




Never Going to Let You Down: Preventing Predictive Shrinkage via the STRONG-R Assessment Method

Zachary Hamilton, Alex Kigerl, Baylee Allen, John Ursino & Amber Krushas

To cite this article: Zachary Hamilton, Alex Kigerl, Baylee Allen, John Ursino & Amber Krushas (07 Aug 2024): Never Going to Let You Down: Preventing Predictive Shrinkage via the STRONG-R Assessment Method, Justice Quarterly, DOI: [10.1080/07418825.2024.2386637](https://doi.org/10.1080/07418825.2024.2386637)

To link to this article: <https://doi.org/10.1080/07418825.2024.2386637>

 View supplementary material [↗](#)

 Published online: 07 Aug 2024.

 Submit your article to this journal [↗](#)

 View related articles [↗](#)

 View Crossmark data [↗](#)



Never Going to Let You Down: Preventing Predictive Shrinkage via the STRONG-R Assessment Method

Zachary Hamilton^a, Alex Kigerl^b, Baylee Allen^b, John Ursino^b and Amber Krushas^c

^aDepartment of Criminology & Criminal Justice, University of Nebraska – Omaha, Omaha, Nebraska, USA;

^bUniversity of Nebraska – Omaha, Omaha, Nebraska, USA; ^cDepartment of Criminal Justice 4505S, University of Nevada Las Vegas, Las Vegas, Nevada, USA

ABSTRACT

Risk-needs assessments (RNAs) are an evidence-based practice used by practitioners to assign supervision and programming. While foundational to day-to-day practices, these tools are typically applied ‘off-the shelf,’ and are mistakenly assumed to demonstrate equivalent prediction accuracy regardless of location or population. Although researchers and providers are aware these tools experience performance shrinkage, the issue is commonly ignored. In 2016 the Tennessee Department of Correction (TDOC) collaborated with the developers of the Static Risk Offender Needs Guide – Revised (STRONG-R) to create a staged development of a locally developed risk needs assessment. Using propensity score matching, a proxy sample was used to create an initial version of the STRONG-R, representing a TDOC-like sample. This version was deployed in 2017. Following data collection, developers recrafted the tool with local data to make Version 2.0. Findings demonstrate improved performance using this innovative method, effectively eliminating performance shrinkage for the TDOC STRONG-R.

ARTICLE HISTORY

Received 30 January 2024

Accepted 25 July 2024

KEYWORDS

Assessment; Recidivism; Proxy; Prediction

Introduction

Over the past 40 years, risk-needs assessments (RNAs) have been identified as a key mechanism for classification and programming recommendations for individuals under correctional supervision and recognized as an evidence-based practice (EBP) (Taxman, 2018). The benefits of RNAs are substantial, providing mechanisms of standardization, reducing biases and, in turn, decreasing justice system involvement (Lowenkamp & Latessa, 2004). As a result of the positive consensus around their utility, many states have mandated the adoption of RNAs within state correctional populations (St John et al., 2020).

Given the high demand for RNAs, most agencies have implemented RNAs that were developed elsewhere, outside of their jurisdiction, with the belief that when these instruments are applied “off-the shelf,” they will function similarly regardless of location or population. Unfortunately, we now know this is not true (Hamilton, et al., 2022).

Too often, the providers of these tools, and in some cases even the developers, convey their validation statistics as stable and unwavering, suggesting that tool performance will never lessen, despite jurisdiction distinctions. Thus, many agencies have been misinformed, which means the vast majority of the RNAs implemented in corrections agencies throughout the United States in the last twenty years have been applied to populations they were not originally developed to assess (Duwe, 2014; Hamilton et al., 2021). This is problematic, as the populations upon which these tools were developed often differ from those to which they are later applied, and where these new populations, or “jurisdictions” possess myriad differences, such as, different statutes, policies, recidivism definitions, law enforcement priorities, and a multitude of additional factors, that combine to reduce an assessment’s accuracy (Hamilton et al., 2021). Without adjustment, an “off-the-shelf” application of an RNA will suffer accuracy issues in its new locale—a phenomenon known as “prediction shrinkage” (Casey et al., 2014). Further, “prediction shrinkage,” or inaccuracy, creates misclassification of risk categories and supervision assignments, and ultimately can undermine an agencies’ population management, thereby reducing staff and public safety.

Attempting to avoid issues of shrinkage and to provide an RNA with content tailored to the jurisdiction and local population, some agencies opt to develop tools locally (Schwalbe et al., 2006). The Washington State Department of Corrections (WADOC) developed its own RNA using local data customized to fit their recidivism definition, items, and response weights (Barnoski, 2004). In the last 10 years, several examples of homegrown RNAs have been developed and demonstrated strong predictive strength (Duwe, 2014; Hamilton et al., 2021).

Yet, developing tools locally takes significant time and resources and when adoption timelines are pressured by state mandates, local RNA development may not be feasible, and an agency may feel forced to adopt an off-the-shelf tool. Further, after an assessment is created, developers may sell or allow a tool to be provided by a private or public entity. These providers may also offer software and training for the tool. However, many developers, agencies, and tool providers avoid/neglect local revalidation, potentially turning a blind eye to needed updates that would improve tool performance locally (Fazel et al., 2022).

Recently, some scholars have used “optimization techniques” to overcome the described limitations and improve tools’ prediction accuracy. Optimization broadly refers to removing, adding, or modifying RNA items of existing tools to meet the unique needs and characteristics of a jurisdiction (Hamilton et al., 2017). Practices such as changing item/response language, removing items, and altering risk level thresholds have been used, sporadically, for over a decade (Duwe, 2021). More advanced techniques (e.g. applying statistically developed response weights) have been used to refine assessment models, using local data, which has been shown to improve recidivism prediction (Duwe, 2021; Hamilton et al., 2021).

Despite these advances, there is a lack of research explaining tool development efforts and applications of RNAs in new jurisdictions. The current study seeks to describe optimization stages and techniques used to create the Static Risk Offender Needs Guide—Revised (STRONG-R) assessment tool for the Tennessee Department of Correction (TDOC). Specifically, this study describes how data collected from a

development sample informed a proxy sample, which was used to develop an initial version of the TDOC STRONG-R. After deployment and data collection, the tool's predictive performance was evaluated. Finally, we describe recalibration efforts used to create a new, more accurate version of the TDOC STRONG-R. In describing these endeavors, the current study seeks to outline best practice strategies in the development and refinement of RNA tools.

Risk Assessment Development

Within the evidence-based movement, the Risk-Need-Responsivity (RNR) model provides a foundation for the development of assessments that aid practitioners' decisions (Andrews et al., 1990; Taxman, 2018). This model outlines the utility of assessment for correctional case management. Tools sum response values across static and/or dynamic risk and protective factors, creating a distribution where individuals with higher scores possess a greater likelihood of recidivism. Assessment scores are used to recommend higher risk individuals for more intensive interventions, while system contact for lower risk populations should be limited (Andrews & Bonta, 2010). Thus, risk assessments provide a method of standardizing an agency's supervision guidelines, hopefully reducing bias associated with the idiosyncrasies of human decisions and help guide judgements such as diversion, supervision intensity, and early release (Viglione, 2019).

Contemporary RNAs

Contemporary tools are believed to have begun with the Level of Service (LS) tools, developed by Andrews and colleagues in 1982 to measure the recidivism risk for a sample of 112 halfway house participants in Ontario and Manitoba, Canada (1995). Their initial Level of Service Inventory (LSI) contained 59¹ Burgess scored (0/1) items and was revised twice, creating the LSI- Revised (LSI-R) (1995) and the Level of Service/ Case Management Inventory (LS/CMI) (2004). With few alternatives available, the LS tools were dubbed the "gold standard" of RNAs and proliferated across Canada, the U.S., and throughout the world (Wormith & Bonta, 2018). The tool's proliferation is also facilitated by a provider—Multi-Health Systems (MHS)²—a publishing house, and provider of hundreds of tools across many industries, who purchased the rights to license and sell the LS family of tools, as well as provide software, training, and certification to adopting agencies. Yet, despite their proliferation, the LS tools' scoring is relatively unchanged from the theoretical model piloted on a small group of White, male, Canadian halfway house participants nearly 40 years ago. Unfortunately, when applied off-the-shelf in other jurisdictions, with populations that differed both geographically and demographically from the initial LSI sample, the tools have often suffered from predictive shrinkage (Olver et al., 2014).

