were rearrested and 10 were reconvicted). None of the probabilities were 50 % or higher (i.e., all offenders were predicted to be “negatives” or non-recidivists), which is why the values for the presumptive and assigned risk levels were the same for each of the three accuracy metrics.

For the five remaining metrics, we see slightly worse performance when the MnSOST-3 assessments are collapsed into one of three risk levels. Still, the results show the MnSOST-3 outperformed the ECRC’s across each of these metrics for the two recidivism measures. As with the results for the MnSOST-3-Static 99R comparison, there was relatively
little difference with respect to calibration. There were marked differences, however, when
we focus on predictive discrimination and our consolidated metric—SHARP. For example,
the ROC curve comparison showed the AUC value for the MnSOST-3 (lower confidence
interval = 0.614; upper confidence interval = 0.832) was significantly higher than the AUC
value for ECRCs (lower confidence interval = 0.419; upper confidence interval = 0.664) for
rearrest (p = 0.014; Z = 2.462). For reconviction, the difference between AUC values for the
MnSOST-3 (lower confidence interval = 0.565; upper confidence interval = 0.837) and
ECRCs (lower confidence interval = 0.352; upper confidence interval = 0.678) approached
conventional statistical significance (p = 0.089; z = 1.702). In addition to having higher H-
measure and PRC values, the MnSOST-3’s SHARP value (0.520) was 0.049 higher for
rearrest and 0.044 higher for reconviction (0.517).

Table 4. Sex Offense Recidivism Results by Risk Level Overrides

<table>
<thead>
<tr>
<th>Overrides</th>
<th>Risk Levels</th>
<th>Rearrest</th>
<th>Reconviction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MnSOST-3/Presumptive</td>
<td>ECRC/Assigned</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>None</td>
<td>Level 1</td>
<td>Level 1</td>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Level 2</td>
<td>Level 2</td>
<td>3</td>
<td>4.91</td>
</tr>
<tr>
<td></td>
<td>Level 3</td>
<td>Level 3</td>
<td>2</td>
<td>6.25</td>
</tr>
<tr>
<td></td>
<td>No Override Total</td>
<td></td>
<td>6</td>
<td>2.79</td>
</tr>
<tr>
<td>Downward</td>
<td>Level 2</td>
<td>Level 1</td>
<td>2</td>
<td>3.77</td>
</tr>
<tr>
<td></td>
<td>Level 3</td>
<td>Level 2</td>
<td>5</td>
<td>9.26</td>
</tr>
<tr>
<td></td>
<td>Level 3</td>
<td>Level 1</td>
<td>3</td>
<td>11.11</td>
</tr>
<tr>
<td></td>
<td>Downward Override Total</td>
<td></td>
<td>10</td>
<td>7.46</td>
</tr>
<tr>
<td>Upward</td>
<td>Level 2</td>
<td>Level 3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Level 1</td>
<td>Level 2</td>
<td>1</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>Level 1</td>
<td>Level 3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Upward Override Total</td>
<td></td>
<td>1</td>
<td>1.09</td>
</tr>
</tbody>
</table>

What these results indicate is that when ECRC’s deviated from the MnSOST-3’s
presumptive risk levels, the assigned risk levels resulted in inferior predictive performance.

In Table 4, we take a closer look at the ways in which professional judgment produced worse
classification decisions. If the ECRC overrides led to better predictive performance, we should expect to see a lower recidivism rate for the downward overrides and a higher rate for the upward overrides. Instead, we see the opposite in Table 4.

In Table 4, we see higher recidivism rates—for both measures—for the downward overrides and lower rates for the upward overrides. Among the 215 cases in which there was risk level agreement between the MnSOST-3 and ECRC, the overall rearrest rate was 2.79 % and the reconviction rate was 1.40 %. Among the 134 cases in which the ECRC’s considered the offender’s risk to be less than that recommended by the MnSOST-3 (i.e., the downward overrides), the rates were 7.46 % for rearrest and 4.48 % for reconviction. For the 92 cases in which the ECRC’s believed the offender’s risk was greater than that recommended by the MnSOST-3 (i.e., the upward overrides), the rearrest and reconviction rate was 1.09 %.

As noted above and in Table 4, 17 (3.9 %) of the 441 offenders were rearrested for a new sex offense while 10 (2.3 %) were reconvicted. Therefore, even though downward overrides accounted for 30% of the 441 cases, nearly 60 % of the offenders rearrested for a new sex offense (10 of the 17) were given a downward override. More specifically, five were presumptive level 3 offenders who were assigned a level 2, three were presumptive level 3 offenders who were given a level 1, and two were presumptive level 2 offenders who were assigned a level 1. Similarly, for reconviction, we see that six of the ten recidivists (60 %) were downward overrides, and each of the three downward override categories had two recidivists.

