Public Safety Risk Assessment CLEARINGHOUSE

Why Inter-Rater Reliability Matters for Recidivism Risk Assessment

Summary

One of the important first steps in implementing a risk assessment instrument is to ensure that the instrument is administered consistently by those who collect and score risk factors. This brief introduces the concept of Inter-Rater Reliability (IRR), discusses different factors that influence it, and illustrates the importance of reliably administering assessments.

Key Takeaways

- IRR refers to the level of consistency between different raters conducting a risk assessment instrument. The extent to which the instrument may be administered (in)consistently across raters should be carefully examined because it may compromise the instrument's performance.
- Certain risk factors (i.e., antisocial personality traits) are more subject to interpretations than other risk factors (i.e., official criminal history), thereby lowering IRR. The advantage of adding such risk factors should be balanced by a potential increase in model performance.
- Similarly, the promise of automation in risk assessment can potentially lead to an increase in model performance and is therefore worthy further exploration.

Introduction

Risk assessment is often used to identify persons who are likely to violate the rules of prison or jail, the conditions of community supervision, or the laws of society. Correctional authorities use risk assessments to guide a host of decisions that are intended to enhance public safety and make better use of scarce resources. For example, in "low stakes" risk assessment, instruments have been used to help determine institutional custody levels, prioritization for programming, and the type of community supervision. In "high stakes" risk assessment, where an individual's liberty is at issue, tools have been used to inform decisions related to pretrial release, sentencing, whether individuals should be paroled from prison and whether persons convicted of a sex offense should be civilly committed[1] after serving their sentence.

Automated tools automatically create risk assessment reports based on available data and therefore help ensure that all individuals are assessed consistently. Automated tools can be used at different stages of the criminal justice system including at pretrial, however may require particular data systems in order to be implemented correctly.

Although the uses of risk assessment instruments vary, the way they are scored typically does not vary as much. The scoring method refers to how the items on a risk assessment instrument are populated, which contributes to a total score that is used to determine an individual's risk level (e.g., low, moderate, high). Items on an instrument—for example, age—can be scored manually, usually by

[1] Civil commitment allows a judge or jury to place an individual determined to be a "violent sexual predator" into a secure social/health services facility, even after serving their court-ordered sentence.

correctional staff, or they can be populated through an automated process. Manually-scored risk assessment instruments are, on occasion, mischaracterized as “automated” when they are web-based or used within a software application that calculates a risk score based on the values entered by hand for each item. However, as correctional staff must interpret and input the values for each item on the instrument based on a database review and/or a face-to-face interview with the person being assessed, those instruments are considered manual in this brief.

With very few exceptions, correctional risk assessment instruments rely on a manual scoring method. For example, widely-used instruments such as the Ohio Risk Assessment System (ORAS) and the Level of Service (LS) family of tools utilize a manual scoring process. Under this process, correctional staff would be required to cognitively process the information they gather through database review and/or an interview with the individual, make decisions on what the appropriate response or answer is for each item, and then correctly enter the values for these items on the risk assessment instrument. There are often many staff who administer risk assessment tools regarding the individuals in their custody or control. Even with extensive initial and “booster” training and monitoring for quality assurance, differences among staff in scoring a manual risk assessment tool are inevitable due to a variety of factors, including the subjectivity of the items on the tool, the extent to which staff have been trained, staff workloads, the amount of time it takes to complete an assessment, data entry errors, and so on.

These differences among the staff, or raters, who administer a manually-scored assessment are what is known as inter-rater disagreement. And, inter-rater reliability (IRR) is a measure of how consistently different raters score the same individuals using assessment instruments. This brief reviews the role of IRR within the context of recidivism risk assessment. While IRR has been recognized as a critical component to recidivism risk assessment, mainly because it can potentially affect a tool’s performance in predicting recidivism, there has been very little research on IRR to date.

The main objective of this whitepaper is to raise awareness of the implications of IRR for current practice in risk assessments and its relationship with the predictive validity of risk assessment tools. The next section begins by reviewing the concepts of reliability and validity in the context of risk assessments.

Reliability and Validity

Reliability and validity are the two main properties commonly used to assess the precision and accuracy of measurement. Reliability refers to consistency between raters in scoring an instrument or how well items in an instrument correlate with one another. Both forms of reliability—inter-rater and internal consistency—are important for risk assessment tools, but perhaps the most critical for tools that require manual scoring is IRR (Baird, 2009). Validity, meanwhile, assesses whether an instrument properly measures what it is supposed to measure—in this instance, the likelihood of recidivism.