¹The original LSI contained 59 items after adding criminal history items and removing some demographic items. Later, four items delineating between probation and parole conditions were removed to create the 54 item LSI-R (Wormith & Bonta, 2018).

²More information on MHS can be found at <https://storefront.mhs.com/collections/areas-of-assessment>.

Historically, correlation coefficients (r) were used to assess the association between a tool's risk score and recidivism. More recently, Area Under the Curve (AUC) statistics are used to measure predictive validity. Rice and Harris (2005) converted r and Cohen's d values, and provided effect size ranges that are commonly used to evaluate the strength of a tool's predictive accuracy, where 0.50 to 0.55 is "negligible", 0.56 to 0.63 is "weak", 0.64 to 0.70 is "moderate", and 0.71 and above is considered a "strong" level of predictive accuracy. Further, Hamilton and colleagues (2022) indicated that AUC effect sizes, on average, represent a 6-point range. Thus, when comparing RNA models an AUC increase of six points (or greater) is considered "substantial" improvement and reductions of six points (or more) represent substantial shrinkage.

Notably, in the original development sample, the LSI possessed "strong" predictive accuracy (AUC = 0.77) (Wormith & Bonta, 2018). Yet, in a 2014 meta-analysis, Olver et al. (2014) found that when the LS tools were applied to other Canadian jurisdictions, the predictive accuracy decreased 8% (AUC = 0.69) indicating only a "moderate" level of predictive accuracy. This 8% reduction in the AUC, or predictive strength, represents a substantial, reduction in predictive accuracy. Moreover, when the LS tools were applied in correctional agencies in the United States, predictive shrinkage dropped even more dramatically, and found to reduce by 16 percentage points, and possessing only "weak" predictive accuracy (AUCs = 0.61). Additionally, the LS tools are poor predictors of violent recidivism, regardless of location, possessing only moderate predictive accuracy among Canadian samples (AUC = 0.65), and validated to be weaker for US samples (AUC = 0.56). The authors note that the reduction in predictive accuracy is likely due to reliability and training issues perceived to be more problematic in US applications. Further, they outline the decentralized nature of recidivism outcome collection, where Canadian samples make use of a centralized criminal history data base (Olver et al., 2014).

Developers of another widely used tool, the Ohio Risk Assessment System (ORAS), created a community supervision tool (CST), on a relatively small sample of probationers in Ohio ($N=678$) (Latessa & Lovins, 2010). The CST possessed moderate predictive accuracy (AUC = 0.70) on the ORAS development sample. The ORAS tools are provided by the University of Cincinnati Correctional Institute (UCCI)³, a university institute that licenses and sells the software, training, and certification services to agencies adopting the tools. Like the LS family of tools, this tool has been applied off-the-shelf in several other jurisdictions. For example, in Indiana, the ORAS (or IRAS, as they titled there) identified weak predictive performance (AUC = 0.59) (Latessa et al., 2013).⁴ When the tool was applied in Texas (or TRAS, as they titled there), it was not applied strictly as off-the-shelf, but was customized, where the item content was modified and questions were added to better reflect the risk and needs of the comparatively larger Hispanic population. While the predictive performance still shrank

³More information on UCCI can be found at <https://cech.uc.edu/about/centers/ucci.html>.

⁴We note that another tool, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), has far less published findings available. Yet, where available, the COMPAS was similarly constructed with respective development samples and applied off-the-shelf in other jurisdictions, and with few adjustments to improve its prediction of recidivism locally (Northpointe, 2015). Notably, the COMPAS tools commonly identify 'strong' predictive accuracy ratings with development samples, and moderate prediction strength in sites following off-the-shelf application (Brennan and Dieterich, 2018). The COMPAS is provided by Equivant, a private entity that delivers the assessment, training, and certification for a licensing free (see <https://www.equivant.com/practitioners-guide-to-compas-core/>).

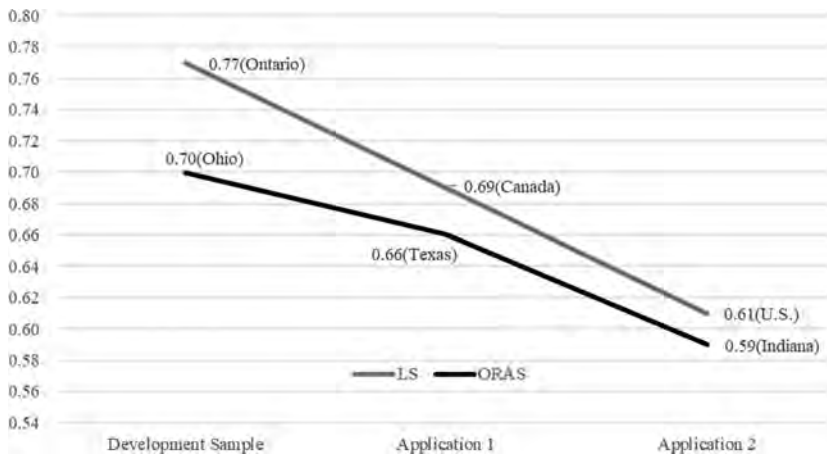


Figure 1. LS and ORAS development & application AUCs. *Note:* LS: Level of Service; ORAS: Ohio Risk Assessment System.

(AUC = 0.66), the customization seemed to play a role in reducing the impact of the shrinkage witnessed in the Indiana application (Lovins et al., 2018).

A common theme among risk assessment development and application is that tools work best in the original development sample. When applied elsewhere, predictive shrinkage occurs. The shrinkage is more substantial if the new jurisdiction's population differs considerably from the development population. However, when the tool is adjusted for the local population, shrinkage is less pronounced. In [Figure 1](#), we provide an illustration of the predictive accuracy of development samples and additional LS and ORAS applications.

RNA Best Practice

RNA development is commonly completed in stages, where 1) developers generate a pool of potential items, 2) an initial tool is created and piloted with a development sample, 3) following data collection, unnecessary items are removed and difficult responses clarified, 4) a final model is established and deployed, and 5) after sufficiently tracking recidivism outcomes (i.e. two-plus years), the tool is validated (Hamilton et al., 2017). Best practice would dictate that all, or a portion, of these steps (i.e. 3 through 5) are repeated every three-to-five-years to ensure the tool is still performing as expected (Hamilton & Campbell, 2013). Following validation, developers may create updated versions of the tool to improve performance (Bucklen et al., 2010). Unfortunately, most agencies adopt tools off-the-shelf, infrequently validating their tools locally. Further, RNA providers and developers rarely adjust tools, failing to acknowledge that the tools and services they provide must be adjusted to meet local agency needs (Fazel et al., 2022).

RNA Innovations

Recently, research has demonstrated the importance of developing and adjusting tools to fit the local population (Hamilton et al., 2021). In 1999, the Washington

Legislature mandated the adoption of an RNA to specify graduated supervision levels. With limited time and options, the WADOC adopted the LSI-R (Barnoski & Drake, 2007). However, after local validation the LSI-R underperformed, and a homegrown tool was created (Barnoski, 2004). In the years to follow, several “customized” tools have similarly been developed, increasing predictive accuracy (Duwe, 2014; Hamilton et al., 2021).