Just as this evidence demonstrates the downward overrides did not improve predictive performance, the same is true for the upward overrides. Of the 92 upward overrides, there was only one recidivist (i.e., this offender was rearrested and reconvicted for a new sex
offense). More specifically, of the 63 presumptive level 1 offenders who were assigned a level 2, only one recidivated. Meanwhile, none of the 29 presumptive level 1 or 2 offenders who were assigned a level 3 sexually reoffended.

Discussion & Conclusion

In the last several years, much progress has been made with respect to risk assessment of sex offenders. As is the case for offenders in general, assessment for sex offenders has progressed from purely clinical (and to some, subjective) judgment to several iterations of actuarial tools. One of the earliest actuarial tools developed was the MnSOST, which has been updated to its current MnSOST-3 version. The most popular tool in North America among criminal justice agencies is the Static-99, developed in the 1990s and updated to its Static-99R version. Research on each of these tools has indicated adequate to good predictive performance, yet practitioners considering these (or other) tools still face several questions concerning which may be most appropriate for their population.

First, the experience in Minnesota indicates that professional judgment still plays a large role in classifying sex offenders even with the use of the MnSOST-3. That is, the ECRC’s consider both the MnSOST-3 and additional factors that may conceivably increase or decrease an offender’s risk of re offending when placing them into risk categories. Second, whether particular tools perform better than others when implemented in populations for which they were developed remains an open question. Is there a “home-field advantage?” An instrument’s development sample is important in determining the populations for which its use is appropriate. Indeed, the degree to which the base rate and attributes of the assessment population mirror those of the tool’s development sample is critical for prediction purposes.
This study directly compared the MnSOST-3 and the Static-99R within a sample of Minnesota sex offenders who were scored with each tool. Findings demonstrated that the MnSOST-3 performed better than the Static-99R on virtually all of the metrics we used for both measures of sexual recidivism. Moreover, we examined the impact of professional judgment or clinical override on classification decisions by comparing the performance of presumptive and assigned risk levels in predicting sexual recidivism. If the ECRC overrides, which are professional judgment supplements to the actuarial tool, add incremental predictive validity, this would be evidence of the value of professional judgment. However, our results indicated unequivocally that clinical judgment in the form of overrides decreased predictive performance, which offers additional evidence that empirically-based actuarial tools are superior to professional judgment.

It is interesting to note the literature seems clear that professional judgment performs worse than actuarial methods irrespective of whether that judgment is structured or unstructured, or the background of the professional making the observation (Hanson & Bussiere, 1998; Grove et al., 2000; Mills & Kroner, 2006; Walters et al., 2014). This is true even for clinical judgement used in combination with actuarial tools (Abbott, 2011). Some research has noted that raters are unfamiliar with or do not use base rate information appropriately in assigning risk (Mills & Kroner, 2006; Walters et al., 2014). Another possibility is that judgement, whether structure or not, necessarily involves a higher degree of subjectivity than actuarial measures and so they are poorer in terms of prediction. Finally, it may be the case, as Hanson and Morton-Bourgon (2005) found, that clinical judgement often utilizes factors that are not related to recidivism (e.g., “low victim empathy”).

25
Due to several limitations, however, these findings should be regarded as somewhat preliminary. First, because our study was confined to sex offenders from a single jurisdiction, it is unclear the extent to which the findings are generalizable. Second, the sample we used was relatively small (N = 650), and it was limited to releases over one calendar year. Third, similar to prior research (Andrews et al., 2011; Blair, Marcus & Boccaccini, 2008; Singh, Grann & Fazel, 2013), the better findings for the MnSOST-3 may reflect an “allegiance effect” in which its scoring and use by MnDOC staff has been more consistent with its design in comparison to the Static-99R. Although we were unable to examine inter-rater reliability (IRR) among the cases scored in our sample for either tool, there is some evidence the MnSOST-3 can be consistently scored by RACN staff on Minnesota sex offenders (Duwe and Freske, 2012). Previous research indicates the Static-99R has adequate IRR (Phenix et al., 2015), although it is not clear whether the Static-99R was scored consistently by RACN staff. As prior research has demonstrated, the presence of inter-rater disagreement can have a significant impact on predictive performance (Duwe & Rocque, 2017).

