Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States

Sarah L. Desmarais, Ph.D.

Department of Psychology, North Carolina State University

Jay P. Singh, Ph.D.

Department of Justice, Psychiatric/Psychological Service, Canton of Zürich, Switzerland

March 27, 2013

ACKNOWLEDGMENTS

We gratefully acknowledge the research assistance and contributions of Kiersten Johnson, Krystina Dillard and Rhonda Morelock to this report, as well as Grace Seamon for her research assistance. We also thank Mr. David D'Amora and Dr. Fred Osher for their guidance in the preparation of this report.

This project was funded by the Council of State Governments Justice Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor.

TABLE OF CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

Overview

The rates of crime, incarceration and correctional supervision are disproportionately high in the U.S. and translate into exorbitant costs to individuals, the public and the state. Though many offenders recidivate, a considerable proportion do not. Thus, there is a need to identify those offenders at greater risk of recidivism and to allocate resources and target risk management and rehabilitation efforts accordingly. Doing so necessitates accurate and reliable assessments of recidivism risk. There is overwhelming evidence to suggest that assessments of risk completed using structured approaches produce estimates that are both more accurate and more consistent across assessors compared to subjective or unstructured approaches. More and more, structured risk assessment approaches are being used in correctional agencies.

In this review, we summarize the research conducted in the United States examining the performance of instruments designed to assess risk of recidivism, including committing a new crime and violating of conditions of supervision, among adult offenders. We focus specifically on performance of tools validated and currently used in correctional settings in the United States.

Methodology

We identified instruments designed to assess risk of recidivism by searching academic research databases and Google. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues. Criteria for instruments to be included in the review were: a) designed to assess the likelihood of general recidivism (i.e., new offenses and violation of conditions); b) intended for assessing adult offenders (18 years of age and older); c) used in correctional settings in the United States; and d) validated in the United States. Instruments were excluded from our review if they: a) were designed to assess the likelihood of adverse outcomes for specific offenses (e.g., sexual offenses, violent offenses, spousal assault); b) were intended for assessing juvenile offenders (less than 18 years of age); c) were not used in correctional settings in the United States; d) had not been validated in the United States.; or e) were developed for use in a specific institution or ward.

We then identified studies examining the validity of these instruments using the same databases, search engine and secondary sources as above, using both the acronyms and full names of the instruments as search criteria. We searched for studies published between 1970 and 2012 in peer-reviewed journals, as well as government reports, doctoral dissertations, and Master's theses. Using this search strategy, an initial total of 173 records was filtered to a final count of 53 studies, representing 72 unique samples.

Information about the characteristics of the instruments, assessment process, and studies was collected. We also recorded information on inter-rater reliability and predictive validity, overall and by offender sex, race/ethnicity, study context, and recidivism outcome, where possible.

Findings

There were very few U.S. evaluations examining the predictive validity of assessments completed using instruments commonly used in U.S. correctional agencies. In most cases, validity had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument. Also, only two of the 53 studies reported evaluations of inter-rater reliability. There was no one instrument that emerged as systematically producing more accurate assessments than the others. Performance within and between instruments varied depending on the assessment sample, circumstances, and outcome.

Some instruments performed better in predicting particular recidivism outcomes than others. Other instruments were developed to assess for specific populations (e.g., parolees) or appeared to perform better for some subgroups of offenders than others (e.g., male versus female offenders). Finally, the information and amount of time required to complete assessments varied considerably. Some instruments could be completed based solely on offender self-report; other instruments used information derived from a variety of sources, including self-report, interview, and review of official records. Still other instruments could be completed based on file review alone. The number of items included the instruments also varied considerably: from four to 130.

Conclusion

When deciding which recidivism risk assessment instrument to implement in practice, we recommend first narrowing the potential risk assessment instruments by answering the following questions: *What is your outcome of interest? What is your population? What resources are required to complete the assessment?* We then recommend careful consideration of the research evidence, including the amount and strength of the empirical support for inter-rater reliability and predictive validity, generalizability of findings, and possible sources of bias that may have impacted results. Finally, it is important to remember that the goal of risk assessment is not simply predict the likelihood of recidivism, but, ultimately, to reduce the risk of recidivism. To do so, the risk assessment tool must be implemented in a sustainable fashion with fidelity; findings of the risk assessment must be communicated accurately and completely; and, finally, information derived during the risk assessment process must be used to guide risk management and rehabilitation efforts.

BACKGROUND

Prevalence of General Offending and Recidivism in the U.S.

The crime rate in the U.S. is high, estimated at 3,295 crimes per 100,000 residents in 2011 (FBI, 2012). With 743 in 100,000 U.S. adults incarcerated at the end of 2009 (Glaze, 2011), the rate of incarceration is over four times the rate found in more than that of half the world's countries (Walmsley, 2010). Indeed, though the U.S. has less than 5% of the global population, it has more than 25% of the world's prisoners (Liptak, 2008). Further, approximately one out of every 30 adults is under some form of correctional supervision (Pew Center on the States, 2009).

These high rates of crime, incarceration and correctional supervision translate into exorbitant costs. Approximately $74 billion was spent on corrections in 2007 (Kyckelhahn, 2012). When both direct and indirect costs are considered, estimates of annual costs have reached as high as $1.7 trillion (Anderson, 1999). Though almost two-thirds of offenders recidivate following release, another third do not go on to reoffend (Langan & Levin, 2002). Criminal justice expenditures, however, typically are distributed equally among offenders, regardless of risk level. It would be more cost-effective to allocate funding based on consideration of other factors, such as risk of recidivism and treatment needs. Indeed, correctional programs that adhere to the Risk-Need-Responsivity (RNR) model for offender assessment and rehabilitation have increased efficacy in reducing recidivism (e.g., Lowenkamp, Pealer, Smith & Latessa, 2006).

The RNR model represents an idiographic approach to risk management and rehabilitation. First, the *risk* principle dictates that treatment and intervention should be proportionate to each offender's recidivism *risk*, with more restrictive and intensive efforts used for high-risk offenders. The *need* principle calls for consideration of individual criminogenic needs to tailor treatment to each offender. Finally, the *responsivity* principle requires adapting treatment according to the individual offenders' learning styles, motivation, personalities and strengths, and use of approaches that are known to be responsive to the identified needs (Bonta & Andrews, 2007). Adherence to the principles of the RNR model necessitates accurate and reliable assessments of recidivism risk.

ISSUES IN RISK ASSESSMENT

Risk Assessment in Correctional Settings in the U.S.

Risk assessment can be defined as the process of estimating the likelihood of future offending to identify those at higher risk and in greater need of intervention. Conducting risk assessments also may assist in the identification of treatment targets and the development of risk management and treatment plans. There is overwhelming evidence to suggest that assessments of risk completed using structured approaches produce estimates that are both more accurate and more consistent across assessors compared to subjective or unstructured approaches (Ægisdóttir et al., 2006). Importantly, the use of structured approaches to classify higher risk individuals within the general offender population also produce better outcomes compared to unstructured approaches (Mamalian, 2011). More and more, correctional agencies are recommending—and many now require—the use of structured risk assessment approaches (Skeem & Monahan, 2011).

Evolution of Risk Assessment

The focus and structure of risk assessment tools have shifted significantly over time. The general characteristics of four distinct generations are summarized below.

First Generation

The first generation of risk assessment is best described as unstructured professional judgment, in which the assessor relies on their professional training and information gathered from the offender, official records or other sources to inform their evaluation of risk for recidivism. It is "unstructured" insofar as there is no set checklist or protocol for completing the risk assessment, though assessors may indeed complete structured interviews during the risk assessment process. This method of assessment was widely accepted for decades prior to the development of structured risk assessment tools in the 1970s. Today, it is less frequently used, but nonetheless remains a prominent risk assessment strategy, despite evidence that accuracy of unstructured assessments risk are less accurate than chance.

Second Generation

Following decades of research focused on identifying factors that increase risk of recidivism, second generation tools represent a drastic advance in risk assessment technology. Second tools are actuarial in nature and comprised primarily of historical and static factors (e.g., sex, age and criminal history). Rather than subjective judgments of recidivism risk, instruments such as the Salient Factor Score (SFS) and Violent Risk Appraisal Guide (VRAG) instead guide assessors to consider a set list of risk factors to arrive at a numerical risk of recidivism. Actuarial instruments are described more fully in the following section.

Third Generation

The third generation of risk assessment is characterized by the development of tools that include dynamic factors and criminogenic needs, and may use an actuarial or structured professional judgment approach. Third generation tools, such as the Level of Service Inventory-Revised (LSI-R), the Self-Appraisal Questionnaire (SAQ), and the Historical-Clinical-Risk Management-20 (HCR-20), still guide assessors to consider static factors; however, by including potentially dynamic items, such as attitude and substance use, they may be sensitive to change in risk levels over time and can assist in identification of treatment targets. These tools are sometimes referred to as "risk-need" instruments and, unlike second generation assessments, tend to be theoretically- and empirically-based as opposed to wholly data driven.

Fourth Generation

Most recently, fourth generation risk assessments explicitly integrate case planning and risk management into the assessment process. As such, the primary goal of the fourth generation extends beyond assessing risk and focuses on enhancing treatment and supervision. Examples of fourth generation tools include the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Ohio Risk Assessment System (ORAS), and Wisconsin Risk and Needs tool (WRN). Like the third generation, this generation of risk assessment instruments allows for the role of professional judgment while remaining grounded in research and theory.

Structured Approaches to Conducting Risk Assessments

There are two broad categories that distinguish between the structured approaches used to conduct risk assessment in the second, third and fourth generations: actuarial and structured professional judgment. We briefly review the strengths and limitations of each below.

Actuarial Risk Assessment

The actuarial approach represents a mechanical model of risk assessment, largely focused on historical or unchanging risk factors. When an actuarial instrument is used to assess risk, an offender is scored on a series of items that were most strongly associated with recidivism in the development sample. The offender's total score is cross-referenced with an actuarial table that translates the score into an estimate of risk over a specified timeframe (e.g., 10 years). This estimate represents the percentage of participants in the instrument's development study who received that score and recidivated. For example, if an offender receives a score of +5 on an instrument which is translated into a risk estimate of 60% over 10 years, this means that 60% of those individuals who received a score of +5 in the instrument's original study went on to recidivate within that time. This does not mean that the offender has a 60% chance of recidivating over a period of 10 years. This is an important distinction that is frequently overlooked in practice.