Recent innovations include 1) customization, 2) optimization, 3) localization, 4) gender-responsivity, 5) outcome-specific modeling, and 6) automation (Hamilton et al., 2017). While customization refers to adapting RNA items to fit the agency’s local diction, crime categories or statutes, and add/remove items to address system variations, optimization refers to selecting and weighting item responses, using statistical models to provide differing scoring values to items with greater prediction importance. Localization is the process of recalibrating the items and weights of an off-the-shelf tool using locally assessed subjects. Related to localization, some tools adjust risk level thresholds to fit agency needs, termed cut point “norming.” Gender-responsive tools isolate male and female samples, creating separate tools by gender. Outcome-specific RNAs create separate statistical models by selecting and weighting items to predict general and more distinct recidivism types (i.e. violent, property, drug). Finally, automation can refer to several processes that can result in tool improvements, including for example funneling routinely collected data (i.e. court records, infractions, programming), to auto-populate assessment responses, thereby reducing assessor labor, increasing reliability of data collected, and improving the accuracy of answers chosen and their associated score.

Proxy Sampling

As indicated, many agencies are required to adopt assessment tools. Timelines following mandates may force an agency to adopt a tool off-the shelf, limiting opportunities to use local data to better inform tool performance. However, social scientists have long employed sampling methods when confronted with similar limitations (Cook et al., 2002). Proxy sampling creates a statistical model of similar individuals to mimic the target population (e.g. Mara, 2002). For instance, Rivlin et al. (2012) used individuals with near-lethal attempts as proxies for those who had completed suicide to examine the causes and prisoner suicide prevention strategies. Additionally, given the limited data for long-term care prison populations, Merianos et al. (1997) used proxy samples to assess care considerations for aging inmates by surveying elderly inmates across demographic and health outcomes and then administering the survey to people in the community that matched the incarcerated sample on key demographic characteristics.

The Tennessee STRONG-R

In 2016, the Tennessee State Legislature enacted the Public Safety Act (PSA) (House Bill 2576), requiring the Tennessee Department of Correction (TDOC) to use a validated RNA to evaluate supervised individuals’ needs and their risk posed to public safety. Specifically, the PSA compelled TDOC, community corrections agencies, the board of

parole, and the courts to use the RNA in determining programming, treatment, release decisions, and post-prison supervision (Norris et al., 2016). With a six-month implementation deadline, developing a homegrown tool from scratch was not feasible.

Although TDOC had its choice of contemporary RNAs, a decision was reached to build a customized tool that would better reflect their population personal conversation with TDOC director. Specifically, stakeholders raised concerns regarding the PSA's requirement to determine who was appropriate for parole release and community supervision, and where the adoption of an off-the-shelf tool would likely struggle to accurately identify the risk cut points needed when making these important decisions. Further, the PSA provided an opportunity to create graduated sanctions, where again, a customized tool design would be better equipped to establish behavioral/risk thresholds used to set sanctioning levels.

The rationale for building a customizable instrument was based on the understanding that a homegrown tool would possess a greater ability to accurately predict risks and needs for TDOC supervised individuals. Moreover, by developing a tool that utilizes the Tennessee's population, predictive accuracy will continue to grow, as data is collected from the tool, and newer versions are developed. Of all the contemporary assessments available, the STRONG-R represents the only fully customizable tool (Hamilton et al., 2018). Notably, to ensure a strong assessment of recidivism risk, the STRONG-R developers redesign the assessment in each jurisdiction, using agency records and vetting the tool with subject matter experts to ensure local item relevance. This process ensures each application of the tool is optimized for the local population.

Thus, in 2016, TDOC contracted with the creators of the STRONG-R - Vant4ge - to collaboratively develop, automate, implement, and support a version of the tool specified to the Tennessee population. The research team created a hybrid RNA development process, using indicators gathered from TDOC administrative records. Existing responses were cross-walked and then used to create a matched selection of individuals previously assessed on nearly 100 STRONG-R items.⁵ This proxy sample represented STRONG-R assessed subjects that were statistically similar to TDOC supervised individuals. Proxy subjects were then used to develop an RNA that was customized, optimized, and made use of local data to create a gender-responsive, outcome-specific Tennessee version of the STRONG-R. In reference to the five-stage development process described, the proxy sample procedure skips the requirement of generating new items and pilot testing with a development sample, which provided TDOC a more feasible alternative to creating a homegrown tool from scratch.

Using this hybrid process, an initial version (1.0) was created and deployed in 2017. Assessment and recidivism data was collected for three years, allowing for a validation of Version 1.0. This newly collected data was then used to develop eight new recidivism prediction models, representing Version 2.0. Notably, the STRONG-R was created by incorporating each of the innovations described above (i.e. customization, optimization, localization, gender-responsivity, outcome-specific modeling, and automation, including programming the entirety of a subject's criminal history into a conviction record, which auto-populates this content within the STRONG-R assessment) (Hamilton et al., 2017).

⁵While the STRONG-R response pool is larger than most RNAs, it retains more items to cover a wider range of non-criminal history items that can be selected during the development process.

The current study describes the staged development, methods, and findings of the Tennessee STRONG-R. Comparing multiple versions of the STRONG-R, we provide researchers and agencies with a set of best practices to improve RNA accuracy.

Methods

In this section, we describe the phased development and evaluation of the TDOC STRONG-R, Versions 1.0 and 2.0. Development consisted of data gathering and the use of TDOC measures and subjects, cross-walked with existing STRONG-R items. We then describe the creation of a proxy sample and development of Version 1.0. Next, we describe data gathered following deployment of Version 1.0 and the creation of Version 2.0. Finally, we outline the methods used to evaluate the STRONG-R's performance.

Proxy Sample Development

The initial STRONG-R was developed with a WADOC sample of individuals convicted of felonies and gross misdemeanors and supervised in the community from 2008 through 2012. The WADOC development sample consisted of 56,606 eligible subjects, who were assessed *via* the STRONG-R and possessed a two-year recidivism follow-up. In 2016, we first gathered TDOC administrative data to create our proxy sample. We identified a sample of TDOC subjects that were released or supervised in the community in 2013 and 2014, allowing for a two-year recidivism follow-up period. Propensity score modeling (PSM) was utilized to match a contemporary sample of TDOC offenders with the STRONG-R development sample. Based on demographics, prior criminal history, and other correctional predictors, a total of 16 STRONG-R items were available and similarly measured in the WADOC community development sample ($N=28,153$). PSM was selected as the matching technique to create the proxy sample, as this method has the ability to select a sample that is statistically similar to the experimental condition (Guo et al., 2020), or in this case select a WADOC group that is similar to the TDOC sample. Bivariate comparisons were completed and standardized difference (STD)⁶ tests were assessed, where an absolute value greater than 20 indicated imbalance (Rosenbaum & Rubin, 1985). Further, AUCs were assessed pre- and post-match, used to identify predictors' ability to assess differences between WADOC and TDOC subjects.

The PSM was completed with a one-to-one, "greedy" matching procedure using a selection caliper (< 0.1 SD). A total of 26,364 STRONG-R development sample of subjects were selected and matched to their TDOC counterparts. While 43% of indicators substantially differed prior to the match, 0% of STDs exceeded 20 in the post-match comparisons. Also, pre-match indicators collectively identified a "strong" ability to predict TDOC versus WADOC (AUC = 0.71) prior, and a negligible prediction post-match (AUC = 0.52). Due to space considerations, a more detailed description of the PSM procedure and pre/post-match descriptives are provided in [Appendix A, supplementary material](#).

⁶The following formula, created by Rosenbaum and Rubin (1985), was used to calculate the standardized absolute differences in percentages, $\frac{|X_t - X_c|}{\sqrt{\frac{s_{2t} + s_{2c}}{2}}}$, where X_t and X_c are the means for the treatment and control groups, respectively, and s_{2t} and s_{2c} are the variances.