Although reliability and validity are distinct, the two properties are intertwined. An instrument that is entirely unreliable would, by necessity, have lower validity (Jackson, 2012). For a risk assessment tool to show optimal performance in identifying those most at risk for offending, the instrument must be used consistently by raters scoring the instrument. To the extent that it is not, validity or the ability of the instrument to correctly classify individuals can be compromised (Baird et al., 2013).

Despite the impact it may have on predictive performance, very few studies have evaluated the IRR of recidivism risk assessment instruments. For example, in their review, Desmarais and Singh (2013) found that less than four percent of the studies evaluating risk assessment tools examined IRR. Even worse, Baird (2009) observed, when reliability was reported, it often lacked details or examined the wrong form of reliability (e.g., internal consistency). When IRR has been evaluated, the most commonly used metric has been the Intra-class Correlation Coefficient (ICC). The general rule-of-thumb in assessing ICC is that a score of less than .40 is inadequate; .40-.59 is adequate; .60-.74 is good; and .75 and higher is excellent (Hallgren, 2012).

Of the studies that have evaluated IRR, the findings have varied. Some have reported relatively low IRR for recidivism risk assessment tools (Austin et al., 2003; Baird et al., 2013; Rocque and Plummer-Beale, 2014), whereas others have found good to excellent reliability (Duwe, 2014; Lowenkamp et al.,

2004; van der Knaap et al., 2012). The variation across IRR findings may be due to several factors, such as the nature and type of information collected in the assessment tool being evaluated, the quality of data reporting and recording, the extent of staff training and buy-in into risk assessment, and the data analyses and methods used to evaluate IRR.

All these factors can also influence the predictive validity of assessment tools. In particular, what information is collected and how that information is utilized (i.e., the classification method, which refers to the process in which the values for each item on an instrument are translated into a risk score or a predicted recidivism probability) have recently attracted increased interest among scholarly and practitioner communities alike. An increasing number of instruments rely on computational algorithms for the classification method. These instruments have recently gained popularity for their potential to improve the accuracy of risk predictions (Berk and Bleich, 2013; Breitenbach, Dieterich, Brennan, and Fan, 2009; Duwe and Kim, 2015; Hess and Turner, 2013). These instruments rely on advanced statistical techniques and a large amount of data. As such, these instruments tend to promote a data collection system that does not require a great deal of manual data collection, entry, and processing, which would in turn improve reliability.

The Relationship between Inter-rater Reliability and Predictive Validity

In response to a growing demand to better understand the implications of IRR, Duwe and Rocque (2017) evaluated the impact of reliability on predictive performance for recidivism risk assessment. Relying on assessment data from the MnSTARR, a manually-scored recidivism risk assessment instrument the Minnesota Department of Corrections (MnDOC) developed and began using in 2013, Duwe and Rocque compared the reliability of a manual scoring approach with a fully automated assessment process. Using multiple performance metrics, Duwe and Rocque then evaluated the predictive validity of the two scoring methods—manual and automated—across males and females for four measures of recidivism.

The results showed the MnSTARR was scored with a relatively high degree of consistency by MnDOC staff. Indeed, the ICC values, which ranged from 0.81 to 0.94, would be considered "excellent" according to past research (Hallgren, 2012). But despite this level of IRR, Duwe and Rocque still found the automated assessments had better performance in predicting recidivism for all three dimensions of predictive validity—discrimination, accuracy, and calibration—than those which had been scored manually.[2]

To better understand the relationship between IRR and predictive validity, Duwe and Rocque arranged the male assessment data into quintiles so as to test whether lower ICC values (i.e., lower than the "excellent" threshold of 0.75) have a greater impact on predictive performance. The results from these analyses showed that as inter-rater disagreement increased (i.e., the ICC value decreased), predictive performance significantly decreased. In particular, when the ICC value was in the "good" range (0.60-0.74), the AUC, for example, was .05 lower for the manual scoring method in comparison to the automated process. Applying these findings to the approximately 8,000 imprisoned persons released each year from Minnesota prisons, Duwe and Rocque estimated that, compared to an automated process, a manually-scored instrument with "good" reliability would result in more than 1,000 classification

[2] Predictive discrimination measures the degree to which the instrument separates the recidivists from the non-recidivists. Predictive accuracy assesses how well a model makes correct classification decisions. For example, if a recidivist had a predicted recidivism probability less than a certain threshold, which is typically 50 percent unless otherwise noted, then this individual would be incorrectly classified as a non-recidivist (i.e., false negative). Conversely, if this individual had not recidivated, then she or he would be accurately classified (i.e., true negative). Lastly, calibration measures how well the predicted probabilities from an instrument correspond with the observed outcome (recidivism) being predicted. Largely ignoring the concepts of predictive accuracy and calibration, the extant literature has evaluated the performance of risk and needs assessment instruments by focusing on predictive discrimination, often relying solely on statistics such as the correlation coefficient or the area under the curve (AUC). With values that range from 0 to 1, the AUC 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 and Laws, 2009).

errors each year. That is, the classification errors would consist of “false positives” (i.e., higher-risk individuals who did not recidivate) or “false negatives” (i.e., lower-risk individuals who recidivated).