Strengths of the actuarial approach include:

- *Objectivity*. No human judgment is involved in estimating risk once items have been rated. Items are typically straightforward and easy to rate (e.g., age, sex, number of prior offenses).
- *Accuracy.* Actuarial assessments are more accurate than unstructured assessments.
- *Transparency*. Information used to inform risk estimates is explicitly included in the instrument. Items are weighted in a pre-determined manner to compute total scores and estimate risk.
- *Speed*. Items included in actuarial instruments can usually be scored using information available in official records.

Drawbacks include the application of group-based statistics and norms to individual offenders. Beyond potential statistical issues (see Hart, Michie & Cooke, 2007), this is a concern because we do not know where any given offender falls within a risk bin. Using the same example provided earlier, if 60% of the individuals who received a score of +5 recidivated over a 10-year period, then 40% did not. Actuarial assessments cannot help distinguish whether an offender receiving a score of +5 is among the 60% or 40%. Additionally, with invariant item content comes the potential exclusion of case specific factors that do not systematically increase (or decrease) recidivism risk across the population but are relevant to a particular offender's level of risk. Finally, actuarial assessments speak to level of risk and may inform decisions regarding risk classification and allocation of resources. However, their utility in guiding the development and implementation of individualized risk reduction and rehabilitation plans is limited due to their focus largely on historical or unchangeable factors that cannot be addressed in treatment.

Structured Professional Judgment

In contrast to the mechanistic, actuarial approach, the structured professional judgment approach focuses on creating individualized and coherent risk formulations and comprehensive risk management plans. These instruments act as *aide-mémoires,* guiding assessors to estimate risk level (e.g., low, moderate or high) through consideration of a set number of factors that are empirically and theoretically associated with the outcome of interest. Although offenders are scored on individual items, total scores are not used to make the final judgments of risk. Instead, assessors consider the relevance of each item to the individual offender, as well as whether there are any case specific factors not explicitly included in the list.

Strengths of the structured professional judgment approach include:

- *Professional discretion.* Assessors consider the relevance of factors to the individual offender to inform final estimates of each. Case specific factors also can be taken into consideration.
- *Accuracy.* Structured professional judgment assessments are more accurate than unstructured assessments (and comparable in accuracy to actuarial assessments).

- *Transparency*. Assessors rate a known list of factors according to specific guidelines. Additional items considered are added to the assessment form.
- *Risk communication and reduction*. Risk formulations provide information regarding the anticipated series of stressors and events that lead to the adverse outcome and over what period time, which can inform risk management strategies and identify treatment targets.

Drawbacks include the potential re-introduction of decision-making biases in the final risk judgments. Structured professional judgment instruments also take comparatively longer to administer than actuarial assessments; item ratings often are more nuanced and information might not be readily available on file to code all items. That said, recent reviews show that actuarial and structured professional judgment instruments produce assessments with commensurate rates of validity in predicting recidivism (Fazel, Singh, Doll & Grann, 2012).

Types of Items and Content Domains

Risk assessment instruments include items that represent characteristics of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or circumstances (e.g., living situation, employment status) that are associated with the likelihood of offending. *Risk factors* are those characteristics that increase risk of offending, whereas *protective factors* are those that reduce risk. Inclusion of protective factors is risk assessment instruments—designed to assess recidivism risk or otherwise—is relatively rare; however, there is mounting evidence that they contribute unique information and improve predictive validity above and beyond consideration risk factors (e.g., Desmarais, Nicholls, Wilson, & Brink, 2012).

Most frequently, recidivism risk assessment instruments focus on biological, psychological and social characteristics; however, more macro-level factors—such as service, system and societal variables—also may affect risk, but are rarely included in recidivism risk assessment instruments.

In a relatively recent review of the literature, Andrews, Bonta and Wormith (2006) identified a shortlist of the most "powerful" risk factors for recidivism across offenders and situations. These include:

1. History of antisocial behavior
2. Antisocial personality pattern
3. Antisocial cognition
4. Antisocial associates
5. Family and/or marital problems
6. School and/or work problems
7. Leisure and/or recreation problems
8. Substance abuse

These "Central Eight" have been widely accepted as the most important domains to be assessed and targeted in risk assessment and management efforts.

Finally, risk and protective factors can either be static or dynamic in nature. *Static factors* are historical or otherwise unchangeable characteristics (e.g., history of antisocial behavior) that help establish absolute level of risk. In contrast, *dynamic factors* are changeable characteristics (e.g., substance abuse) that establish a relative level of risk and help inform intervention; they can be either relatively *stable*, changing relatively slowly over time (e.g., antisocial cognition) or *acute* (e.g., mood state) (Hanson & Harris, 2000). Research shows that dynamic factors add incrementally to the predictive validity of static factors and that the former may be more relevant to short-term outcomes and rehabilitation efforts (Wilson, Desmarais, Nicholls, Hart, & Brink, in press), whereas the latter to longer term outcomes and risk classification (Hart, Webster, & Douglas, 2001). Thus, there are important benefits to considering both static and dynamic factors in assessing recidivism risk.

Focus of the Present Review

In this review, we summarize the research conducted in the U.S. examining the performance of instruments designed to assess risk of recidivism among adult offenders, including new offenses and violation of conditions. We focus specifically on performance of tools validated and currently used in correctional settings in the United States.[1] By identifying those instruments that produce the most consistent and accurate assessments, decision makers may be able to make more informed choices regarding which measure(s) to implement and how they should invest financial and staff resources.

[1] For meta-analytic reviews of instruments used in other jurisdictions and research outside the United States see Fazel et al., 2012; Gendreau, Goggin, & Little, 1996; Smith, Cullen, & Latessa, 2009).

METHODS OF THE CURRENT REVIEW

Search Criteria and Process

Identifying Risk Assessment Instruments Used in Correctional Settings in the U.S.

Instruments designed to assess risk of recidivism were identified by searching academic research databases (PsycINFO and the U.S. National Criminal Justice Reference Service Abstracts) and Google using combinations of the following keywords: *risk assessment, instrument, tool, general, recidivism*, *offending*, *probation revocation, parole violation*, and *prediction*. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues.

We identified instruments designed to assess risk of recidivism by searching academic research databases and the Google search engine. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues. Criteria for instruments to be included in the review were: a) designed to assess the likelihood of general recidivism (i.e., new offenses and violation of conditions); b) intended for assessing adult offenders (18 years of age and older); c) currently or recently used in correctional settings in the United States; and d) validated in the United States.

Instruments were excluded from our review if they: a) were designed to assess the likelihood of specific offenses (e.g., sexual offenses, violent offenses, spousal assault); b) were intended for assessing juvenile offenders (less than 18 years of age); c) were not used in correctional settings in the United States; d) had not been validated in the United States; or e) were developed for use in a specific institution or ward.

We also excluded violence risk assessment instruments (e.g., Historical, Clinical, Risk Management-20, Violence Risk Appraisal Guide), clinical inventories (e.g., Beck Depression Inventory, Novaco Anger Scale), personality assessments (e.g., Psychopathy Checklist-Revised, Personality Assessment Inventory), and criminal thinking scales (e.g., TCU Criminal Thinking Scales, Psychological Inventory of Criminal Thinking) from our formal review. These instruments were not designed to assess risk for general offending *per se*; however, they frequently are used for that purpose in correctional settings in the U.S. Thus, we briefly review their validity in predicting general offending later in this report.

Using these inclusion and exclusion criteria, we identified 19 instruments:

1. Community Risk/Needs Management Scale (CRNMS)
2. Correctional Assessment and Intervention System (CAIS)
3. Correctional Offender Management Profile for Alternative Sanctions (COMPAS)
4. Dynamic Factors Intake Assessment (DFIA)
5. Inventory of Offender Risks, Needs, and Strengths (IORNS)

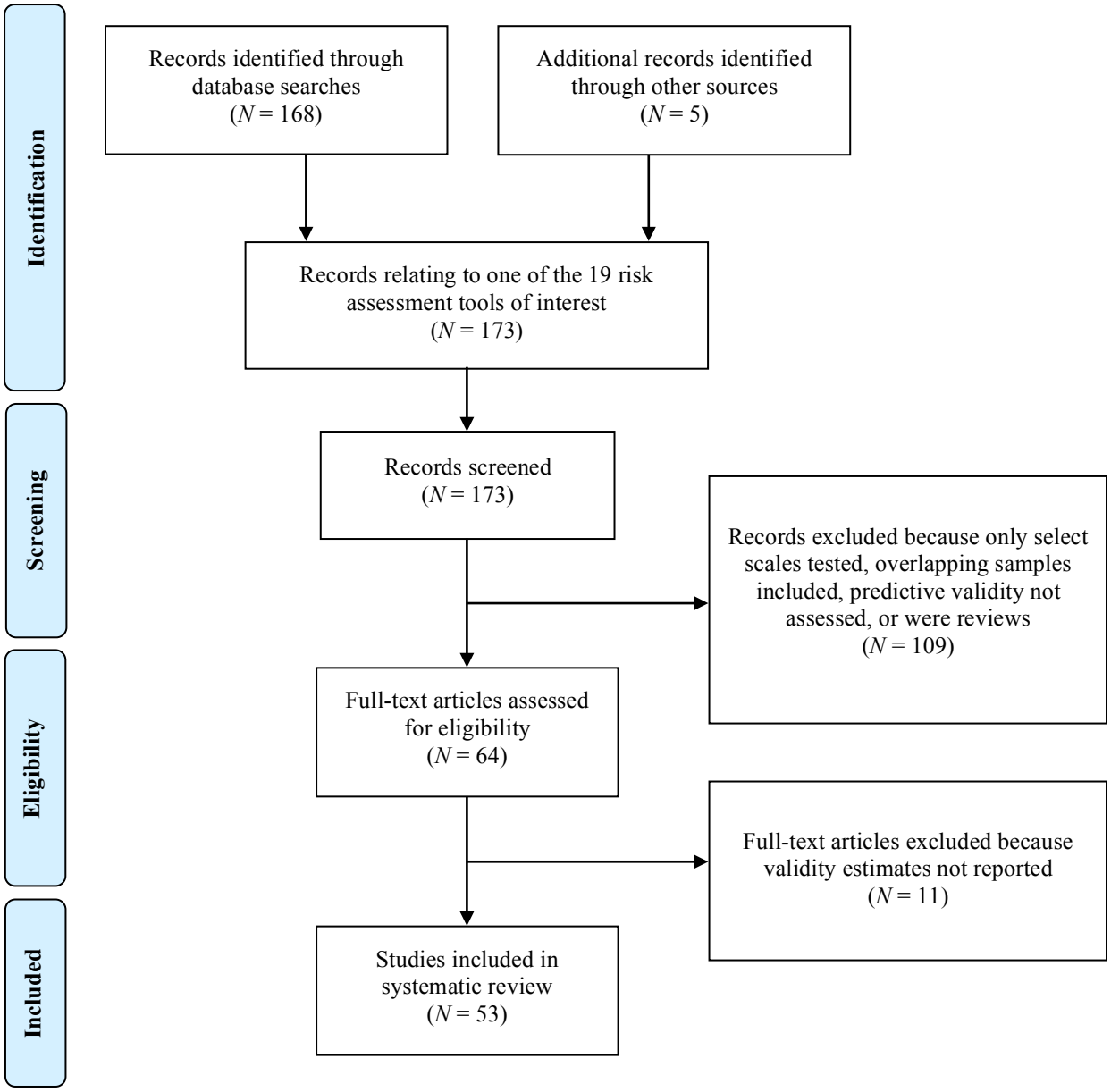
6. Level of Service instruments, including Level of Service/Case Management Inventory (LS/CMI), Level of Service/Risk, Need, Responsivity (LS/RNR), Level of Service Inventory (LSI), Level of Service Inventory-Revised (LSI-R), and Level of Service Inventory-Revised: Screening Version (LSI-R:SV)
7. Offender Assessment System (OASys)
8. Offender Group Reconviction Scale (OGRS)
9. Ohio Risk Assessment System, including the Ohio Risk Assessment System-Pretrial Assessment Tool (ORAS-PAT), Ohio Risk Assessment System-Community Supervision Tool (ORAS-CST), Ohio Risk Assessment System-Community Supervision Screening Tool (ORAS- CSST), Ohio Risk Assessment System-Prison Intake Tool (ORAS-PIT), and Ohio Risk Assessment System-Reentry Tool (ORAS-RT)
10. Federal Post Conviction Risk Assessment (PCRA)
11. Recidivism Risk Assessment Scales (RISc)
12. Risk Management System (RMS)
13. Risk of Reconviction (ROC)
14. Statistical Information of Recidivism Scale (SIR)
15. Salient Factor Score instruments, including the Salient Factor Score-1974 Version (SFS74), Salient Factor Score-1976 Version (SFS76), and Salient Factor Score-1998 Version (SFS98)
16. Self-Appraisal Questionnaire (SAQ)
17. Service Planning Instrument (SPIn) and Service Planning Instrument-Women (SPIn-W)
18. Static Risk and Offender Needs Guide (STRONG)
19. Wisconsin Risk and Needs (WRN) and Wisconsin Risk and Needs-Revised (WRN-R)