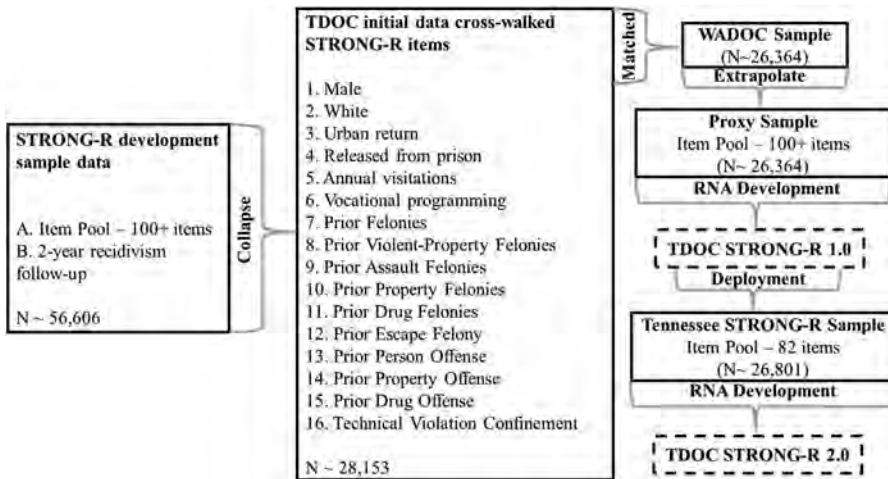


Figure 2. TDOC STRONG-R Proxy sample creation process. *Note:* TDOC: Tennessee Department of Corrections; WADOC: Washington State Department of Corrections.

Extrapolated, the matched WADOC subjects possessed the full pool of STRONG-R items and provided the closest approximation of TDOC subjects assessed using the STRONG-R and formed our proxy sample. This process is outlined in [Figure 2](#).

TDOC STRONG-R Development

In 2016, the described proxy sample was used to statistically weight STRONG-R items to predict four outcomes—violent, property, drug, and “any” felony recidivism. This process was completed for subsamples of males and females, for a total of eight models. To create models, we used ridge regression to select and weight items, optimized for their given sample. The ridge regression method optimizes prediction in high dimensionality datasets, removing prediction noise while maximizing performance (Harrell, 2001). Further, we developed customized processes in R to eliminate items that with negative coefficients, where all protective responses (items/responses anticipated to reduce recidivism) were reverse coded. Selected items’ coefficient values were used as response weights, with larger values indicative of greater importance, and models removed items with no contribution to a model’s AUC.

All models were developed with k-fold validation, where processes were completed by randomly partitioning the dataset into 10 parts/folds (see Kohavi, 1995). A model is trained to nine parts, while the remaining part is used for testing. This process is repeated 10 times, with a different tenth of the dataset set aside for testing, and predictive metrics from each of the 10 test parts are then averaged. A total of 82 items were selected to be included in at least 1 of the 8 computed models. To conserve space, STRONG-R item descriptive statistics are provided in [Appendix B, supplementary material](#)⁷ and response weights are provided in [Appendix C, supplementary material](#).

⁷With 96 items it is not feasible to operationalize each response here. We refer readers to source materials (see Hamilton et al., 2016) or contact the Washington State or Tennessee DOCs for manuals and supporting materials.

TDOC STRONG-R 1.0 Validation & 2.0 Development

Once the development of the TDOC STRONG-R 1.0 was complete, Vant4ge implemented the tool and its supporting software platform statewide, including all of TDOC's divisions, which include probation, prison, parole, and community corrections. Specifically, Vant4ge's tool automation software platform maximized assessment efficiency, reduced redundancies, and supported local practice and PSA mandates to improve overall system outcomes. Some of the key software features that help to ensure the highest quality and fidelity in the administration and utilization of the STRONG-R include: an application program interface (API) to auto-populate criminal history item responses, a design that allows for scoring of multiple prediction models (i.e. violent, property, drug)⁸, inter- and intra-item logic that reduces incongruence of item responses and minimizes redundancy⁹, embedded user-guidance to support inter-rater reliability, an automation of needs assessment results into real-time programming recommendations, and sophisticated reporting technologies that support agency-level decision making related to rehabilitative programming resource allocation.

Following deployment of Version 1.0, assessments were conducted between January 4, 2017, and August 31, 2020, which resulted in sample of 46,516 individuals. Item descriptives are provided in [Appendix D, supplementary material](#). The study outcome—recidivism—was defined as a felony offense that resulted in a conviction, occurring within 24 months of release from incarceration, or the initial assessment for those individuals supervised in the field. Outcome categories (i.e. violent, property, drug) were coded to be in line with prior TDOC offense-statute categorizations and were coded “1” for the presence of a recidivistic event, and coded “0” otherwise. To allow charges that occurred within the 24-month follow-up window time to result in a conviction, a six-month lag was used, creating a 30-month observation period. Using this follow-up period reduced the available data set to 17,689 individuals, of which 3,597 were female and 14,092 were male. Using newly collected data following the 2017 deployment, we computed eight ridge regressions to select new coefficient weights for the TDOC STRONG-R 2.0 models. K-fold validation procedures were again used to compute model performance metrics.

Analysis Plan – Model Comparison

The focus of the current study was to provide a model comparison. The goal of a “model comparison approach” is to contrast the explanatory power of two or more models, demonstrating the best approach (Judd et al., 2008). Specifically, we sought to compare the phased development of our STRONG-R Proxy 1.0 and TDOC 2.0¹⁰ to that of a more traditional, off-the-shelf approach. The original Washington State

⁸Multi-band scoring allows an assessor to complete a single set of items, where responses are differentially weighted and scored by software to specifically predict each of the eight recidivism models.

⁹The software flags response pairs that are inconsistently scored, such as “No threatening, aggressive, or violent behaviors in individual's lifetime” and “Prior felony assault offense.”

¹⁰We note that item coding was updated as part of TDOC'S STRONG-R 2.0 development, optimizing response categories to provide stronger recidivism prediction. A brief description of response changes is provided in Appendix D, supplementary material.

development sample was used to select and “weight” items using WADOC subjects. By contrast, many tools use an “unweighted” approach, selecting items from the development sample, allowing the raw values (rather than weighted responses) to be summed for composite risk scoring. Thus, to provide a more robust comparison, we developed Unweighted Development models of the STRONG-R, using items selected from the Weighted models, and computing scores from the raw response values. Next, we applied the Unweighted and Weighted Development models, created with the Washington State sample, and applied their scoring off-the-shelf (OTS) in the TDOC sample collected following the 2017 deployment. These OTS applications were used to replicate a more traditional RNA development processes, such as those used by the ORAS and LS tool providers. Thus, for the current study we provide a model comparison of six STRONG-R versions, including the 1) Unweighted Development, 2) Unweighted OTS, 3) Weighted Development, 4) Weighted OTS, 5) TN 1.0 Proxy, and 6) TDOC 2.0.

The methods here were used to contrast RNA development and application methods. The TN 1.0 Proxy and TDOC 2.0 models represent experimental, where OTS models were computed as control conditions, holding all other development procedures constant (i.e. sample, available predictors, outcome, and feature selection algorithm). The result of this model comparison is the isolation of the assessment development method, and in turn, identification of the best performing procedure for RNA design and application. The aims of this model comparison formed three study hypotheses:

H1: Predictive shrinkage is observed, comparing the Development to OTS models.

H2: Predictive improvement is observed, comparing the Proxy to OTS models.

H3: Predictive improvement is observed, comparing the TDOC 2.0 to OTS models.

We computed an average of AUC values across the four modeled outcomes (violent, property, drug, & felony), separated by gender, to compare model performance between the six STRONG-R versions. The AUC was selected as it represents the field-standard predictive performance statistic (Fawcett, 2004). We also made use of the previously referenced effect size ranges (see Rice & Harris, 2005)¹¹ and described model comparison criterion where AUC differences of 6% (or greater) represent “substantial” predictive shrinkage/improvement (see Hamilton et al., 2021).¹² Next, we compared the performance of the six STRONG-R versions applied in Washington and Tennessee.

Results

As outlined, we selected items to create the TDOC STRONG-R 2.0. Model coefficients values were used to weight responses and combined to create domain and composite risk scores. For a list of items and coefficient values see [Appendix E, supplementary](#)

¹¹As described, Rice and Harris (2005) provided effect size ranges that are commonly researcher to evaluate the strength of tool’s predictive accuracy, where 0.5 to 0.55 is ‘negligible’, 0.56 to 0.63 is ‘weak’, 0.64 to 0.70 is ‘moderate’, and 0.71 and above is considered a ‘strong’ level of predictive accuracy.

