Washington State Institute for Criminal Justice:

Research in Brief

Research Brief
Inter-Rater Reliability (IRR) of the Static Risk Offender Need Guide for Recidivism (STRONG-R)

Douglas Routh, M.A.
Zachary Hamilton, Ph.D.
Washington State University

Achieving a high level of inter-rater reliability is an important performance metric for risk assessment tools. A high level of IRR indicates that correctional staff members similarly score offenders using a given tool. For practice, IRR translates to staff members being well-trained and a tool with clearly interpretable content. This is a metric that is of particular importance for new tools like the STORNG-R, where training modules are newly established. The purpose of the current study was to assess the IRR between raters on the scoring items of the STRONG-R.

Method

Data collection

IRR analyses were completed for WADOC staff recently trained using the STRONG-R. Video recorded interviews were created for four offenders. On June 1, 2016 raters (staff members) observed the recorded interviews. Each rater scored all four interviews. Raters completed paper copies of the assessment, which were then digitized for analysis.

Analytic Plan

While there are several ways to assess IRR, a two-way random-effects intra-class correlation (ICC) coefficient with absolute agreement was selected as the most appropriate method based of the sample characteristics (Shrout & Fleiss, 1979). First, the ICC accounts for more than two raters. Second, both the offenders and raters were selected from a larger population, which introduces variance. The two-way random-effects model accounts for this variance.

The STRONG-R is completed in two stages. First a Criminal Conviction Record (CCR) is auto-populated from an automated data draw from several criminal history data sources. This record review process possesses a near 100% accuracy in addressing 27 criminal history items. Thus, it was important to separate the IRR of the full tool with that of the interview items alone. The ICC coefficients are provided for scoring items of the STRONG-R, both including and excluding the criminal history measures. The subjective items of the interview portion were isolated to determine correctional staff members' understanding when scoring the STRONG-R.

Results

Results from the IRR test are presented in Table 1. The ICC coefficients are provided for scoring items of the STRONG-R, both including and excluding the criminal history measures. The criminal history measures are automatically populated by software following a file review and posses a near 100% accuracy rating due to routinized WADOC record reviews. Essentially, the criminal history measures posses no subjectively compared to the interview portion of the STRONG-R, therefore subjective items were isolated to determine correctional staff members' understand when scoring the STRONG-R.

Table 1 – ICC Coefficients for STRONG-R IRR (N=33)

	Scoring Items	Scoring Items (Criminal History Removed)
Offender 1	.89	.53
Offender 2	.88	.59
Offender 3	.91	.67
Offender 4	.87	.63
Average IRR	.89	.61

Note: The average IRR was calculated by averaging all four offenders' ICC values.
ICC: acceptable agreement .40-.59, good agreement .60-.74, excellent/strong agreement .75-1.00 (Cicchetti, 1994).

As seen from Table 1, the level of consistency between all thirty-three raters differs is different based on the inclusion or exclusion of the criminal history measures. With the inclusion of the criminal history measures, the level of agreement ranges between .87 and .91, indicating excellent or strong consistency between raters. The average ICC was found also found to be excellent (Mean ICC = 0.89). As indicated, this is likely due to the fact that information for the criminal history measures are automated and thoroughly obtained from criminal records and the remaining domains are assessed via an interview. When the criminal history measures are excluded, the level of agreement ranges between .53 to .67 and the mean is 0.61, which indicates acceptable-to-good agreement between raters, acceptable agreement for Offenders 1 and 2 and good agreement for Offender 3 and 4. There was little difference in ratings between the male offenders in prison or under community supervision (Offenders 2 and 4). However, there was noticeable difference in ratings for female offenders in prison and community supervision (Offender 1 and 3). Community supervision has a higher level of agreement than in prison.

Conclusion

This study has demonstrated acceptable to excellent reliability between raters utilizing the STRONG-R risk assessment tool. It is important to note that the interaction between the rater and the offender can be a source of variance between raters. Personal judgments must be made

based on the interview section of the STRONG-R will differ between raters. Similarly, a rater's ability to utilize the tool, interact and communicate with offenders, and allowing offenders to open up about their personal history will also be a source of variation.

With that said, the WADOC is still in the initial stages of implementing the STRONG-R and training efforts. This initial test of IRR will be used to improve these efforts prior to the initial launch of the STRONG-R. Future research is needed to determine the stability of the inter-rater reliability. One possible avenue of research will implement a test-retest approach similar to Farabee and colleagues' (2010) approach to validating the COMPAS risk assessment tool. Reassessing the IRR is critical to establishing consistent results over time. Also, reassessing the IRR can detect any issues that raters may be experiencing, and modify training to improve consistency. Continuing to improve assessor's accuracy will improve risk category and efficiency of supervision and treatment.

References

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4): 284–290

Farabee, D., Zhang, S., Roberts, R.E.L., & Yang, J. (2010). COMPAS validation study: final report. Technical Report. Semel Institute for Neuroscience and Human Behavior: Los Angeles, CA.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

For further details about the STRONG-R IRR research findings, WSU Researchers can be contacted at zachary.hamilton@wsu.edu