We also identified 47 instruments designed for use in specific jurisdictions. Detailed review is beyond the scope of the current report, but these instruments are listed in Appendix A.

Identifying Predictive Validity Studies

Studies investigating the predictive validity of the 19 above instruments were identified using the same databases, search engine and secondary sources as above, using both the acronyms and full names of the instruments as search criteria. We searched for studies published between 1970 and 2012 in peer-reviewed journals, as well as government reports, doctoral dissertations, and Master's theses. Studies were included in our review if their titles, abstracts, or methods sections described evaluations of validity in predicting general offending (including the violation of probation or parole conditions) conducted in the U.S. Studies were excluded if they only included some items or scales of an instrument. Using this search strategy, an initial total of 173 records was filtered to a final count of 53 studies (k samples = 72), including 26 journal articles (k = 30), 16 government reports (k = 31), nine doctoral dissertations (k = 9), and two Master's theses (k = 2). This systematic search process is visually depicted in the figure on the following page. A full list of the included studies is available from the authors upon request.

Systematic Search Conducted to Identify U.S. Predictive Validity Studies



Evaluation Criteria and Process

Three research assistants collected information about the characteristics of the risk assessment instruments (approach, number of items, types of items, domains measured, intended population and outcome) and studies (geographic location, context, design, population, sample size, sex, race/ethnicity, age, diagnostic composition, outcome, length of follow-up), as well as characteristics of the assessment process (setting, timing, format, assessor, sources of information, time needed to administer and score) from the included studies. They recorded information on inter-rater reliability and predictive validity, overall and by offender sex, race/ethnicity, study context, and recidivism outcome, where possible.

To evaluate performance, we computed the median performance indicators reported across studies for inter-rater reliability and predictive validity. For inter-rater reliability, we used the criteria presented in Table 1 to determine the practical significance of the median indicators.

Table 1. Criteria Used to Determine Practical Significance of Aggregate Inter-Rater Reliability Findings

INTER-RATER RELIABILITY	PERFORMANCE INDICATOR		
	Kappa (κ)	**Intra-class Correlation Coefficient (ICC)**	**Observed Agreement (%)**
Poor	.00 – .40	.00 – .40	< 70
Fair	.40 – .59	.40 – .59	70 – 79
Good	.60 – .74	.60 – .74	80 – 89
Excellent	.75 – 1.00	.75 – 1.00	90 – 100

Note. Table adapted from Cicchetti (2001, p. 697).

We also computed the median performance indicators for predictive validity. We used the criteria presented in Table 2 to determine the practical significance.

Table 2. Criteria Used to Determine Practical Significance of Aggregate Predictive Validity Findings

PREDICTIVE VALIDITY	PERFORMANCE INDICATOR				
	Cohen's *d*	**Correlation (r_{pb})**	**Area Under the Curve (AUC)**	**Odds Ratio (OR)**	**Somer's *d***
Poor	< .20	< .10	< .55	< 1.50	< .10
Fair	.20 – .49	.10 – .23	.55 – .63	1.50 – 2.99	.10 – .19
Good	.50 – .79	.24 – .36	.64 – .71	3.00 – 4.99	.20 – .29
Excellent	≥ .80	.37 – 1.00	.71 – 1.00	≥ 5.00	.30 – 1.00

Notes. Criteria were anchored to Cohen's *d* (1988) and based upon the calculations of Rice and Harris (2005) for AUC values, and Chen, Cohen, and Chen (2010) for the odds ratios. Somer's *d* values, as well as those for other performance indicators reported less frequently, also were interpreted in relation to Cohen's *d*.

In following sections of this report, we first summarize findings across instruments and then present findings of this review by instrument, respectively. We report only the interpretations of the practical significance of the performance indicators for both inter-rater reliability and predictive validity, but detailed statistical results are available upon request. We did not find any studies investigating the predictive validity of the CAIS, CRNMS, DFIA, LS/CMI, LS/RNR, LSI, OGRS, OASys, RISc, ROC, SFS98, SIR, or SPIn that met our inclusion criteria.

For a glossary of terms used in this report, including a brief explanation of the performance indicators included in Tables 1 and 2, see Appendix B.

SUMMARY OF FINDINGS ACROSS INSTRUMENTS

Characteristics of the Risk Assessment Instruments

Table 3 summarizes the characteristics of the risk assessment instruments. The number of items ranged from four for the ORAS-CSST to 130 for the IORNS. All instruments were intended for use across offender populations, with the exception of the SFS74, SFS76 and SFS81. Most were intended to be used to assess risk of new offenses, excluding violations). Of the nine instruments for which estimates were provided in the manual, length ranged from 5-10 minutes for the ORAS-CSST up to 60 minutes for the COMPAS. All were actuarial instruments.

Table 3. Characteristics of Risk Assessment Instruments

INSTRUMENTS		CHARACTERISTICS				
	k	**Items**	**Generation**	**Intended Population(s)**	**Intended Outcome(s)**	**Time (minutes)**
COMPAS	3	70	4th	Any Offender	Offenses & Violations	10-60
IORNS	1	130	3rd	Any Offender	Offenses & Violations	15-20
LSI-R	25	54	3rd	Any Offender	Offenses & Violations	30-40
LSI-R:SV	2	8	3rd	Any Offender	Offenses & Violations	10-15
ORAS-PAT	3	7	4th	Any Offender	Offenses	10-15
ORAS-CST	1	35	4th	Any Offender	Offenses	30-45
ORAS-CSST	1	4	4th	Any Offender	Offenses	5-10
ORAS-PIT	1	31	4th	Any Offender	Offenses	Unknown
ORAS-RT	1	20	4th	Any Offender	Offenses	Unknown
PCRA	2	30	4th	Any Offender	Offenses & Violations	15-30
RMS	2	65	4th*	Any Offender	Offenses	Unknown
SAQ	2	72	3rd	Any Offender	Offenses	15
SFS74	3	9	2nd	Parolees	Offenses	Unknown
SFS76	4	7	2nd	Parolees	Offenses	Unknown
SFS81	8	6	2nd	Parolees	Offenses	Unknown
SPIn-W	2	100	4th	Any Offender	Offenses	Unknown
STRONG[a]	1	26	4th	Any Offender	Offenses	Unknown
WRN	9	53	4th	Any Offender	Offenses	Unknown
WRN-R	1	52	4th	Any Offender	Offenses	Unknown

Notes. k = number of samples; Offenses = new arrest, charge, conviction, or incarceration; Violations = violations of conditions of supervision. [a]The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism. *The authors of the RMS describe it as being a 5th generation risk assessment instrument due to its exemplar-based approach.

Table 4 summarizes the types of factors included in the instruments. Only two instruments, the IORNS and the SPIn-W, include protective factors; all others include risk factors exclusively. The majority include static and dynamic factors, with the exception of the SFS instruments and the STRONG, both of which only include static factors. None only include only dynamic factors.

Table 4. Types of Items Included in the Risk Assessment Instruments

	TYPES OF ITEMS			
INSTRUMENTS	**Risk**	**Protective**	**Static**	**Dynamic**
COMPAS	X		X	X
IORNS	X	X	X	X
LSI-R	X		X	X
LSI-R:SV	X		X	X
ORAS-PAT	X		X	X
ORAS-CST	X		X	X
ORAS-CSST	X		X	X
ORAS-PIT	X		X	X
ORAS-RT	X		X	X
PCRA	X		X	X
RMS	X		X	X
SAQ	X		X	X
SFS74	X		X	
SFS76	X		X	
SFS81	X		X	
SPIn-W	X	X	X	X
STRONG[a]	X		X	
WRN	X		X	X
WRN-R	X		X	X

Note. [a]The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism.

Table 5 summarizes the content domains considered in the risk assessment instruments. All instruments include items assessing history of antisocial behavior and substance use problems. Slightly more than half of the instruments have items assessing mental health problems. Nine instruments include items assessing personality problems. Roughly two-thirds of the instruments consider attitudes, and similar proportions consider the influence of peers and relationships. The COMPAS and the LSI-R consider the most content domains. The ORAS-CST, ORAS-PIT, RMS, and SPIn-W evaluate all but one of the domains included in Table 5; the exception varied for each instrument. The SFS81 and STRONG instruments considered the fewest domains.