¹²We note that with large sample sizes, statistical significance is easily obtained and thus, we assess substantive rather than significant differences when comparing model versions.

Table 1. STRONG-R base rate.

Model Outcome	WADOC	Proxy	TDOC
Violent	11%	8%	7%
Property	9%	9%	5%
Drug	9%	9%	4%
Felony	25%	22%	16%

Note: WADOC: Washington Department of Corrections; TDOC: Tennessee Department of Corrections.

Table 2. STRONG-R version model performance by gender.

Model Outcome	Washington Sample			Tennessee Sample		
	Weighted Development	Unweighted Development	Unweighted OTS Application	Weighted OTS Application	TN 1.0 Proxy	TDOC 2.0
Male						
Violent	0.74	0.69	0.66	0.69	0.71	0.76
Property	0.78	0.74	0.67	0.71	0.72	0.77
Drug	0.76	0.72	0.64	0.69	0.71	0.74
Felony	0.74	0.71	0.64	0.66	0.68	0.73
<i>Average</i>	<i>0.76</i>	<i>0.72</i>	<i>0.65</i>	<i>0.69</i>	<i>0.71</i>	<i>0.75</i>
Female						
Violent	0.74	0.72	0.65	0.65	0.71	0.76
Property	0.74	0.72	0.65	0.67	0.71	0.75
Drug	0.73	0.72	0.56	0.58	0.64	0.72
Felony	0.72	0.70	0.63	0.64	0.66	0.73
<i>Average</i>	<i>0.73</i>	<i>0.72</i>	<i>0.62</i>	<i>0.64</i>	<i>0.68</i>	<i>0.74</i>

Note: TN: Tennessee; TDOC: Tennessee Department of Corrections; OTS: off-the-shelf.

material. We then compared the performance of each of the six STRONG-R versions created using three distinct samples—WADOC Development, WADOC Proxy, and TDOC 2.0 Development. Each of these three samples have recidivism base rate descriptions across the four outcomes measured provided in Table 1. For the STRONG-R 2.0, a range of 27 to 59 items were selected across eight models, where a varying number of dimensions were identified as predictors for each recidivism outcome and by gender. Regarding predictive performance, all models exceeded the “strong” level of accuracy ($AUC > 0.71$) (see Appendix E, supplementary material).

In Table 2, we provide models’ AUC performance for the four model types, computed across the six STRONG-R versions, by two genders, for a total of 48 AUC statistics. We also provide a summary (or average) AUC across the four outcome-specific models by gender. Readers should note that, compared to Felony models, the outcome-specific models demonstrate relatively consistent and modest AUC improvement (1% to 5%), with only the Unweighted Development and Unweighted OTS application demonstrating outcome-specific model inconsistencies. Predictive performance is slightly better for male compared to female models, with average AUC differences ranging from 1% to 5% across the six STRONG-R versions. These findings are consistent with prior comparisons of outcome-specific and gender-responsive models (Duwe, 2014; Hamilton et al., 2021).

Notably, computed only for the purposes of this study, the Unweighted Development model was created with the Washington sample to represent a contemporary, off-the-shelf version of the STRONG-R. When comparing performance, the Unweighted Development models were the worst performers, where the Weighted Development models performed consistently better in both the Washington and Tennessee samples.

Specifically, the Weighted model for males demonstrated a 4% average AUC improvement and the female models indicated a 1% improvement in Washington and a 4% increase in the Tennessee sample.

When comparing Development to OTS applications, a substantial drop in performance was observed when models developed in Washington were applied to the Tennessee sample. Comparing the Unweighted Development and OTS applications, a 7% and 10% performance shrinkage were observed for males and females, respectively. When comparing the Weighted Development and OTS application, the average AUCs were greater than Unweighted models, identifying a 7% and 9% performance shrinkage for males and females, respectively. However, compared to the Unweighted and Weighted Development models, the TN Proxy identified improved performance in Tennessee, where average AUCs for male models improved by 7% and 2% and females increased by 6% and 4%, respectively.

Finally, the TDOC 2.0 models demonstrated a consistent and substantial improvement over the other three model types. Regarding the Unweighted OTS applications, the TDOC 2.0 provided a 10% improvement for males and a 12% increase for females. We note that this improvement represents more than an effect size magnitude improvement of 1.5 and 2 for males and females, respectively. When compared to the Weighted OTS applications, the TDOC 2.0 improved predictive performance by 6% for males and 10% for females, again representing an effect size improvement for both genders. Lastly, a more modest improvement was observed when comparing the Proxy and TDOC 2.0, where the male model performance increased by 4% and the average female AUC improved by 6%.

To better illustrate the change in predictive performance, we provide the AUC model average as a trend line in [Figure 3](#), by gender. Our Weighted Development models are listed first, which we compare to the Unweighted Development models. The grey trendline indicates the modest drop between weighted and unweighted models in the development sample. However, both development models were built using local Washington data. As the dashed trend indicates, when the OTS models are applied in a new jurisdiction, a substantial drop is indicated. This drop in performance is predictive shrinkage and the reduction is similar to the OTS applications of the LS and ORAS tools described.

However, our novel approach in creating the TN 1.0 Proxy model provided a better approximation of a tool developed in Tennessee. As indicated, the improvement is modest when compared to the Weighted Development model (2% and 4%, respectively) and the performance increase is substantial, when compared to the Unweighted Development model (6%). Further, the TN 1.0 Proxy model provided a “hybrid” version of an OTS tool and a Tennessee version of STRONG-R to be used until the locally weighted TDOC 2.0 could be developed. Notably, the observed improvements of the TDOC 2.0 “returns” the STRONG-R performance to levels observed in the original Washington State development sample.

Discussion

Over the last 40 years, the RNR model has been established as an evidence-based practice (Andrews & Bonta, 2010). The model asserts that RNA tools form the

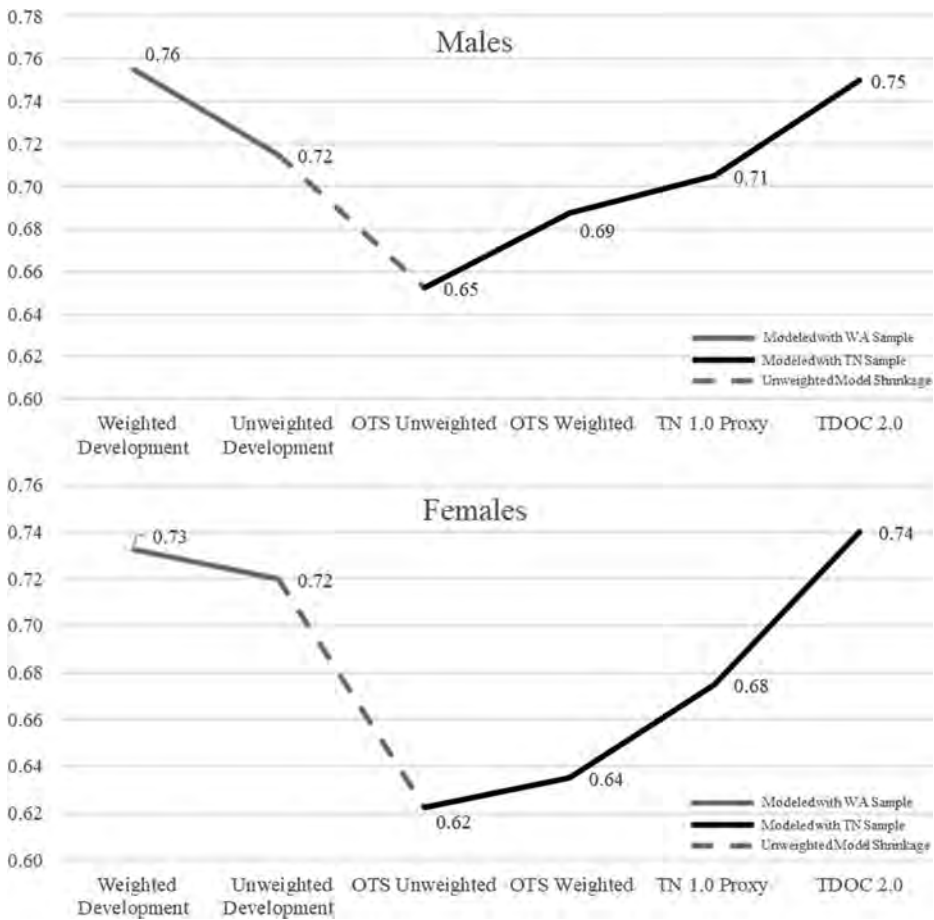


Figure 3. STRONG-R Development & TDOC application AUCs. Notes: TN: Tennessee; TDOC: Tennessee Department of Corrections; OTS: off-the-shelf; WA: Washington.

