Customizing Offender Assessment

Zachary K. Hamilton, Ph.D.

Jacqueline van Wormer, Ph.D.
Introduction

Since their seminal work “Psychology of Criminal Conduct”, offender assessment has become a staple of the criminal justice system (Andrews & Bonta, 1994). From arrest to community supervision, it is nearly inconceivable for an individual to have contact with the criminal justice system and not be assessed for any number of attributes. These attributes may include risks, needs, protective factors, skills, deficits, or responsivity an on a variety of outcomes. Whatever the intent, a risk assessment may be built to suit the needs of an agency. Most agencies are either unaware that instruments can be customized to meet their needs and may have difficulties expressing their needs with regard to potential modifications available. To address this challenge, this paper provides a helpful guide to explore the variety of assessment types and considerations needed when customizing a tool to meet an agency’s needs.

A Systems Approach to Assessment

Given the expansion and use of risk tools in the criminal justice system, commonalities have been uncovered to allow for a system’s approach to assessment. Many initial instruments were built for adult probation populations. These instruments were created to guide supervision strategies. As early utility was heralded a success, use soon expanded to parole, juvenile justice, prison and pretrial jurisdictions. Despite the initial intent, which was to predict recidivism for a probation population, it was quickly discovered that when “taken off the shelf” and applied to an adjacent criminal justice population, these instruments still provided a satisfactory prediction of criminogenic outcomes.

Although one may be weary to apply a risk assessment to a new population, it is easy to understand why many of the risk factors used to examine one population would be similar to that of another and why outcomes of pretrial risk are correlated with outcomes important to community supervision. However, these predictions can be further added by adopting additional “specialty items” used to predict important outcomes of interest at a given point in the system, such as prior infractions, technical violations, failure to appear warrants, negative UAs, and treatment compliance. This understanding of a general “Assessment Item Pool” that is adaptable to variant populations has inspired assessment developers (i.e. the COMPAS, ORAS, and STRONG) to develop Assessment Systems to be used to assist agency staff and track offender’s progress through the criminal justice stages.
Issues of concern

Because risk assessments may guide and inform criminal justice staff at each stage of the system, choosing an assessment is not a simple task. There are several issues to consider that will impact the performance and ultimately the buy-in that staff/stakeholders have following implementation. Included here is a list of several noted items that policy makers should consider before adopting an assessment tool.

1. Risks and Needs Assessments

Seen as a generational advancement in offender assessment, that addition of dynamic needs items was viewed nearly universally to be an improvement in both quality and predictive ability. Defined as the first “R” of the Risk, Need, and Responsivity principles (RNR), risk assessments make use of any and all items that logically predict the offender outcome of interest, typically recidivism. These can be static items, such as “total number of prior convictions”, or dynamic items, such as “use of heroin in the past six months”. There are a variety of methods of combining offender responses to compute a final risk score. Ultimately, the score is totaled and then cut points divide potential scores into risk category bins (i.e. low, moderate, or high). These category bins may then be used to guide supervision practices, such as contact standards in the community.

Although the same/similar items may be used, needs assessments serve a different purpose – to prioritize individuals for interventions and services. One of the guiding principles of RNR is that greater reductions in the probability of recidivism and other negative outcomes will be achieved when moderate and high risk/need offenders are prioritized for interventions and services. While items used to assess risk may include a variety of measures, those used to assess need are notably designed to be temporary and changeable. In this way, needs are typically grouped into domains, in which those identified to have a high need in a given domain (i.e. Employment, Criminal Attitudes, Substance Abuse), will be given priority for intervention an service resources over those scoring lower for said domain.
A primary distinction between the two assessment types is that risk assessments tend to focus on static and often primarily criminal history items. These assessments may be automated, while a needs assessment requires and interview, costing assessment labor. However, simply using dynamic needs items does not necessarily equate to better prediction of recidivism. Depending on the intended purpose of the agency, needs assessments may be of less importance or potentially not feasible. One may wish to only perform a static risk assessment due to cost, feasibility, and increased objectivity.

If prediction of recidivism is the only goal, a modification such as this may come with little loss to prediction strength. In fact, in Washington State Barnoski and Drake (2007) actually found a more specified static assessment could increase instrument predictive strength over and beyond more notable risk-need assessments. Alternatively, some prefer to use a static assessment to screen out, or divert, lower risk offenders; thus, reserving the labor demands of a need assessment interview for only higher risk offenders. While still others prefer a more holistic model that focuses on assessing needs of the entire population, monitoring micro-level changes within offenders over time. This is an important distinction as some instruments require an offender (need assessment) interview in order to be scored, while others offer more flexible options such as those described.