Table 5. Content Domains Assessed across the Risk Assessment Instruments

	ITEM CONTENT DOMAINS									
INSTRUMENTS	**Attitudes**	**Associates/ Peers**	**History of Antisocial Behavior**	**Personality Problems**	**Relationships**	**Work/ School**	**Recreation/ Leisure Activities**	**Substance Use Problems**	**Mental Health Problems**	**Housing Status**
COMPAS	X	X	X	X	X	X	X	X	X	X
IORNS	X	X	X	X	X	X		X	X	
LSI-R	X	X	X	X	X	X	X	X	X	X
LSI-R:SV	X	X	X		X	X		X	X	
ORAS-PAT			X			X		X		X
ORAS-CST	X	X	X	X	X	X	X	X		X
ORAS-CSST		X	X			X		X		
ORAS-PIT		X	X	X	X	X	X	X	X	X
ORAS-RT	X		X	X	X	X		X	X	
PCRA	X	X	X		X	X		X		
RMS	X	X	X	X	X	X		X	X	X
SAQ	X	X	X	X				X		
SFS74			X			X		X		X
SFS76			X			X		X		
SFS81			X					X		
SPIn-W	X	X	X		X	X	X	X	X	X
STRONG			X					X		
WRN	X	X	X		X	X		X	X	
WRN-R	X	X	X		X	X		X	X	

Note. [a]The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism.

Study Characteristics

Population and Sample Characteristics

More than a third of samples (40%) comprised inmates and roughly a quarter (22%), probationers. The remainder included at either parolees only (11%) or inmates and parolees (7%) or probationers and parolees (11%). Legal status was not reported in six samples (8%).

Studies generally provided few details regarding sample characteristics. Below we summarize findings regarding size, age, sex, race/ethnicity and mental health, when reported.

Sample size. The average sample size after attrition was 5,032.

Age. The average offender age at the time of risk assessment was 33.5 years.

Sex. In samples where sex was reported, the vast majority of offenders (86%) were male.

Race/ethnicity. In samples where race/ethnicity was reported, almost two-thirds (61%) were White and close to one-third (29%) were Black, with 14% identified as Hispanic. It is important to note that racial/ethnic categories were not consistent across studies. For instance, in some cases, authors reported the proportion of offenders identified as White, Black, or Hispanic (Farabee et al., 2010), while others reported prevalence of Hispanic and non-Hispanic offenders (Tillyer & Vose, 2011).

Mental health. Mental health characteristics were rarely reported. Only five studies--one evaluating the SFS74, one evaluating the SFS81, two evaluating the SPIn-W and one evaluating the WRN--described prevalence of major mental disorder (MMD), substance use disorder (SUD), or personality disorder. All offenders in the Howard (2007) study of the SFS81 were diagnosed with an MMD; slightly under half (46%) an SUD, and 11% had a personality disorder. This was the only study reporting prevalence of personality disorders. In one study of the SPIn-W all offenders had an SUD and three-quarters, a MMD (Meadon, 2012), whereas in the other study of the SPIn-W, just over half (53%) had a MMD (Millson et al., 2010). Only the WRN study reported prevalence by diagnosis. Bipolar disorder was the most prevalent MMD (36%) and schizophrenia, the least (16%), and alcohol abuse was the most prevalent SUD (48%) and amphetamines, the least (13%) (Castillo & Alardi, 2011). Finally, in the SFS74 study (Robuck, 1976), just under half of the sample (47%) suffered from alcohol abuse and 15%, illicit drug use.

Assessment Process

Table 6 shows the characteristics of the assessment process used in the studies. Risk assessments were complete by professionals in forensic services for over three-quarters of the studies (82%); the remaining assessments were conducted by the researchers (15%) or, in two studies, were self-administered. These assessments most often took place in a prison (28%) or in the community (38%), but at times were administered in jail (10%), a clinic or hospital (4%), or at another facility (6%). In terms of timing, roughly one third of assessments (36%) were conducted during community supervision, a quarter were completed pre-release (26%), and the remainder were conducted either prior to incarceration (11%) or at admission (10%). The source of information

used to complete the assessments were file reviews in 24 samples (33%), interviews in 12 samples (17%), and offender self-report in two samples (3%).

Table 6. Characteristics of the Assessment Process Used in Studies Included in this Review

CHARACTERISTICS	NUMBER OF SAMPLES *k (%)*
Assessor	
Researcher	11 (15.3)
Professional	59 (81.9)[a]
Offender (self-report)	2 (2.8)[b]
Assessment Setting	
Jail	7 (9.7)
Prison	20 (27.8)
Clinic/Hospital	3 (4.2)
Community	27 (37.5)
Other	4 (5.6)
Unstated/Unclear	11 (15.3)
Timing of Assessment	
Prior to incarceration	8 (11.1)
At admission	7 (9.7)
Prior to release	19 (26.4)
During community supervision	26 (36.1)
Unstated/Unclear	13 (18.1)
Source(s) of Information	
File review	24 (33.3)
Interview	12 (16.7)
Self-report	2 (2.8)
Mixed	18 (25.0)
Unstated/Unclear	16 (22.2)

Notes. Overall $k = 72$ samples. [a]Correctional officer ($k = 35$, 48.6%), parole officer ($k = 2$, 2.8%), probation officer ($k = 1$, 1.4%), other trained staff ($k = 14$, 19.4%), unstated/unclear ($k = 7$, 9.7%). [b]The SAQ, designed to be self-administered, was the only tool not administered by a researcher or professional.

Administration time was reported for only five instruments in a total of nine studies. For the LSI-R administration time ranged from 30 to 60 minutes for assessments conducted in the context of 'real world' practice (Holsinger et al., 2004; Lowenkamp et al., 2009), and 45 to 90 minutes in research studies (Evans, 2009; Latessa et al., 2009). The LSI-R:SV was reported to have a mean administration time of 10 minutes when completed in practice (Miller, 2006). In the same study, the IORNS required 15 minutes to complete; however, this estimate included only the interview portion of the assessment. Across three studies, administration time for the COMPAS varied

from 43 to 165 minutes (Brennan et al., 2009; Farabee et al., 2010; Farabee & Zhang, 2007). In the study evaluating SAQ assessments, assessments were reported to take approximately 20 minutes (Mitchell & McKenzie, 2006).

Study Designs and Procedures

More than two-thirds of studies (70%) used a prospective study design, an optimal approach for examining predictive validity, and the average length of follow-up was almost two years (23.5 months). Studies were most frequently conducted in midwestern states (38%) followed by the southwestern and northeastern (11% each) regions of the U.S.

Close to 70% of the studies examined general recidivism as the outcome; roughly a quarter (26%) considered a variety of outcomes, and the remainder (18%) focused specifically on violations. As a result, our knowledge of the validity of recidivism risk assessment instruments in predicting violations as opposed to other forms of recidivism is limited. The threshold for recidivism varied across studies, but arrest was used as an indicator in close to a third of studies (31%), followed in order by conviction (13%), incarceration (10%), revocations (4%), and charge (3%). Finally, assessments for the majority of samples (65%) were conducted in the context of 'real world' practice rather than for the purposes of research.

Nearly a third of the studies included in our review (31%, $k = 22$) were conducted by the author of the tool being studied. In fact, for many instruments, all of the studies included in our review were completed by the same people who developed the instrument under investigation. This was true for the IORNS (Miller, 2006), the PCRA (Johnson et al., 2011), the ORAS instruments (Latessa et al., 2008, 2009), the STRONG (Barnoski & Drake, 2007), and the WRN-R (Eisenberg et al., 2009). The authors of the RMS conducted one of two studies evaluating predictive validity of RMS assessments (Dow et al., 2005), and the authors of the COMPAS conducted one of three samples evaluating COMPAS assessments (Brennan et al., 2009). The authors of the SFS74, SFS76, and SFS81 evaluated two of three samples for the SFS74 (Hoffman & Beck, 1974), two of four for the SFS76 (Hoffman, 1980; Hoffman & Beck, 1980), and four of eight for the SFS81 (Hoffman, 1983, 1994; Hoffman & Beck, 1985).

Table 7. Design Characteristics and Procedures of Studies Included in this Review

CHARACTERISTICS	NUMBER OF SAMPLES *k (%)*
Study Context	
Research	25 (34.7)
Practice	47 (65.3)
Temporal Design	
Prospective	50 (69.4)
Retrospective	22 (30.6)
Geographical Region	
Northwest	2 (2.8)
Southwest	8 (11.1)
Midwest	27 (37.5)
Northeast	8 (11.1)
Southeast	5 (6.9)
Non-continental	1 (1.4)
Mixture	1 (1.4)
Unstated/Unclear	20 (27.8)
Type of Outcome	
General recidivism	50 (69.4)
Violation/Breach of conditions	13 (18.1)
Mixed	19 (26.4)
Threshold for Recidivism	
Arrest	22 (30.6)
Charge	2 (2.8)
Conviction	9 (12.5)
Incarceration	7 (9.7)
Revocation	3 (4.2)
Mixed	29 (40.3)

Note. k = number of samples

Inter-Rater Reliability

Inter-rater reliability was evaluated in only two studies, one examining the LSI-R and the other, the LSI-R:SV. In both cases, inter-rater reliability was excellent. Assessments were conducted by professionals rather than research assistants, providing evidence of *field* reliability, specifically.

Predictive Validity

Overall

Table 8 presents the practical significance of predictive validity performance indicators across studies. Overall, and consistent with prior research reviews, no one instrument stands out as producing more accurate instruments than the others, with validity varying with the indicator reported. Odds ratios generally suggested poor performance for the majority of instruments, with only one instrument (the SFS81) demonstrating good predictive validity. In contrast, Somer's *d* values ranged from good to excellent. AUCs and point-biserial correlations each ranged from fair to excellent across instruments. Below, we describe predictive validity by instrument.

COMPAS. The predictive validity of COMPAS assessments ranged from poor to good, as a function of performance indicator; more studies used the AUC and, thus, reported good validity.

LSI instruments. LSI-R assessments were evaluated in the most samples. Predictive validity was good across studies and indicators, with the exception of odds ratios. Validity of LSI-R:SV assessments ranged from fair to good.

ORAS instruments. Across instruments and studies, ORAS assessments demonstrated excellent point-biserial values. No other performance indicators were reported.