foundation of correctional practice, using scores to classify individuals to be supervised and provided programming and services according to their risk to reoffend. Understanding the foundational importance of risk assessment, agencies often seek to adopt tools identified to be “valid” in the prediction of recidivism risk. Yet, establishing a valid tool is a low bar, requiring little more than prediction beyond a coin flip, or random chance (Bucklen et al., 2010). More recent research has established the need to assess the magnitude of prediction, striving for better accuracy of assessments (Powers, 2011). Unfortunately, predictive shrinkage is a well-known concept among RNA researchers, where tool’s accuracy is strongest for the development sample in which an assessment was created and drops precipitously when adopted and deployed off-the-shelf by new jurisdictions (Fazel et al., 2022). Aside from diminished accuracy, poorly performing tools can cause a loss in stakeholder confidence, reducing their use in case management practices, and cause drift from the RNR model and effective supervision practices (Viglione, 2019). Unfortunately, LS/CMI researchers and evaluators have speculated that training and data collection as a primary rationale that Canadian tools have reduced performance in American applications (Olver et al.,

2014), ignoring the environmental, demographic, and jurisdictional variations universally shown to impact predictive accuracy (Hamilton et al., 2021).

Although recent research has demonstrated that optimization techniques improve the accuracy of homegrown tools, their adoption can be daunting (Duwe, 2021). While it is common for tools adopted off-the-shelf to require substantial time and investment, developing a homegrown tool requires considerably more effort. In particular, when creating a tool from scratch, developers must establish item pools and pilot test initial versions (Hamilton et al., 2017). These processes slow RNA deployment, and for agencies statutorily mandated to incorporate an RNA, tight adoption timelines can make homegrown tools seem infeasible.

Following the passage of the 2016 Public Safety Act, the TDOC was mandated to adopt a validated RNA to guide supervision and programming in the community. Seeking to avoid predictive shrinkage, they contracted with the STRONG-R developers in the creation of a hybrid approach to local tool development. Creating a staged process, a repository of STRONG-R assessment responses and reoffending data was statistically matched to create a Proxy sample of individuals that represented TDOC-like individuals. This proxy sample was then used to weight responses to be more reflective of Tennessee's population, creating version 1.0 of TDOC's assessment. Following deployment and data gathering, items were reweighted with a TDOC sample to create a homegrown version of a tool (2.0).

The current study describes the performance loss associated with traditional RNA development processes. We advanced the process by demonstrating performance gained *via* our staged RNA development approach, comparing six STRONG-R versions to test three study hypotheses. First, we assessed the AUC model change between the Development models created with the WADOC sample. We also created an Unweighted version of the Development model to provide more generalizable findings and a model that is roughly comparable to the ORAS and LS development processes. Study findings indicated that, when applied off-the-shelf in the TDOC sample, substantial predictive shrinkage was observed for both Weighted and Unweighted versions, providing support for *H1: Predictive shrinkage is observed, comparing the Development to OTS models*.

Next, we compared TDOC's STRONG-R Proxy models (Version 1.0) to the OTS versions, finding support for *H2: Predictive improvement is observed, comparing the Proxy to OTS models*. Notably, while the Proxy models indicate substantial improvement over the OTS Unweighted models, a more modest increase was observed for the Weighted OTS comparison. Therefore, we would expect that those that have, or are considering, implementing the LS, ORAS, or similarly unweighted tools, these agencies would greatly benefit by implementing a similar staged development process.

Finally, we compared the performance of the TDOC 2.0 to OTS versions, finding support for *H3: Predictive improvement is observed, comparing the TDOC 2.0 to OTS models*. Here, we find consistent and substantial predictive performance improvement when comparing the 2.0 models to the OTS versions. While others have posited the advantages that homegrown tools provide (Duwe, 2014; Hamilton et al., 2021), the current study findings provide the clearest evidence of this effect to date. Our findings indicate that the described homegrown RNA development yielded a 10-point improvement over the OTS tools, representing a potential 20% accuracy improvement on the

AUC scale. Therefore, creating a homegrown tool, like the TDOC STRONG-R 2.0, has the potential of making a “weak” predicting model “strong” and providing countless benefits for classification, supervision, programming, and service referrals.

Limitations

As mentioned, the rationale for our staged development approach was to reduce the well-known and notable impacts of prediction shrinkage. While the process of updating assessment tools based on the findings of new data is suggested as part of a multi-year routine for all agencies using RNAs (Bucklen et al., 2010), this best practice is rarely completed. Although the need for further refinements should be assessed every two-to-three-years, we anticipate additional refinements are likely to be minor and provide less dramatic impacts on predictive performance going forward. Thus, TDOC’s STRONG-R 2.0 currently provides the most accurate prediction of recidivism for individuals under supervision in Tennessee and given the custom and localized versioning, it is likely unrivaled by any contemporary tool currently available. With this said, applying the TDOC STRONG-R 2.0 in another agency/jurisdiction, without adjustment, would likely create substantial predictive shrinkage and is not advisable.

Related to issues of shrinkage across jurisdiction, populations within a jurisdiction are also subject to change. As many contemporary tools were constructed decades prior, item content requires updating, and response prevalence will vary over time. In 2021 Duwe updated the Minnesota Screening Tool. Assessing Recidivism Risk (MnSTARR), a homegrown tool used by the Minnesota Department of Corrections. Findings indicated that item weights required adjustment to retain the tool’s high level of predictive validity. Thus, performance shrinkage has several causes, and can be prevented with routine revalidation and adjustment.

However, an innovative development process, such as the one described here, is not easily accomplished, and requires investment of time and resources. Both the proxy model and STRONG-R 2.0 required substantial data to build and software to facilitate each model’s logic and scoring complexities. The performance and fidelity of the Tennessee STRONG-R is due, in no small part, to the efforts made by TDOC staff and their software and training contractor—Vant4ge. Specifically, when developing the tools, Vant4ge created a 35-person subject matter expert (SME) group of the TDOC agency, including administrators, line staff, wardens and clinical staff to ensure stakeholder involvement.

Further, both the 1.0 and 2.0 Versions were implemented with weighted scoring across eight models that were both outcome-specific and gender-responsive. Thus, the STRONG-R is a far cry from the original LSI tool, designed to be hand-scored in the field (Andrews & Bonta, 1995). The TDOC development process required an array of software creations and updates, changing the background logic of assessment scoring functionality for updated response weights. Finally, the finished product requires training and quality assurance (QA) procedures to insure reliable and accurate scoring. Further, accurate completion of the tool was assessed *via* a TDOC QA team, and assessment administrators were required to maintain accuracy to retain their position/employment with the agency. A flowchart of the implementation process is provided in [Figure 4](#).