2. Weighting, Norming, and Tailoring

Instruments are developed using a single population. Items are selected and weighted based on the importance of each item in the prediction of the recidivism and/or other criminogenic outcomes. Classic methods of instrument construction will provide equal weight to each item. For example, being male, having a prior felony conviction, and a diagnosed mental illness would each receive a point (or equal weight). This style of scoring schematic is referred to as Burgess weighting. Alternatively, statistical models may be built to determine the importance of each item, providing more points for items with greater prediction strength (i.e. age) and fewer points for items with lesser importance in the prediction model (i.e. leisure activities). This is referred to as analytic weighting. While Burgess weighted instruments provide simplicity and are less likely to require scores to be calculated via software, analytically weighted models provide greater face validity and are notably more precise.

Whatever the weighting structure desired, following implementation data must be gathered to determine how accurate the instrument is at predicting risk. Often it is found that cut point used to determine risk categories (i.e. low, moderate, and high risk) must be adjusted for the jurisdiction. A process referred to as “norming”. This process is important for every implementation effort, as populations can (and do) change overtime. However, this process is especially important for instruments “taken off the shelf” or initially developed for another jurisdiction or population. All agencies should be aware of the importance of norming to increase assessment accuracy and create a timeline for routine revalidation and norming of implemented assessments.

Tailoring an instrument to a jurisdiction is rarely discussed. This process can be viewed as an assessment of the jurisdiction’s responsivity to the tool. Tailoring can be viewed as a process by which one adjusts an assessment for the localized population. Some instrument developers view their product as set, or static, in the items and the response categories provided. These instruments are often created to be generalized for all jurisdictions. A new trend in assessment involves the process of identifying new items and/or response categories of the instrument that are
better tailored to the needs of the jurisdiction. For example, the Positive Achievement Change Tool (PACT) for youth was developed in Washington State. It was later implemented in its original form in Florida. Over time, modifications were made to the items, responses, and scoring, creating what was called a “Florida-ized” version of the instrument. The process of tailoring can be seen as a complex process and may result in several versions of the same instrument, depending on the location in which it is implemented. However, consideration should be given to the uniqueness of the population in which the assessment will be applied, as tailoring, if completed with the assistance of experts, will likely improve instrument prediction strength and stakeholder buy in.

3. Outcome Type and Specificity

One of the biggest practical problems in validating risk assessments is having reliable outcome data that can readily be linked to the assessment. The most comprehensive metrics of recidivism are typically available at the state-level. However, not all jurisdictions have access to an automated comprehensive statewide criminal record database.

Assuming applicable data sources, the two most common criminogenic outcomes predicted are rearrests and reconvictions. Although substantially correlated, the outcome used for prediction is important and will vary based on the agency. An agency early in the systems timeline (i.e. juvenile justice or drug court) might focus efforts on preventing rearrests, while a department of corrections (DOC) may focus on reconvictions and reincarcerations. Furthermore, not all agencies will focus on the same crime type, as probation departments may not want to discriminate with regard to crime severity, while a DOC will seek to focus recidivism prevention efforts on felony convictions. Regardless of a developer’s preference in outcome type, one should be aware that assessments originally developed using one outcome may perform quite differently when utilized to predict another.

Another concern is the specificity of the outcome. Several instruments have attempted to improve prediction specificity by modeling separate prediction models for Violent, Property and Drug offenders. This detailed understanding of an offender’s primary offense type aids in the prioritization of contact standards based on an agency’s priority. Furthermore, several instruments have attempted to increase prediction specificity by providing assessments of prison infraction to assist with custody designation and measures of non-compliance to inform case management in the community. As an agency seeks to adopt an instrument, consideration should be given to the type and specificity of the prediction desired.

4. Narrow vs. Broadband Assessments

Agencies commonly provide a multitude of assessments, which are often used for a variety of purposes during case management. As mentioned, risk assessments make use of any and all measures that are theoretically and statistically related to recidivism and other criminogenic outcomes. Many of the nationally renowned offender risk assessments used today are conceptualized as broadband instruments. Most of these instruments will specify subgroups of items (i.e. domains) that will indicate an individual’s needs. Focused on identifying how a given need domain predicts future criminality, these items subsets may not provide the detailed assessment needed to specify treatment in a given area.