PCRA. PCRA assessments were evaluated in only two samples, with AUC values suggesting excellent predictive validity in both. No other performance indicators were reported.

RMS. In three samples, RMS assessments showed good performance according to the AUC values. No other performance indicators were reported.

SFS instruments. SFS74, SFS76, and SFS81 assessments showed predictive validity ranging from good to excellent, with the SFS81 outperforming the previous versions.

SPIn-W. SPIn-W assessments showed good performance according to the AUC but poor performance according to the odds ratio.

STRONG. In one study, predictive validity of STRONG assessments was excellent according to the AUC. No other performance indicators were reported.

WRN instruments. Predictive validity for WRN and WRN-R assessments ranged from poor to good, depending on the performance indicator used.

No studies reported predictive validity of IORNS or SAQ assessments using these indicators.

Table 8. Summary of Predictive Validity Findings by Performance Indicator across Studies

INSTRUMENT	MEDIAN PERFORMANCE INDICATOR							
	k	**AUC**	***K***	r_{pb}	***k***	**OR**	***k***	**Somer's *d***
COMPAS	3	Good	1	Fair	1	Poor	–	–
LSI-R	5	Good	21	Good	6	Poor	2	Good
LSI-R:SV	1	Fair	1	Good	–	–	–	–
ORAS-PAT	–	–	5	Good	–	–	–	–
ORAS-CST	–	–	1	Excellent	–	–	–	–
ORAS-CSST	–	–	1	Excellent	–	–	–	–
ORAS-PIT	–	–	1	Excellent	–	–	–	–
ORAS-RT	–	–	1	Excellent	–	–	–	–
PCRA	2	Excellent	–	–	–	–	–	–
RMS	3	Good	–	–	–	–	–	–
SFS74	–	–	–	–	–	–	2	Good
SFS76	–	–	1	Excellent	–	–	2	Good
SFS81	–	–	4	Excellent	2	Good	5	Excellent
SPIn-W	1	Excellent	–	–	1	Poor	–	–
STRONG	1	Excellent	–	–	–	–	–	–
WRN	3	Good	6	Fair	1	Poor	–	–
WRN-R	1	Good	–	–	–	–	–	–

Notes. *k* = number of samples; AUC = area under the receiver operating characteristic curve; r_{pb} = point-biserial correlation coefficient; OR = odds ratio. Medians were calculated using either total scores or risk bins. There were no studies reporting predictive validity of the IORNS or SAQ using these performance indicators.

Validity of Total Scores in Predicting Different Forms of Recidivism

Table 9 presents the validity of total scores in predicting different forms of recidivism. For general offending *including* violations, predictive validity ranged from poor for SPIn-W assessments to excellent for SFS76 and SFS81 assessments. For general offending *excluding* violations, total scores for over two-thirds of instruments had either good or excellent predictive validity. Specifically, predictive validity ranged from fair for ORAS-PAT assessments to excellent for the ORAS-CST, ORAS-CSST, PCRA, and STRONG assessments. For *violations*, predictive validity ranged from fair COMPAS assessments to excellent WRN assessments. Below, we describe predictive validity by instrument.

COMPAS. The COMPAS total scores demonstrated good validity in predicting general offending *excluding* violations, but was only fair for violations only.

LSI instruments. LSI-R total scores showed good predictive validity for both general offending *including* violations and violations only, and ranged from fair to good validity in general offending *excluding* violations.

ORAS instruments. With the exception of the ORAS-PAT, the total scores on the ORAS instruments all demonstrated predictive validity ranging from good to excellent for general offending *excluding* violations. ORAS-PAT total scores, however, were only fair at predicting general offending outcomes, though predictive validity was good for violations only.

RMS. RMS total scores demonstrated good validity in predicting general offending *excluding* violations, as well as violations only.

SFS instruments. SFS76 and SFS81 total scores showed excellent validity in predicting general offending *including* violations. No studies reported predictive validity of SFS74 total scores by outcome.

SPIn-W. SPIn-W total scores had poor validity in predicting general offending *including* violations.

STRONG. STRONG total scores demonstrated excellent validity in predicting general offending *excluding* violations.

WRN instruments. WRN total scores ranged from fair to good in their ability to predict general offending excluding violations. Predictive validity was excellent for violations only. WRN-R total scores showed good validity in predicting general offending *excluding* violations.

Overall, total scores of SFS76 and SFS81 total scores stood out as excellent predictors of general offending *including* violations. Total scores on the ORAS-CST, ORAS-CSST, PCRA, and STRONG were excellent predictors of general offending *excluding* violations. WRN total scores stood alone as excellent in predicting violations only. It is important to note, however, the small number of studies examining these outcomes; SFS76, ORAS-CST, ORAS-CSST, STRONG, and WRN assessments were evaluated in only one sample, compared to the 26 samples evaluating LSI-R assessments.

Table 9. Validity of Total Scores in Predicting Different Forms of Recidivism

INSTRUMENTS	OUTCOMES					
	k	General Offending (including Violations)	*k*	General Offending (excluding Violations)	*k*	Violations Only
COMPAS	–	–	5	Good	1	Fair
LSI-R	3	Good	26	Fair-Good	7	Good
LSI-R:SV	–	–	2	Fair-Good	–	–
ORAS-PAT	1	Fair	2	Fair	2	Good
ORAS-CST	–	–	1	Excellent	–	–
ORAS-CSST	–	–	1	Excellent	–	–
ORAS-PIT	–	–	1	Good	–	–
ORAS-RT	–	–	1	Good	–	–
PCRA	–	–	2	Excellent	–	–
RMS	–	–	1	Good	1	Good
SFS74	–	–	–	–	–	–
SFS76	1	Excellent	–	–	–	–
SFS81	6	Excellent	–	–	–	–
SPIn-W	1	Poor	–	–	–	–
STRONG	–	–	1	Excellent	–	–
WRN	–	–	8	Fair-Good	1	Excellent
WRN-R	–	–	1	Good	–	–

Notes. k = number of samples. General Offending = new arrest, charge, conviction, or incarceration; Violations = violations of conditions of supervision.

Predictive Validity of Risk Classifications

Table 10 presents the validity of risk classifications in predicting different forms of recidivism. Validity of risk classifications in predicting general offending *including* violations was excellent for SFS74, SFS76, and SPIn-W assessments. For general offending *excluding* violations, the predictive validity was fair for WRN assessments and excellent for RMS and SFS81 assessments. Validity of SFS risk classifications in predicting general offending *including* violations also was excellent.

No U.S. studies examined the predictive validity of risk classifications for violations alone. There also were no U.S. studies reporting predictive validity of the risk classifications for the COMPAS, IORNS, LSI-R, LSI-R:SV, ORAS-PAT, ORAS-CST, ORAS-CSST, ORAS-PIT, ORAS-RT, PCRA, SAQ, STRONG, or WRN-R for any of the recidivism outcomes.

Table 10. Validity of Risk Classifications in Predicting Different Forms of Recidivism

INSTRUMENTS	OUTCOMES			
	k	General Offending (including Violations)	*k*	General Offending (excluding Violations)
RMS	–	–	1	Excellent
SFS74	2	Excellent	–	–
SFS76	2	Excellent	–	–
SFS81	4	Excellent	1	Excellent
SPIn-W	1	Excellent	–	–
WRN	–	–	1	Fair

Notes. *k* = number of samples. There were no studies that reported the predictive validity of the risk classifications for the COMPAS, IORNS, LSI-R, LSI-R:SV, ORAS-PAT, ORAS-CST, ORAS-CSST, ORAS-PIT, ORAS-RT, PCRA, SAQ, STRONG, or WRN-R using these performance indicators. The risk bins used to classify offenders were those recommended by instrument authors.

Predictive Validity across Offender Subgroups

Sex. Table 11 presents the validity of total scores in predicting recidivism by the offender's sex. Overall, predictive validity ranged from fair to excellent for both male and female offenders. Some instruments performed equally well for male and female offenders; for instance, COMPAS assessments demonstrated good predictive validity for both sexes. STRONG assessments also demonstrated excellent validity for both male and female offenders. Finally, predictive validity for the ORAS instrument for which comparisons were possible—namely, the ORAS-CST, ORAS-CSST, ORAS-PIT, and ORAS-RT—ranged from good to excellent for both male and female offenders.

Other instruments showed differential performance by offender sex. In particular, LSI-R assessments showed good predictive validity for male offenders, but predictive validity was only fair for female offenders. Similarly, LSI-R:SV assessments showed only fair predictive validity for female offenders, but ranged from fair to good in its predictions for male offenders.

Other instruments were evaluated in exclusively male or female offenders. Predictive validity of SFS76 and SFS81 assessments, for example, were only evaluated for male offenders; SFS76 total scores demonstrated excellent validity, while validity of SFS81 assessments ranged from good to excellent. WRN total scores also were evaluated for male offenders and showed fair validity. Designed for women, the SPIn-W has only been evaluated for female offenders and showed good validity.

No studies reported predictive validity of assessments by offender sex for the IORNS, ORAS-PAT, PCRA, RMS, SAQ, SFS74, or WRN-R.

Table 11. Validity of Total Scores in Predicting Recidivism by Offender Sex

INSTRUMENTS	OFFENDER SEX			
	k	**Male**	***k***	**Female**
COMPAS	2	Good	2	Good
LSI-R[a]	9	Good	8	Fair
LSI-R:SV	2	Fair-Good	1	Fair
ORAS-CST	1	Excellent	1	Good
ORAS-CSST	1	Good	1	Excellent
ORAS-PIT	1	Good	1	Good
ORAS-RT	1	Good	1	Excellent
SFS76[b]	1	Excellent	–	–
SFS81[c]	–	Good-Excellent	–	–
SPIn-W[d,e]	–	–	2	Good
STRONG	1	Excellent	1	Excellent
WRN	1	Fair	–	–

Notes. k = number of performance indicators. No studies reported predictive validity estimates by sex for the IORNS, ORAS-PAT, PCRA, RMS, SAQ, SFS74, or WRN-R using the included performance indicators.
[a]One LSI-R sample specifically included technical violations in the operational definition of recidivism.
[b]One SFS76 sample specifically included technical violations in the operational definition of recidivism.
[c]One SFS81 sample specifically included technical violations in the operational definition of recidivism.
[d]Both SPIn-W samples were composed entirely of women.
[e]One SPIn-W sample reported predictive validity of the risk categorizations rather than total scores.