Implications of MnSTARR Automation Study for Recidivism Risk Assessment

Given these findings, Duwe and Rocque highlighted four notable implications for recidivism risk assessment. First, Duwe and Rocque demonstrated that inter-rater disagreement significantly impacts predictive performance, and the size of the impact is similar to that found for the classification method (Duwe and Kim, 2016). To illustrate, let us revisit the example discussed above regarding the 8,000 released individuals. If a manually-scored tool with “good” reliability yields approximately 1,000 more classification errors than a fully-automated process, then an instrument based on Burgess methodology (the worst performing classification method) would produce roughly the same number of errors compared to a tool developed with a machine learning algorithm (the best performing classification method). Therefore, use of a fully-automated instrument developed with a high-performing algorithm could result in at least 2,000 fewer classification errors, or about one-fourth of the individuals released from prison.

Second, Duwe and Rocque suggest that prior thresholds for evaluating inter-rater reliability may be overly optimistic. For example, if an ICC value in the “good” range results in a significant drop in predictive performance, then perhaps an ICC value in this range might not be so good after all. They proposed the following thresholds for assessing IRR within the context of manually-scored recidivism risk assessment tools: 0.95 and above is excellent; 0.85-0.94 is good; 0.75-0.84 is adequate; and below 0.75 is poor.

Level of Agreement	ICC Level
Poor	Less than .74
Adequate	0.75-0.84
Good	0.85-0.94
Excellent	Higher than 0.95

Third, comparing the IRR results for the MnSTARR with a recent study that found an ICC value of 0.65 for the LSI-R (Rocque and Plummer-Beale, 2014), Duwe and Rocque maintain that using objective items (as opposed to more subjective items that require more interpretation) is critical for achieving adequate reliability. For instance, rather than asking an individual in a one-on-one interview whether she or he is a “social isolate”, it may be better to simply measure whether she or he received any visits in prison. Or, rather than asking an individual how many of his/her friends have criminal records, it may be better for recidivism prediction purposes to record whether s/he has been identified as an active gang member.

Finally, Duwe and Rocque suggest that rather than redoubling efforts to improve the scoring of manual instruments, it would be more prudent, over the long term, to invest in automation. Compared to manual scoring, an automated process improves predictive performance by eliminating inter-rater disagreement, increases assessment capacity, and greatly reduces the staff time needed for ongoing training, quality assurance checks and, most important, administering the assessment. Due to the increased efficiency of automation, Duwe and Rocque found that automating the MnSTARR will deliver a highly favorable return on investment. For every dollar spent on automating the MnSTARR, there will be an estimated return of nearly $22 within five years, totaling $2.8 million. Given these results, Duwe and Rocque proposed that automated risk assessment may be regarded as a reinvestment strategy in which the time and cost savings can be diverted into other areas to improve correctional policy and practice.

Conclusion

Other industries that commonly make risk assessment decisions (e.g., financial lending, insurance, healthcare, etc.) have, over the last few decades, increasingly abandoned manual assessment processes in favor of automated ones. Reasons for the shift towards automation include not only more objective, reliable, and valid risk assessment decisions, but also greater efficiency and cost-effectiveness (Matthews and Hodach, 2012; Straka, 2000). Given that automated risk assessment is still a relatively new idea in criminal justice, the vast majority of risk assessments administered are still scored by hand. As long as manual scoring remains the prevailing practice in criminal justice settings, then IRR should warrant much more attention than it has received in the past. The following areas, in particular, deserve careful consideration for future research.

First, when prior studies have evaluated the performance of risk assessment tools, they have typically focused only on how accurately an instrument predicts recidivism. However, evaluations of risk assessment tools should include IRR. There have been relatively few IRR evaluations of manually-scored risk assessment instruments, perhaps because they are generally more time-consuming than validation studies. Yet, given the findings from the Duwe and Rocque (2017), future studies should begin examining the relationship between IRR and predictive performance.