With regard to substance abuse and mental health issues, domains within broadband assessments are usually inadequate. Agencies often use psychiatric and chemical dependency professionals to provide narrowband assessments. Through a combination of structured judgment and specified instruments used to get at symptom severity, individuals are assessed for medication needs and/or matched to an appropriate treatment regimen. It is important for agencies to understand that the implementation of a broadband risk-needs assessment does not remove the necessity for the narrowband assessments and with proper policy integration both assessment types can work together to achieve common goals.
5. Special Populations and Overrides

Related to the issue of narrowband assessments, special populations can obscure the mechanics and delivery of an assessment. Developed on the aggregate, assessments are used to create a prediction that minimizes inaccuracy, averaged across the entire population in which it was developed. Example cases that exist in the extremes of the prediction landscape will not be appropriate for broadband assessments and procedures should be included to provide for overrides. Offenders with severe mental illnesses, physically disabled, and offenders released after long-term sentences (e.g. low risk yet high need) are just a few examples of cases that will require override procedures. Assessment manuals and proper training should be provided to ensure that case managers are not confined to results that do not match the logic and experiential knowledge of skilled criminal justice staff. Providing clear exceptions for assessment results will improve stakeholder buy in and ultimately increase the utility of the assessment.

6. Gender Responsivity

In the last decade pioneering research has been completed, demonstrating the variant pathways of crime female offenders take and how those pathways differ from male offenders (Van Voorhis et al., 2010). In recent years the concept of gender responsivity in offender assessment has accelerated understanding of the variations in prediction items and interventions needing specified female components.

As all criminal justice populations are vastly male, instruments developed using an agency’s general population will notably select and weight instrument items that trend toward male dominated models. Many of the nationally recognized instruments were created in this way and are referred to as gender neutral. Perceived to be less responsive for female offenders, methods of instrument development have been outlined to create new, or alter previously developed gender neutral models.

There are four primary methods for accounting for gender in an assessment. First, one may use gender as a predictor, which has the effect of combining the potential variations between males and females into a single measure. While concise, this method will limit contextual information that may be provided to case managers to assess intervention and service provision. Second, an instrument may manually adjust cut points that determine what range of risk/need scores are categorized as high risk. Typically, these adjustments will serve to reduce the amount of females in high risk categories. Similarly, this method will also lack contextual information. Third, an instrument can be developed to be gender specific, selecting and weighting items separately for each gender. This method provides the contextual information lacking in previous method. However, instruments built with this method often make use of items initially designed for gender neutral assessments and are simply weighted separately for the two distinct subpopulations. A final method used to increase gender responsivity is to assemble and assess the impact of scales thought to be more predictive of female and male risks and needs. For example, the Women’s Risk Needs Assessment (WRNA) utilizes scales such as parental stress and trauma as factors known to be of higher prevalence in female offender populations.

<table>
<thead>
<tr>
<th>Table 1. Gender Responsive Assessment Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender responsive method</strong></td>
</tr>
<tr>
<td>1. Use of gender as a predictor</td>
</tr>
<tr>
<td>2. Adjust cut points</td>
</tr>
<tr>
<td>3. Subsample item selection &amp; weighting</td>
</tr>
<tr>
<td>4. Gender responsive scales</td>
</tr>
</tbody>
</table>
7. Most Validated vs. Strength of Validation

Validation is a term that denotes acceptance and confidence. However, it is important for criminal justice agencies to understand that “validated instrument” can be presented in different forms and can be a somewhat misinformed term. The process of instrument development begins with the selecting and then weighting of items to be scored for each offender. The sample in which the instrument is created on is referred to as the construction sample. To assess the likely future stability of the assessment, the items are then scored on a new group of individuals, called the validation sample. Statistical tests results reported, often use the validation sample to determine how well the instrument predicts the outcome of interest (i.e. recidivism).

To measure prediction strength a value is computed using the Area Under the Curve (AUC) statistic. The value of the statistic ranges from 0.5 to 1.0, where .5 reflects being right/incorrect in one’s prediction 50% of the time. This is the equivalent of flipping a coin to predict if an individual will recidivate, or what some refer to as “betting the base rate”. An instrument is reported to be “validated” if it is identified to have an AUC somewhat greater than 0.5. However, with only a few good items (i.e. age, gender, and prior number of convictions) and a sufficient sample, the mark of “validated” is a criterion that is rather easily achieved.