Race/ethnicity. Comparisons by offender race/ethnicity were only possible for assessments completed using the COMPAS and LSI-R. For COMPAS assessments, predictive validity was good for White and Black offenders. For LSI-R assessments, predictive validity ranged from poor to good across White, Black, Hispanic, and non-White offenders, with performance varying largely depending on sample size and performance indicator rather than race/ethnicity. Together, these findings fail to provide evidence of differential performance of COMPAS and LSI-R assessments as a function of offender race/ethnicity.

Diagnostic categories. No comparisons of predictive validity within or across instruments as a function of mental, substance use or personality disorders were possible. Even when these sample characteristics were reported, predictive validity was not provided by subgroup. As for race/ethnicity, there is a critical need for research examining risk assessment accuracy between mentally disordered and nondisordered offenders as well as across diagnostic subgroups. That said, prior meta-analytic work has found the predictors of recidivism to be comparable for mentally disordered offenders (Bonta, Law, & Hanson, 1998), suggesting that assessments also may perform comparably.

Predictive Validity in the Context of Research versus 'Real World' Practice

Recently there has been a focus on the need to establish the performance of risk assessment instruments *in the field*. Much of our knowledge stems from research-based studies, in which researchers can carefully train and monitor assessors. In 'real world' practice, however, such training and oversight is not necessarily present (Douglas, Otto, Desmarais, & Borum, in press).

Comparisons between the performance of assessments completed in the context of research and practice were possible for the LSI-R, RMS, SPIn-W, and WRN. Whereas both LSI-R and WRN total scores performed comparably whether conducted in research studies or in the context of 'real world' practice, RMS risk classifications had better predictive validity when completed by researchers rather than practitioners (though performance was still good). SPIn-W assessments also seemed to perform better in research studies than in practice, though predictive validity in both contexts was excellent. However, in the research context, predictive validity of the SPIn-W was evaluated vis-à-vis the total scores while in practice, the risk classifications were examined, preventing direct comparisons of the results.

No comparisons were possible for the other risk assessment instruments. Specifically, COMPAS, IORNS, SFS76, and SFS81 assessments have only been evaluated in the context of 'real world' practice, and the LSI-R:SV, ORAS tools, PCRA, SAQ, SFS74, STRONG, and WRN-R assessments have only been evaluated in research studies.

SUMMARY OF FINDINGS BY INSTRUMENT

In this section describe each risk assessment instrument and summarize findings of U.S. studies examining predictive validity. Instruments are presented in alphabetical order.

Correctional Offender Management Profiling for Alternative Sanctions

Description

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Brennan et al., 2009).

The COMPAS contains static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing, divided across 22 scales (Blomberg, Bales, Mann, Meldrum, & Nedelec, 2010). Scores on the self-report assessment, data from official records, and information from interview are used to arrive at an overall risk score for each offender. The COMPAS is a 4th generation risk assessment instrument.

COMPAS assessments are completed through a combination of a computer-assisted self-report questionnaire, an interview conducted by a trained assessor, and data collected from the offender's records. The instrument can be purchased from Northpointe at www.northpointeinc.com. Assessors must complete a 2-day training session that covers practical use, interpretation of results, and case planning strategies in order to administer the COMPAS. Advanced training options that focus on the theoretical underpinnings of offender assessments, gender responsivity, motivational interviewing, and other topics are available.

U.S. Research Evidence

In total, four studies have evaluated predictive validity of COMPAS assessments in U.S. samples. Blomberg and colleagues (2010) found that those identified as higher risk were indeed more likely to recidivate; specifically, 7% of those identified to be low risk recidivated, 16% of those identified as medium risk, and 27% of those identified as high risk. In other samples, predictive validity was good for general offending (Brennan, Dieterich, & Ehret, 2009) and fair for violations (Farabee & Zhang, 2007). Predictive validity for male and female offenders has ranged from good to excellent (Brennan et al., 2009).

There were no studies published between 1970 and 2012 comparing predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

Practical Issues and Considerations

For the self-report portion of the assessment, the computer upon which the offender completes the questionnaire must have Internet access and run on Windows. The assessor must complete training to be qualified to administer the structured interview.

Selected References and Suggested Readings

Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *Validation of the COMPAS risk assessment classification instrument.* City, ST: publisher. Retrieved from http://www.criminologycenter.fsu.edu/p/pdf/pretrial/Broward%20Co.%20COMPAS%20Validation%202010.pdf

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior, 36*, 21-40.

Farabee, D., Zhang, S., Roberts, R. E. L., & Yang, J. (2010). *COMPAS validation study: Final report.* California Department of Corrections and Rehabilitation. Retrieved from http://www.cdcr.ca.gov/adult_research_branch/Research_Documents/COMPAS_Final_Report_08-11-10.pdf

Federal Post Conviction Risk Assessment

Description

The Federal Post Conviction Risk Assessment (PCRA) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011).

The PCRA contains 30 static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at work or school, and substance use problems. Self-report assessment scores are combined with probation officer assessment scores to arrive at an overall risk score. The PCRA is a 4th generation risk assessment instrument.

PCRA assessments comprise two components: 1) the Officer Assessment, and 2) Offender Self-Assessment. The self-report questionnaire consists of items that are "scored" and "unscored". The 15 scored items are those that have been shown in studies conducted by the Administrative Office of U.S. Courts (Administrative Office) to predict recidivism and contribute to the overall risk score. The 15 unscored items have been shown in other research to predict recidivism, but have not been evaluated by the Administrative Office. They are included to inform intervention strategies, but do not contribute to the risk scores. Assessments must be administered by probation officers who have passed the online certification test created and offered by the Administrative Office; the Administrative Office prohibits uncertified assessors from accessing the PCRA. Prior to the online certification, probation officers must complete 16 hours of

training. They also must renew their certification every year. The PCRA is available through the Administrative Office at www.uscourts.gov.

U.S. Research Evidence

One study has assessed the predictive validity of PCRA assessments in two large U.S. samples. Johnson, Lowenkamp, VanBenschoten, and Robinson (2011) found excellent predictive validity in both. As of December 2012, there were no studies comparing predictive validity between assessments completed in research and practice contexts, by offender sex or by offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

Practical Issues and Considerations

Though promising, research evidence is limited to date. As noted above, there were no published evaluations of the reliability and predictive validity of PCRA assessments that met our inclusion criteria beyond the initial construction and validation study. However, a study published early this year by the instrument's authors (Lowenkamp, Johnson, VanBenschoten, & Robinson, 2013) compared predictive validity between research and practical contexts and reported high rates of inter-rater agreement. Independent replication is needed.

Selected References and Suggested Readings

Administrative Office of the United States Courts, Office of Probation and Pretrial Services. (2011, September). *An overview of the Federal Post Conviction Risk Assessment.* Retrieved from http://www.uscourts.gov/uscourts/FederalCourts/PPS/PCRA_Sep_2011.pdf

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The construction and validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, *75*, 16-29.

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). *Psychological Services, 10,* 87-96.

Inventory of Offender Risk, Needs, and Strengths

Description

The Inventory of Offender Risk, Needs, and Strengths (IORNS) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Miller, 2006a).

The IORNS contains 130 static, dynamic, risk, and protective factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, substance use problems, mental health problems,

and housing. Individual item responses are summed to create Static, Dynamic and Protective indexes as well as an Overall risk index. There also are two validity scales. The IORNS is a 3rd generation risk assessment instrument.

The IORNS is a true/false self-report questionnaire completed by the offender and requires 3rd grade reading level. The IORNS manual indicates that assessments take 15 to 20 minutes to administer, and 20 to 25 minutes to score. There are no training requirements for assessors, provided the purchaser of the exam has a degree in forensic or clinical psychology or psychiatry as well as certification in psychological testing. The purchaser also is responsible for overseeing the scoring of the assessment. IORNS assessments are available through Psychological Assessment Resources (parinc.com). Costs include those associated with the manual, interview guides, and assessment forms. For further information on pricing, see www.parinc.com.

U.S. Research Evidence

Predictive validity of IORNS assessments have been evaluated in only one U.S. sample conducted by the author of the instrument. Miller (2006b) found that offenders with higher Overall Risk Indices were in jail more frequently and had more non-violent arrests than those with lower scores. Similarly, those offenders who had more half-way house rule violations have significantly lower Overall Risk, and Dynamic Needs Indices.

As of December 2012, there were no published studies comparing predictive validity in U.S. samples between assessments completed in research and practice contexts, by recidivism outcome, offender sex, or offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

Practical Issues and Considerations

Though findings are promising, predictive validity of IORNS assessments has only been evaluated in one study conducted by the instrument developer that met our inclusion criteria; independent replication is needed.

Selected References and Suggested Readings

Miller, H. A. (2006a). *Manual of the Inventory of Offender Risk, Needs, and Strengths* (IORNS). Odessa, FL: Psychological Assessment Resources.

Miller, H. A. (2006b). A dynamic assessment of offender risk, needs, and strengths in a sample of pre-release general offenders. *Behavioral Sciences & the Law*, *24*, 767-782.

Level of Service Instruments

Description

The Level of Service family of instruments includes the Level of Service Inventory-Revised (LSI-R) and Level of Service Inventory-Revised: Screening Version (LSI-R:SV), actuarial risk assessment instruments intended to assess risk for general offending and violations across offender populations (Andrews & Bonta, 1995; 1998).

The LSI-R contains 54 static and dynamic risk factors. Content areas include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing. Item responses are scored and summed for a total score from 0 to 54 that is used to classify risk as: Low = 0-23; Medium = 24-33; and High = >34. The LSI-R is a 3rd generation risk assessment instrument.

The LSI-R:SV contains eight static and dynamic items selected from the LSI-R. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, and substance abuse problems. Individual item responses are scored and summed for a total score ranging from 0-9. This score is used to determine if the offender requires a full LSI-R assessment. Like the interview-based version, the LSI-R:SV is also a 3rd generation risk assessment instrument.

LSI-R and LSI-R:SV assessments are completed through interview and file review, a process estimated to require approximately 30-40 minutes for the LSI-R and 10-15 minutes for the LSI-R:SV (though studies we reviewed reported longer completion times – see below). The assessor does not need formal training, but scoring must be overseen by someone who has post-secondary training in psychological assessment. The LSI-R and LSI-R:SV materials are available through Multi-Health Systems (www.mhs.com). Costs include those associated with the manual, interview guides, and assessment forms. For further information on pricing, see www.mhs.com.