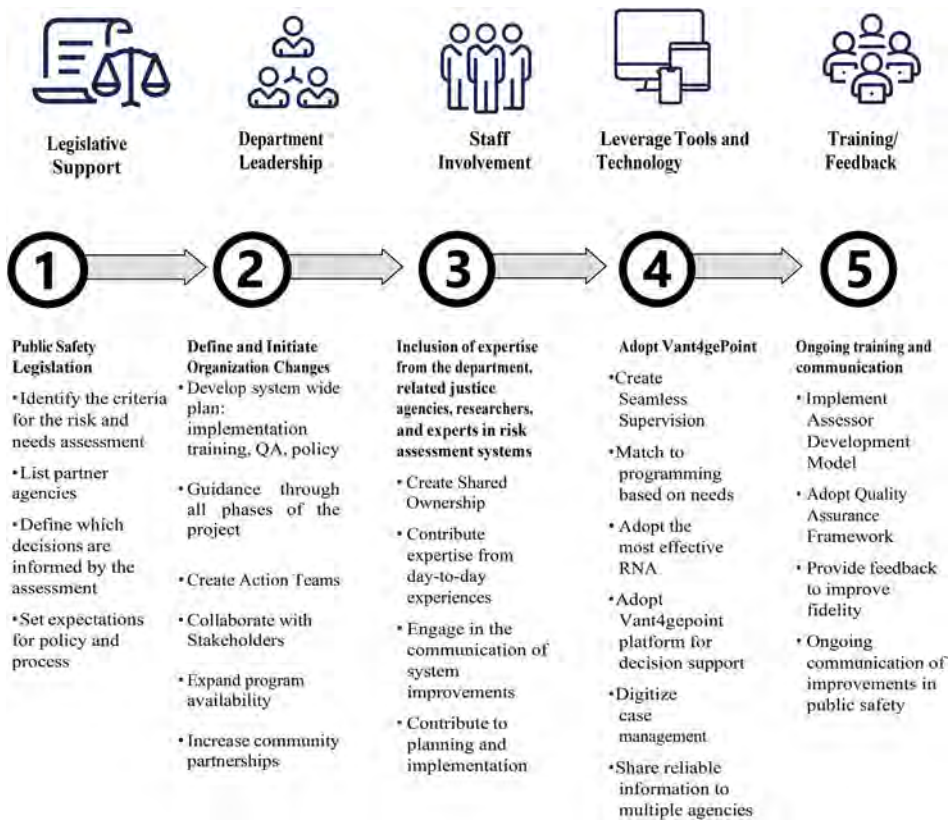


Figure 4. Decision support strategies. This flowchart describes TDOC strategies for successfully implementing Vant4gePoint™, which lead to seamless supervision from prison to community, more effective programming, and tailored solutions that meet agency needs. While in some instances each of these five steps may be applicable, other agencies may only require two or three strategies. For more detailed information on implementation, please visit www.vant4ge.com/implementation.

The entirety of the TDOC STRONG-R development process required a substantial commitment of time and resources to ensure proper implementation and accuracy of assessment data collected. This investment was successful and should not be overlooked. As many agencies have committed to justice reinvestment initiatives, and may be mandated to implement a validated assessment, agency resources may be more limited in future applications. Therefore, evaluating the scope of the RNA adjustment is a necessary first step for a project such as the one described.

Related to this point, the STRONG-R includes substantially more items than other contemporary RNAs. While the tool possesses a similar administration duration to that of contemporary tools (roughly 45 min), assessing an individual on 96 items may seem daunting for practitioners with large caseloads. Although it is an initial commitment to collect the entire pool of STRONG-R items, we have demonstrated that it is necessary to first collect more items than needed, so that an agency may customize and have the flexibility to reduce the content to fit their needs in subsequent versions. Therefore, it may not be feasible for tools that possess only a few dozen items (i.e.

the ORAS) to fully customize their assessment without first adding to the off-the-shelf content provided. With that said, several tool developers have reduced the time it takes to collect assessment information by automating data collection processes, reducing the time needed for semi-structured interviews (Duwe & Rocque, 2017). Thus, we anticipate that the STRONG-R and other assessment tools may be able to gather sufficient information with more efficient processes in the near future.

We further acknowledge that there are other alternatives to our proxy sample method. The development of synthetic data sets represents a complementary method. Using methods, such as multiple imputations, synthetic methods have been used for over 25 years (Bhati et al., 2013). The similarity of the approaches is that both aim to create a data set that is representative of the local population, prior to deploying the assessment that makes use of local data sources. It is unknown if the synthetic approach is superior to our proxy method, but with increased use of synthetic methods in justice research (Brunton-Smith et al., 2023; Dioli, 2022; Zilka et al., 2022), future analyses should seek to describe the optimal approach.

Finally, this study represents but one study, where predictive shrinkage and the methods of removing its effects were evaluated across two state samples. We anticipate that the magnitude of predictive shrinkage observed, and the strength of the Proxy and updated models created, will vary from jurisdiction to jurisdiction. Replication and further examination of updated tools' impact is needed, with additional evaluations of performance and bias by gender and racial/ethnic subgroups necessary. With applications of the STRONG-R in process for Nebraska, Pennsylvania, and South Carolina, we anticipate additional findings supporting the work described here in the years to come.

Conclusion

Outlined as an effective practice by Andrews and Bonta in their seminal piece the *Psychology of Criminal Conduct*, indicate that RNAs are to be routinely evaluated and adjusted to fit the population they assess (1995, 2010). However, when applied in a new jurisdiction, revalidations are not routine and when performed, often demonstrate performance loss (Fazel et al., 2022). Despite the vast repositories of data, likely consisting of millions of assessments to date, many tool providers, such as the MHS, UCCI, and Equivant, are resistant to adjust items, weights, or provide local variations of scoring algorithms shown to improve a tool's accuracy.

This apprehension may be due to time and resource restraints, branding considerations, or theoretical departures with our approach. However, resistance may also result from a lack of effective development and application methods. To solve this long-standing issue, we created our hybrid approach, which allows a developer to adjust any off-the-shelf RNA to approximate the local population. Further, the newest TDOC STRONG-R iteration (2.0), refined and calibrated the tool with TDOC subjects, where findings indicated strong recidivism prediction similar to the performance of the original tool developed and tested in the Washington State development sample. Planned in 2016 and methodically implemented by TDOC, Vant4ge, and the research team, these findings provide a culmination of five years of collaborative work.

While resource intensive, the TDOC investments produced tangible success. Specifically, a nearly 10% in evidence-based programing participation, a 10%

increase in parole releases, and a 22% increase in community corrections programming all contributed to an 15% reduction in recidivism in the years following the STRONG-R deployment (TNDOC, 2023). We feel the final product represents an organic development process, built with TDOC staff input, providing an evidence-based method of creating a homegrown RNA. As a result, our findings demonstrate methods of eliminating performance loss and the establishment of stakeholder ownership and buy-in that contribute to the STRONG-R's current and future success in Tennessee.

Notably, in a recent collaboration with the Counsel of State Governments, Desmarais et al. (2022), provide 13 guidelines for post-conviction risk and needs assessments, which outline the need to revalidate tools every five years, in consultation with university partners. They suggest that if properly incorporated into fiscal planning, even agencies with limited resources can ensure the long-term accuracy of their tool's provision. While these guidelines provide a strong foundation for evaluation, we would offer an extension of these efforts by suggesting that the data collected be leveraged to update the items, response weights, and cut points, to improve local accuracy. If completed with university partners and/or experts, these processes can be incorporated into routine revalidation efforts, iteratively improving the tool for the local population. Further these guidelines outline the need to communicate the importance of tool updates and document how newer versions will impact the day-to-day assessment processes and results (Desmarais et al., 2022).

The effectiveness of STRONG-R method encourages policymakers to provide support for the long-term development and implementation of RNA tools. Furthermore, the increased accuracy of RNA tools improves every aspect of correctional organization. Accurate risk prediction improves programming recommendations, which improves placement matching and programs access, increasing system flow. Thus, more accurate tools help agencies expedite rehabilitation practices, with the potential to reduce prison crowding and recidivism.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

- Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17(1), 19–52. <https://doi.org/10.1177/0093854890017001004>
- Andrews, D. A., & Bonta, J. (1995). *The level of service inventory—Revised*. Multi-Health Systems.
- Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Lexis Nexis/Anderson Pub.
- Barnoski, R. (2004). *Assessing risk for re-offense: Validating the Washington State Juvenile Court Assessment* (Report. No. 04-03-1201). Washington State Institute for Public Policy.
- Barnoski, R. P., & Drake, E. (2007). Washington's offender accountability act: Department of corrections' static risk instrument. Washington State Institute for Public Policy.