In the plots below one can see three examples using data from the male development sample of the STRONG-R risk assessment instrument. Plot 1 predicts reconviction using just “age at assessment”. This measure alone provides an AUC that is a small effect size (AUC = 0.58) and due in part to the large sample size (N=39,155) the AUC is determined to be significantly greater than 0.5 (p<.001). The second plot adds only the measure of “number of prior felony convictions” to the prediction model and provides a prediction model of moderate strength (AUC = 0.67). This value is roughly the strength of prediction similar to that of many nationally utilized instruments. Finally, Plot 3 uses the full risk score generated from all the items included in the STRONG-R felony male model and identifies a strong prediction strength (AUC=0.74). Although all three of these prediction models would be identified as “validated”, one can see that prediction strength varies greatly.
Instruments that have been marketed over the years have had ample opportunities to be validated. However, while the instrument that is the “most validated” indicates historical stability, this distinction does not describe the instrument’s strength in predicting recidivism and other criminogenic events. With regard to strength, researchers often examine the size of the AUC value, as those boasting larger values will provide improve an instrument’s performance in identifying of who is and who is not likely to recidivate. Based on prior findings (Rice & Harris, 2005), and depicted in Table 2 four categories of strength have been established; where an AUC value below 0.55 are considered negligible, values of 0.56 or greater are small, values greater than 0.64 are considered moderate, and those greater than 0.71 are strong prediction instruments.

<table>
<thead>
<tr>
<th>Area Under the Curve (AUC)</th>
<th>Strength of Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 0.55</td>
<td>Negligible</td>
</tr>
<tr>
<td>0.56 to 0.63</td>
<td>Small</td>
</tr>
<tr>
<td>0.64 to 0.70</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.71 and above</td>
<td>Strong</td>
</tr>
</tbody>
</table>

It is interesting to note that many instruments developed (or using similar methodologies as those developed) more than a decade ago often promote the sufficiency of being “validated”, while more recently developed instruments and/or those developed with more advanced methodologies will claim greater “prediction strength”. As agencies decide on the adoption of assessment tools most appropriate for their jurisdiction, strength versus frequency of validation should be a key consideration.

8. Software Application

Initially, offender assessments were designed to be completed with pen and paper; however, with increased use of agency software to track offender progress, the use of software applications of criminal justice assessments has become imperative. Although variants exist with regard to cost and feasibility, many new and impressive applications have been developed to improve staff communication and efficiency. With regard to communication, tracking an individual’s risks and needs over time is important. As an offender transitions from arrest, to pretrial, prison, and to the community, their case will be managed by a variety of staff and likely more than one agency. Utilizing software applications one may simply make minor modifications to the individual’s assessment at each stage. This will allow staff to get a sense of the individual’s progress and streamline the assessment process, updating only those items that register change or require alteration.

9. Case Management Planning

“Responsivity” or effective case management is often the forgotten “R” of the RNR system, as is most difficult to properly implement. Historically this component has been underdeveloped and considered an afterthought. Responsivity is the process of matching offenders to services and interventions that will have the greatest impact. When the tool provides an individual’s risk and needs categories the case manager must then identify the interventions and services most appropriate. If a person has a serious mental illness or language difficulties, these characteristics can impede their progress in a given program, a concept known as “general responsivity”. Identifying that an individual might need a certain sequence of services or that a person might not be motivated or a good match for a given program is a part of case management known as “specific responsivity”. At a certain point the category of risk and the areas of needs will only get an agency so far. Linking this information to a coordinated case plan takes system integration, intervention and service availability, experienced and well-trained staff, and quality assurance practices. Customization at this level takes an agency from well-informed to “evidenced-based”.
Conclusion

As agencies continue to search for the best metric to guide their decision-making, greater reliance on risk and needs assessments have been observed. When an agency looks to change contact standards in the community, custody levels in prison, bail applications at pretrial, or the provision of interventions and services to address needs, the information gathered from assessments should help inform the decision-making process. Apart from individual/offender level data, assessment data can be used to identify the prevalence of supervision and treatment needs, strengthening rationales for reorganization of resources and areas of funding need.

This increase in reliance has taken the field from what was once a simple handful of criminal history and demographic items to a more detailed understanding of an individual’s risks and needs across a variety of recidivism and criminogenic outcomes. The current trend is to provide a systems approach to assessment. Using a general pool of items, assessment tools can track individuals from criminal justice system entry to exit. With modeling variations these tools can be customized to more accurately predict outcomes of interests that are in line with the characteristics of the population, jurisdiction and agency goals. Although more likely exist, this paper provides an overview of many of the primary assessment variations and describes the issues agencies should consider when selecting and implementing an assessment tool.

★ What is the AUC?

The Area Under Curve (AUC) is a widely used measure of predictive utility between recidivism and risk classification. AUCs range can vary between .5 and 1.0 and the larger AUCs represent higher accuracy. According to Rice and Harris (2005), the AUC values of small, medium, and large effects, are .556, .639, .714, respectively.