U.S. Research Evidence

Predictive validity of LSI-R total scores had been evaluated in 25 U.S. samples as of December 2012. Performance in has ranged from poor to good, with the median on the cusp of fair and good. There were no studies examining the predictive validity of the risk classifications (as opposed to total scores) that met criteria for inclusion in this review. LSI-R total scores seem perform slightly better for men than for women, though performance is in the fair-good range for both. U.S. studies have not shown differences in validity as a function of racial/ethnicity. Predictive validity for total scores completed in the context of research and practice also is comparable. Validity in predicting is general offending is slightly better than violations. In the one U.S. study reporting inter-rater reliability data, agreement ranged from poor to excellent across content domains, but was excellent overall (Simourd, 2006).

Predictive validity of the LSI-R:SV has only been examined in two U.S. samples with mixed results: one study showed fair performance (Walters, 2011) and the other, good (Lowenkamp et

al., 2009). The LSI-R:SV seems to perform better for men (good predictive validity) than for women (fair predictive validity). There had been no studies comparing predictive validity between total scores and risk classifications, assessments completed in research and practice, by offender race/ethnicity, or by recidivism outcome as of December 2012. Because the LSI-R:SV is a self-report instrument, inter-reliability is not relevant.

The LSI-R instruments have been evaluated extensively outside of the United States. For example, there have been many evaluations of the predictive validity and inter-rater reliability of the LSI-R conducted in Canada and Europe (see, for example, Vose, Cullen, & Smith, 2008), but none have compared the predictive validity between total scores and risk classifications. Similarly, the LSI-R:SV has been studied outside of the United States (e.g., Daffern et al., 2005; Ferguson et al., 2005), but the research does not address the limitations described above.

Practical Issues and Considerations

Researchers and professionals have reported administration times that deviate considerably from the LSI-R manual's estimate of 30-40 minutes, including 60 minutes in one sample (Holsinger et al., 2004) and 45-90 minutes in two others (Evans, 2009; Latessa et al., 2009).

There is considerable variation in the cut-off scores used for the risk categories. The manual encourages altering cut-off scores based on offense group characteristics, but research should be conducted *prior to* implementation to establish the validity of revised cut-off scores (Kim, 2010).

A recent addition to the Level of Service family of instruments is the Level of Service/Case Management Inventory (LS/CMI), an actuarial risk assessment with 43 items intended to aid professionals in offender management with late adolescent and adult offenders. No studies examining the LS/CMI met our inclusion criteria. However, there have been many evaluations of the predictive validity of the LS/CMI conducted outside of the United States (Andrews et al., 2011). Studies have included samples of male and female, as well as young offenders. Performance estimates for these populations ranged from fair to excellent. Inter-rater reliability has also been evaluated for total scores and found to be excellent (Rettinger & Andrews, 2010).

Selected References and Suggested Readings

Andrews, D. A. & Bonta, J. (1995). *LSI-R: The Level of Service Inventory-Revised user's manual.* Toronto: Multi-Health Systems.

Andrews, D. A., & Bonta, J. L. (1998). *Level of Service Inventory-Revised: Screening Version (LSI-R:SV): User's manual.* Toronto: Multi-Health Systems.

Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with Level of Service general risk and need. *Criminal Justice & Behavior, 38,* 413-432.

Daffern, M., Ogloff, J. R. P., Ferguson, M., & Thomson, L. (2005). Assessing risk for aggression in a forensic psychiatric hospital using the Level of Service Inventory-Revised:Screening Version. *International Journal of Forensic Mental Health, 4*, 201-206.

Ferguson, A. M., Ogloff, J. R. P., & Thomson, L. (2005). Predicting recidivism by mentally disordered offenders using the LSI-R:SV. *Criminal Justice & Behavior, 36,* 5-20.

Lowenkamp, C. T., Lovins, B., & Latessa, E. J. (2009). Validating the Level of Service Inventory−Revised and the Level of Service Inventory: Screening Version with a sample of probationers. *The Prison Journal, 89*, 192-204.

Rettinger, L. J., & Andrews, D. A. (2010). General risk and need, gender specificity, and the recidivism of female offenders. *Criminal Justice & Behavior, 37,* 29-46.

Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation, 72*, 22-29.

Ohio Risk Assessment System

Description

The Ohio Risk Assessment System (ORAS) is comprised of five actuarial risk assessment instruments intended to assess risk for recidivism across offender populations (Latessa et al., 2009): the 7-item Pretrial Assessment Tool (ORAS-PAT), the 4-item Community Supervision Screening Tool (ORAS-CSST), the 35-item Community Supervision Tool (ORAS-CST), the 31-item Prison Intake Tool (ORAS-PIT), and the 20-item Prison Re-entry Tool (ORAS-RT). Each includes static and dynamic risk factors and is designed for use at a specific stage in the criminal justice system; namely, pretrial, community supervision, institutional intake, and community reentry. Assessments identify criminogenic needs and place offenders into risk categories. An additional sixth instrument, the Prison Screening Tool (ORAS-PST), is designed to identify low risk inmates who do not need the full ORAS-PIT assessment.

Item responses are scored and summed to create total scores which are compared against risk classification cut-off values. The ORAS-PAT has a range from 0 to 9, the ORAS-CSST from 0 to 7, the ORAS-CST from 0 to 49, the ORAS-PIT from 3 to 29, and the ORAS-RT from 0 to 28. Each tool considers the offender's history of antisocial behavior, circumstance at school or work, and substance abuse problems; some also evaluate additional domains, such as attitudes (e.g., ORAS-CST, ORAS-RT), and mental health problems (e.g., ORAS-PIT, ORAS-RT). Together, the ORAS system reflects the 4th generation of risk assessment.

The ORAS tools are completed through a structured interview and analysis of official records; the ORAS-CSST, ORAS-PIT, and ORAS-RT additionally use self-report questionnaires. Assessors must complete a 2-day training package that accompanies the tool prior to administering any assessments. The ORAS is published by the Ohio Department of Rehabilitation and Correction (http://www.drc.ohio.gov). The system is non-proprietary and can

be obtained from the Center of Criminal Justice Research, University of Cincinnati (http://www.uc.edu/corrections/services/risk-assessment.html).

U.S. Research Evidence

ORAS-PAT total scores demonstrated fair validity in predicting arrest in the construction sample and good validity in the validation sample (Latessa et al., 2009). A second evaluation found fair predictive validity for ORAS-PAT assessments, good validity for ORAS-PIT and ORAS-RT assessments, and excellent validity for ORAS-CCST and ORAS-CST assessments (Lowenkamp, Lemke, & Latessa, 2008). ORAS-PST assessments have not been included in these evaluations.

Predictive validity of ORAS assessments differs somewhat as a function of offender sex. Specifically, ORAS-CST assessments performed slightly better for male than female offenders, though predictive validity was excellent in both cases. Conversely, ORAS-PIT and ORAS-RT assessments performed better for female (excellent predictive validity) than male offenders (good). ORAS-CSST assessments, in contrast, have shown comparable predictive validity for both male and female offenders. The ORAS-PAT total scores have demonstrated better validity in predicting violations (good) than general offending (fair).

As of December 2012, there had been no U.S. studies comparing predictive validity between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity that met our inclusion criteria. There also had not been any evaluations of inter-rater reliability.

Practical Issues and Considerations

Though findings are very promising, there has been relatively little research on the predictive validity of the ORAS, with only one evaluation of four of the tools and two of the other. Further, studies that met our inclusion criteria did not report inter-rater reliability of the assessments. Finally, all research on the ORAS reviewed in this report had been completed by the study developers; independent replication is needed.

Selected References and Suggested Readings

Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and validation of the Ohio Risk Assessment System: Final report.* Cincinnati, OH: Authors. Retrieved from http://www.uc.edu/ccjr/Reports/ProjectReports/ORAS_Final_Report.pdf

Lowenkamp, C. T., Lemke, R., & Latessa, E. (2008). The development and validation of a pretrial screening tool. *Federal Probation, 72*, 2-9.

Risk Management Systems

Description

The Risk Management Systems (RMS) is an actuarial risk assessment instrument intended for use intended to assess risk for general offending across offender populations (Dow, Jones, & Mott, 2005). The RMS currently contains 67 static and dynamic risk factors; however, when it was validated, the instrument included only 65 items. The assessment is split into four parts: 1) Needs (24 items), 2) Risk (9 items), 3) Mental Health (10 items), and 4) Other-External (24 items). Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, substance abuse problems, mental health problems, and housing. The developers of the RMS describe it as a 5th generation risk assessment instrument due to its exemplar-based approach.

The RMS is administered using a computer-based questionnaire. As such, the assessor is removed from the initial assessment process; individual item responses are statistically analyzed to calculate risk of recidivism. Risk scores for violence and recidivism range from 1.00 (Low) to 2.00 (High), at 0.01 intervals. However, there are no established cut-off scores for risk categories, so the assessor must interpret the subsequent level of risk/supervision required. RMS assessment materials are available through Syscon Justice Systems (www.syscon.net). For information on pricing see www.syscon.net.

U.S. Research Evidence

As of December 2012, predictive validity of RMS assessments had been reported in two U.S. studies; performance ranged from good (Kelly, 2009; later republished in Shaffer et al., 2010) to excellent (Dow et al., 2005). The risk classifications have notably better predictive validity (excellent) compared to total scores (good). Validity is comparable for predicting general offending and violations. RMS assessments appear to have better predictive validity when completed in research studies (excellent) than in the context of 'real world' practice (good); however, risk classifications were used in one study and total scores in the other.

There were no studies of predictive validity conducted in the United States that compared findings across offender sex or racial/ethnic groups. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

Practical Issues and Considerations

In the initial development and validation work, the tool was intended to be used for assessing risk for general offending (Dow et al., 2005), but a later study established the validity of RMS assessments in predicting violations (Kelly, 2009). Overall, further independent research is needed to replicate and establish the generalizability of findings, as well as to determine the validity of different cut-off scores.

Selected References and Suggested Readings

Dow, E., Jones, C., & Mott, J. (2005). An empirical modeling approach to recidivism classification. *Criminal Justice and Behavior, 32*, 223-247.

Kelly, B. (2009). *A validation study of Risk Management Systems* (Master's thesis). Retrieved from UNLV Theses/Dissertations/Professional Papers/Capstones. (Paper 128). http://digitalscholarship.unlv.edu/thesesdissertations/128

Shaffer, D. K., Kelly, B., & Lieberman, J. D. (2010). An exemplar-based approach to risk assessment: Validating the Risk Management Systems instrument. *Criminal Justice Policy Review, 22*, 167-186.

Salient Factor Score

Description

The Salient Factor Score (SFS) is an actuarial risk assessment tool intended to inform decisions regarding whether an offender should be granted parole or not. The SFS is a 2nd generation risk assessment instrument.