- Bhati, A., Crites, E. L., & Taxman, F. S. (2013). RNR simulation tool: A synthetic datasets and its uses for policy simulations. In *Simulation strategies to reduce recidivism: Risk Need Responsivity (RNR) modeling for the criminal justice system* (pp. 197–221). Springer New York.
- Brennan, T., & Dieterich, W. (2018). Correctional offender management profiles for alternative sanctions (COMPAS). *Handbook of recidivism risk/needs assessment tools*, 49–75.
- Brunton-Smith, I., Buil-Gil, D., Pina-Sánchez, J., Cernat, A., & Moretti, A. (2023). *Using synthetic crime data to understand patterns of police under-counting at the local level*. CrimRxiv.
- Bucklen, K. B., Duwe, G., & Taxman, F. S. (2010). *Guidelines for post-sentencing risk assessment*.
- Casey, P. M., Elek, J. K., Warren, R. K., Cheesman, F., Kleiman, M., & Ostrom, B. (2014). *Offender risk & needs assessment instruments: A primer for courts*. National Center for State Courts.
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (Vol. 1195). Houghton Mifflin.
- 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.
- Dioli, M. J. F. (2022). *Fair Imputation: Reducing bias under missing data* [Master's thesis]. University of Oslo
- Duwe, G. (2014). The development, validity, and reliability of the Minnesota screening tool assessing recidivism risk (MnSTARR). *Criminal Justice Policy Review*, 25(5), 579–613. <https://doi.org/10.1177/0887403413478821>
- Duwe, G., & Rocque, M. (2017). Effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy*, 16(1), 235–269. <https://doi.org/10.1111/1745-9133.12270>
- Duwe, G. (2021). Evaluating bias, shrinkage and the home-field advantage: Results from a revalidation of the MnSTARR 2.0. *Corrections*, 9(1), 20–42. 23. <https://doi.org/10.1080/23774657.2021.2011802>
- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31(1), 1–38.
- Fazel, S., Burghart, M., Fanshawe, T., Gil, S. D., Monahan, J., & Yu, R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *Journal of Criminal Justice*, 81, 101902. <https://doi.org/10.1016/j.jcrimjus.2022.101902>
- Guo, S., Fraser, M., & Chen, Q. (2020). Propensity score analysis: Recent debate and discussion. *Journal of the Society for Social Work and Research*, 11(3), 463–482. <https://doi.org/10.1086/711393>
- Hamilton, Z. K., & Campbell, C. M. (2013). A dark figure of corrections: Failure by way of participation. *Criminal Justice and Behavior*, 40(2), 180–202. <https://doi.org/10.1177/0093854812464219>
- Hamilton, Z., Kigerl, A., Campagna, M., Barnoski, R., Lee, S., Van Wormer, J., & Block, L. (2016). Designed to fit: The development and validation of the STRONG-R recidivism risk assessment. *Criminal Justice and Behavior*, 43(2), 230–263. <https://doi.org/10.1177/0093854815615633>
- Hamilton, Z., Campagna, M., Tollefsbol, E., van Wormer, J., & Barnoski, R. (2017). A model consistent application of the RNR model. *Criminal Justice and Behavior*, 44(2), 261–292. <https://doi.org/10.1177/0093854816678032>
- Hamilton, Z., Duwe, G., Kigerl, A., Gwinn, J., Langan, N., & Dollar, C. (2021). Tailoring to a mandate: The development and validation of the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN). *Justice Quarterly*, 39(6), 1129–1155. <https://doi.org/10.1080/07418825.2021.1906930>
- Hamilton, Z., Kigerl, A., & Kowalski, M. (2022). Prediction is local: The benefits of risk assessment optimization. *Justice Quarterly*, 39(4), 722–744.
- Hamilton, Z., Mei, X., & Routh, D. (2018). The static risk offender needs guide—Revised (STRONG-R). In *Handbook of recidivism risk/needs assessment tools* (pp. 199–228). John Wiley & Sons.
- Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. 608). Springer.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2008). *Data analysis: A model comparison approach* (2nd Edn.). Routledge.

- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, 14(2), 1137–1145.
- Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims & Offenders*, 5(3), 203–219. <https://doi.org/10.1080/15564886.2010.485900>
- Latessa, E., Lovins, B., Makarios, M. (2013). *Validation of the Indiana risk assessment system: Final report*. Unpublished technical report. <http://www.in.gov/judiciary/cadp/files/prob-risk-iras-final.Pdf>
- Lowenkamp, C., & Latessa, E. (2004). Understanding the risk principle: how and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections*, 2004 (1), 3–8.
- Lovins, B. K., Latessa, E. J., May, T., & Lux, J. (2018). Validating the Ohio risk assessment system community supervision tool with a diverse sample from Texas. *Corrections (2377-4657)*, 3(3), 186–202. <https://doi.org/10.1080/23774657.2017.1361798>
- Mara, C. M. (2002). Expansion of long-term care in the prison system: An aging inmate population poses policy and programmatic questions. *Journal of Aging & Social Policy*, 14(2), 43–61. https://doi.org/10.1300/J031v14n02_03
- Merianos, D. E., Marquart, J. W., Damphousse, K., & Hebert, J. L. (1997). From the outside in: Using public health data to make inferences about older inmates. *Crime & Delinquency*, 43(3), 298–313. <https://doi.org/10.1177/0011128797043003004>
- Norris, K., Overbey, J., Massey, S. (2016). SB2567. Tennessee General Assembly legislation. <https://wapp.capitol.tn.gov/apps/BillInfo/default.aspx?BillNumber=HB2576&ga=109>
- Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service Scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, 26(1), 156–176. <https://doi.org/10.1037/a0035080>
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*, 2(1), 37–63.
- 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. <https://doi.org/10.1007/s10979-005-6832-7>
- Rivlin, A., Fazel, S., Marzano, L., & Hawton, K. (2012). Studying survivors of near-lethal suicide attempts as a proxy for completed suicide in prisons. *Forensic Science International*, 220(1-3), 19–26. <https://doi.org/10.1016/j.forsciint.2012.01.022>
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38. <https://doi.org/10.1080/00031305.1985.10479383>
- Schwalbe, C. S., Fraser, M. W., Day, S. H., & Cooley, V. (2006). Classifying juvenile offenders according to risk of recidivism: Predictive validity, race/ethnicity, and gender. *Criminal Justice and Behavior*, 33(3), 305–324. <https://doi.org/10.1177/0093854806286451>
- St John, V., Murphy, K., & Liberman, A. (2020). *Recommendations for addressing racial bias in risk and needs assessment in the juvenile justice system*. Child Trends.
- Taxman, F. S. (2018). Risk assessment: Where do we go from here? In *Handbook of recidivism risk/needs assessment tools* (pp. 269–284). John Wiley & Sons.
- TNDOC (2023). *TDOC recidivism drops to lowest level in decade*. Tennessee State Government - TN.gov. <https://www.tn.gov/correction/news/2023/10/9/tdoc-recidivism-drops-to-lowest-level-in-decade.html>
- Viglione, J. (2019). The risk-need-responsivity model: How do probation officers implement the principles of effective intervention? *Criminal Justice and Behavior*, 46(5), 655–673. <https://doi.org/10.1177/0093854818807505>
- Wormith, S., & Bonta, J. (2018). The level of service (LS) instruments. In *Handbook of recidivism risk/needs assessment tools* (pp. 117–145). John Wiley & Sons.
- Zilka, M., Butcher, B., & Weller, A. (2022). A survey and datasheet repository of publicly available US criminal justice datasets. *Advances in Neural Information Processing Systems*, 35, 28008–28022.