Even with these limitations, however, our study holds several important implications for research, policy, and practice. First, existing research on the validation of sex offender risk assessment tools has often relied a single metric—namely, the AUC. As we noted earlier, the AUC has its strengths but it also has some weaknesses. We suggest that future validation research should begin using alternative measures of predictive discrimination such as Hand’s H-measure and the precision-recall curve. But given that predictive discrimination addresses only one dimension of predictive validity, metrics that assess accuracy and calibration should also be used to provide a more comprehensive evaluation of predictive performance. As this study illustrates, accuracy metrics are informative for imbalanced
datasets so long as there are at least some predicted positives in the dataset. Moreover, if researchers and practitioners must rely on a single metric, we suggest that either the SAR or SHARP statistics would be preferable since both tap into multiple dimensions of predictive validity.

Second, the AUC values for both the Static-99R and MnSOST-3.1 were lower in comparison to what most of the existing research has reported for either instrument. Much of this research, as we indicated earlier, consists of assessments that were scored for research purposes. In this study, we used assessments that had been scored by correctional staff for operational purposes, which provide what is arguably a truer test of predictive performance. Compared to field assessments, those administered strictly for the sake of research may yield overly optimistic estimates of predictive performance due to more favorable conditions in which raters are likely to have had more recent, thorough training. To provide a more realistic estimate of how sex offender risk assessment tools perform in practice, future research should begin relying more on assessments performed by field staff. In addition, the results suggest that caution may be warranted in using an instrument whose predictive performance has yet to be evaluated on real-world assessments.

Third, given that the MnSOST-3 outperformed the Static-99R for our sample of Minnesota sex offenders, the results suggest local instruments may have a home-field advantage. To be sure, there are differences between the two instruments in terms of the items included and the classification methods used to develop the tools. In fact, to better demonstrate whether local instruments have a home-field advantage over global assessments, future research should attempt to more effectively isolate the effects of using a customized assessment compared to an imported instrument. Still, the evidence presented here suggests
there may be value in applying an instrument to the same, or at least similar, population on which it was developed and validated.

In our view, home-grown instruments developed and validated within a particular population are the best option when considering tools for that population. Of course, many jurisdictions will not have a validated actuarial tool that was customized specifically for their own offender populations. In that case, universal tools (i.e., those developed using several populations, such as the Static-99 family) may be a good option, although such tools should be developed and validated on samples that are truly universal. Put another way, the population on which an instrument is being used should be very similar to the one on which the assessment was developed and validated.

When a global instrument is used, it cannot be assumed the tool will deliver the same performance for a different assessment population. As Austin (2006) states, “There is a tendency for correctional agencies to simply borrow or buy an instrument that has been developed on another population that may or may not reflect the attributes of their own offender populations” (p. 59). In order to understand whether a particular tool is effective with an agency’s population, the tool’s predictive performance must be evaluated on that population.

Finally, our findings provide one more “nail in the coffin” for the value of clinical judgment in making recidivism predictions. While some evidence exists that certain factors (dynamic ones in particular) may improve tools like the Static-99 (Phenix & Epperson, 2016), the vast majority of empirical research has demonstrated that actuarial tools significantly outperform professional judgment. This does not mean clinical judgment is not
important for the purposes of guiding treatment. Rather, when sex offenders are classified for recidivism risk assessment purposes, actuarial tools should be the preferred method.

Given the consistently demonstrated superiority of actuarial assessments in predicting recidivism, we suggest it may be prudent to limit the extent to which professional judgment is used. Reducing the use of clinical judgment may involve restricting not only the types of cases in which overrides would be admissible, but also how much an override would be allowed to deviate from an actuarial assessment. For example, if an actuarial instrument suggests an offender is lower risk for sexual recidivism (e.g., Level 1 in Minnesota), then the highest risk level this offender could be assigned through an override would be Level 2 (moderate risk). In order to develop guidelines that provide greater structure and clarity on when overrides are permissible, future research is needed to examine the conditions under which clinical judgment actually improves classification decisions or, at a minimum, does no worse than actuarial assessments.


Center for Sex Offender Management (CSOM) (2010). Exploring public awareness and attitudes about sex offender management: Findings from a national public opinion poll. US Department of Justice, Office of Justice Programs.


Duwe, G. (2017). Better practices in the development and validation of recidivism
risk assessments: The Minnesota Sex Offender Screening Tool-4 (MnSOST-4).


Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area
under the ROC curve. *Machine Learning*, 77(1), 103-123.