Second, while the Duwe and Rocque (2017) study showed that inter-rater disagreement leads to worse predictive performance for recidivism, is the same true for tools that also attempt to assess criminogenic needs? Among the manually-scored risk and needs assessment instruments, does inter-rater error compromise the ability to accurately identify needs areas as well? That is, does the inconsistency among raters result in the misclassification of needs for some individuals? In addition to examining whether risk and needs assessments accurately identify needs for individuals, future research should investigate how IRR interacts with needs prediction. In doing so, we may gain an understanding of whether inter-rater disagreement has a similar impact on needs assessment as it does on recidivism prediction.

References

Austin, J., Coleman, D., Peyton, J., & Johnson, K.D. (2003). Reliability and validity study of the LSI-R risk assessment instrument. Washington, DC: The Institute on Crime, Justice and Corrections (Submitted to the Pennsylvania Board of Probation and Parole).

Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime & Delinquency.

Baird, C., Healy, T., Johnson, K., Bogie, A., Dankert, E.W., & Scharenbroch, C. (2013). A comparison of risk assessment instruments in juvenile justice. National Council on Crime & Delinquency. Retrieved from https://www.nccdglobal.org/sites/default/files/publication_pdf/nccd_fire_report.pdf.

Berk, R.A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*, 513-544.

Breitenbach, M., Dieterich, W., Brennan, T., & Fan, A. (2009). Creating Risk-Scores in very imbalanced datasets: Predicting extremely low violent crime among criminal offenders following release from prison. In Y.S. Koh & Rountree, N. (Eds.), *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection* (pp. 231–254). Hershey, PA: Information Science Reference.

Burgess, E.W. (1928). Factors determining success or failure on parole. In A.A. Bruce, E.W. Burgess, J. Landesco, & A.J. Harno (Eds.), *The workings of the indeterminate sentence law and the parole system in Illinois*, (pp. 221–234). Springfield, IL: Illinois State Board of Parole.

Desmarais, S.L., & Singh, J.P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Retrieved from https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf

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

Duwe. G. & Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review*. DOI: 10.1177/0887403415604899.

Duwe, G. & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections: Policy, Practice and Research*.

Duwe, G. & Rocque, M. (2016). A jack of all trades but a master of none? Evaluating the LSI-R's performance in assessing risk and need. *Corrections: Policy, Practice, and Research*. http://dx.doi.org/10.1080/23774657.2015.1111743.

Duwe, G. & Rocque, M. (2017). The effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy*.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, 52(1), 178-200.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23-34.

Hess, J. & Turner, S. (2013). *Risk Assessment Accuracy in Corrections Population Management: Testing the Promise of Tree Based Ensemble Predictions*. Center for Evidence-Based Corrections: The University of California, Irvine.

Jackson, S. L. (2012). *Research Methods: A Modular Approach*. Stamford, CT: Cengage.

Lowenkamp, C.T., Holsinger, A.M., Brusman-Lovins, L., & Latessa, E.J. (2004). Assessing The inter-rater agreement of the level of service inventory revised. *Federal Probation*, *68(3)*, 34–38.

Matthews, M.B., & Hodach, R. (2012). Automation is key to managing a population's health. *Healthcare Financial Management*, *April*, 1-8.

Rocque, M., & Plummer-Beale, J. (2014). In the eye of the beholder? An examination of the inter-rater reliability of the LSI-R and YLS/CMI in a correctional agency. *Journal of Criminal Justice*.

Straka, J.W. (2000). A shift in the mortgage landscape: The 1990s move to automated credit evaluations. *Journal of Housing Research*, 11: 207-232.

Thornton, D. & Laws, D.R. (2009). *Cognitive Approaches to the Assessment of Sexual Interest in Sexual Offenders*. Hoboken, NJ: Wiley.

Van der Knaap, L. M., Leenarts, L. E., Born, M. P., & Oosterveld, P. (2012). Reevaluating interrater reliability in offender risk assessment. *Crime & Delinquency*, 58(1), 147-163.

The **Public Safety Risk Assessment Clearinghouse** is a one-stop resource that provides practitioners and policymakers with up-to-date and objective information about risk assessment as well as training and technical assistance on its use.

For more information, please visit: https://psrac.bja.ojp.gov/

Suggested citation for this publication:

Duwe, Grant (2017). *Why inter-rater reliability matters for recidivism risk assessment (Policy Brief Number 2017-03).* Washington, DC: The Public Safety Risk Assessment Clearinghouse.

This project was supported by Grant No. 2015-ZB-BX-K004 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.