There are at least four versions of the SFS, all of which measure static risk factors. Items have been adapted throughout the years to be consistent with research findings. The SFS74 contains nine items and content areas include history of antisocial behavior, circumstances at work or school, substance use problems, and housing. The SFS76 contains seven items and content areas include history of antisocial behavior, circumstances at work or school, and substance use problems. The SFS81 contains six items and content areas include history of antisocial behavior and substance use problems. The SFS98 includes six items and the only content area included is history of antisocial behavior. Unlike the prior versions, the SFS98 also considers whether the offender was older than 41 at the time of the current offense.

SFS assessments are completed through review of official records. Item ratings are summed to arrive at an overall risk score; a *higher* score indicating *lower* risk. These total scores are then used to place offenders within one of four risk categories: very good risk, good risk, fair risk, and poor risk. For further information contact the United States Parole Commission (http://www.justice.gov/uspc).

U.S. Research Evidence

As of December 2012, predictive validity of SFS74, SFS76, and the SFS81 assessments had been examined in 15 U.S. samples. Validity of SFS74 and SFS76 assessments in predicting general offending has ranged from good to excellent. SFS81 assessments also have shown excellent predictive validity across most studies, though the odds ratio was notably low in one evaluation (Howard, 2007). We did not find any evaluations of the predictive validity of SFS98 assessments that met our inclusion criteria.

To date, there have been no U.S. studies comparing predictive validity of the SFS instruments between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity. We also did not find any evaluations of inter-rater reliability that met our inclusion criteria.

Practical Issues and Considerations

Though items are relatively straightforward to code, investigations of inter-rater reliability are needed to establish the consistency of assessments completed by different assessors.

Jurisdiction-specific adaptations include the Connecticut Salient Factor Score.

Selected References and Suggested Readings

Hoffman, P. (1996). Twenty years of operational use of a risk prediction instrument: The United States Parole Commission's Salient Factor Score. *Journal of Criminal Justice, 22*, 477-494.

Hoffman, P. & Adelberg, S. (1980). The Salient Factor Score: A nontechnical overview. *Federal Probation, 44*, 44-52.

Howard, B. (2007). *Examining predictive validity of the Salient Factor Score and HCR-20 among behavior health court clientele: Comparing static and dynamic variables*. (Unpublished doctoral dissertation).

Self-Appraisal Questionnaire

The Self-Appraisal Questionnaire (SAQ) is an actuarial risk assessment instrument to assess risk for general offending among male offenders (Loza, 2005).

The SAQ contains 72 dynamic and static risk factors. Content areas include attitudes, associates or peers, history of antisocial behavior, personality problems, and substance abuse problems. Items are divided across seven subscales. Scores on six subscales are calculated to provide an overall risk score. A seventh anger subscale is not used to assess risk for recidivism. Therefore, of the 72 total items, 67 items are used to predict recidivism. Total scores are used to place offenders in one of four risk categories: low, low-moderate, high-moderate, and high. The SAQ is a 3rd generation risk assessment instrument.

The SAQ is a true/false self-report questionnaire. Five items can be used to assess the validity of an offender's answers by comparing them against official records. The SAQ takes approximately 15 minutes to administer and five minutes to hand-score. The assessor does not need formal training, but scoring must be overseen by someone who has post-secondary training in psychological assessment. The SAQ can be purchased from Multi-Health Systems Inc. at www.mhs.com. Costs include those associated with the manual and assessment forms. For further information on pricing, see www.mhs.com.

U.S. Research Evidence

Two studies have evaluated the predictive validity of the SAQ in U.S. samples. These studies used low, moderate, and high risk categories rather than the four categories suggested by the assessment developer. Mitchell and Mackenzie (2006) found poor validity of the SAQ

assessments in predicting re-arrest and failed to find differences in total scores between recidivists and non-recidivists. In contrast, using a longer follow-up period and a larger sample, Mitchell, Caudy and Mackenzie (2012) found that SAQ assessments predicted time to first reconviction, though the effect size was small.

As of December 2012, there had been no studies comparing predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice, by offender sex, or race/ethnicity that met our inclusion criteria. Because the SAQ is a self-report instrument, inter-reliability is not relevant.

There have been many evaluations of the SAQ in Canada (e.g., Kroner & Loza, 2001; Loza & Loza-Fanous, 2000; Loza et al., 2005), but none have compared the predictive validity between total scores and risk classifications, research and practice contexts, by offender sex, or race/ethnicity.

Practical Issues and Considerations

The SAQ requires a 5th grade reading level. Prior studies of the validity of SAQ assessments in predicting violent outcomes, including institutional violence and violent recidivism (e.g., Campbell, French & Gendreau, 2009), as well as violent and non-violent recidivism in Canadian samples (e.g., Loza, MacTavish, & Loza-Fanous, 2007) have shown more promising results than those reported herein vis-à-vis validity in predicting non-violent offending in U.S. samples.

Selected References and Suggested Readings

Kroner, D., & Loza, W. (2001). Evidence for the efficacy of self-report in predicting violent and nonviolent criminal recidivism. *Journal of Interpersonal Violence, 16,* 168-177.

Loza, W., & Loza-Fanous, A. (2000). Predictive validity of the Self-Appraisal Questionnaire (SAQ): A tool for assessing violent and nonviolent release failures. *Journal of Interpersonal Violence, 15*, 1183-1191.

Loza, W. (2005). *The Self-Appraisal Questionnaire (SAQ): A tool for assessing violent and non-violent recidivism.* Toronto: Mental Health Systems.

Loza, W., Neo, L. H., Shahinfar, A., & Loza-Fanous, A. (2005). Cross-validation of the Self-Appraisal Questionnaire: A tool for assessing violent and nonviolent recidivism with female offenders. *International Journal of Offender Therapy & Comparative Criminology, 49,* 547-560.

Mitchell, O., Caudy, M., & Mackenzie, D. (2012). A reanalysis of the Self-Appraisal Questionnaire: Psychometric properties and predictive validity. *International Journal of Offender Therapy and Comparative Criminology 20,* 1-15.

Mitchell, O., & Mackenzie, D. (2006). Disconfirmation of the predictive validity of the Self-Appraisal Questionnaire in a sample of high-risk drug offenders. *Criminal Justice and Behavior 33,* 449-466.

Service Planning Instruments

Description

The Service Planning Instrument (SPIn) is an actuarial risk assessment tool intended to assess risk for offending and to identify service needs of male offenders. The SPIn-W was developed for use with female offenders.

Both the SPIn and SPIn-W are self-report, computer-based instruments. The SPIn includes 90 static, dynamic, risk, and protective factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at school or work, substance use problems, mental health problems, and housing. The SPIn-W includes 100 static, dynamic, risk, and protective factors. Content areas include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing. The SPIn and SPIn-W are 4th generation risk assessment instruments.

For both instruments, software is used to calculate an offender's risk score which is presented graphically and narratively. The assessor must compare responses on static items to the offender's official records. Assessors are required to attend a two-day training session. Additional 2-day training program to help administrators better prepare for the case planning process, as well as data workshops, refresher courses, technical support, and quality assurance also are available. The SPIn and SPIn-W can be purchased from Orbis Partners Inc. (www.orbispartners.com). For information on pricing, see www.orbispartners.com.

U.S. Research Evidence

As of December 2012, there were no published studies assessing predictive validity of SPIn assessments in U.S. samples. Two studies have evaluated predictive validity of the SPIn-W assessments; performance ranged from poor to excellent.

There were no comparisons of predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice contexts, by outcome or by offender race/ethnicity that met our inclusion criteria. We also did not identify any U.S. evaluations of inter-rater reliability that met these criteria.

Practical Issues and Considerations

Current evidence regarding the predictive validity of SPIn-W assessments is both limited and mixed. More research is needed.

Selected References and Suggested Readings

Meaden, C. (2012). *The utility of the Level of Service Inventory-Revised versus the Service Planning Instrument for Women in predicting program completion in female offenders*.

(Unpublished Master's thesis). Retrieved from Central Connecticut State University Theses, Dissertations, and Special Projects.

Millson, B., Robinson, D., & Van Dieten, M. (2010). *Women Offender Case Management Model: An outcome evaluation.* Washington, DC: U.S. Department of Justice, National Institute of Corrections. Retrieved from: http://www.cjinvolvedwomen.org/sites/all/documents/Women%20Offender%20Case%20Management%20Model.pdf

Static Risk and Offender Needs Guide

The Static Risk and Offender Needs Guide (STRONG) is an actuarial risk assessment instrument intended to assess risk for general offending across offender populations (Barnoski & Drake, 2007).

The STRONG consists of three parts: 1) the Static Risk Assessment which contains 26 static risk factors; 2) the Offender Needs Assessment which contains 70 dynamic risk and protective factors; and 3) the Offender Supervision Plan, which is auto-populated based on the results of the Offender Needs Assessment. Content areas assessed in the Static Risk Assessment include history of antisocial behavior and substance use problems. Items scores are used to create three separate scores: Felony Risk Score; Non-Violent Felony Risk Score (high property risk/high drug risk); and Violent Felony Risk Score. These three scores are used to classify offenders in one of five categories: high risk violent; high risk property; high risk drug; moderate risk; and low risk. Content areas assessed in the Offender Needs Assessment include attitudes, associates or peers, personality problems, relationships, circumstances at work or school, substance use problems, mental health problems, and housing. Ratings on items included in the Offender Needs Assessment are not used to inform risk assessments, but instead guide the development of interventions designed to reduce risk of future criminal justice involvement. As such, the STRONG is a 4th generation risk assessment instrument.

STRONG assessments are completed by assessors using a web-based interface. Assessors must complete an initial training program as well as routine booster training sessions. The STRONG was developed by Assessments.com in collaboration with the Washington Department of Corrections. A very similar version can be purchased for use in other jurisdictions through www.assessments.com.

U.S. Research Evidence

Only one study that met our inclusion criteria has evaluated the predictive validity of STRONG assessments; assessments demonstrated excellent predictive validity overall as well as for male and female offenders separately (Barnoski & Drake, 2007). There were no U.S. studies comparing predictive validity as a function of offender race/ethnicity, type of recidivism outcome or between assessments completed in the context of research versus practice. We also did not find any evaluations of inter-rater reliability that met inclusion criteria.

Practical Issues and Considerations

Though findings are promising, predictive validity of STRONG assessments has only been evaluated in one study conducted by the instrument developer; independent replication is needed.

Selected References and Suggested Readings

Barnoski, R., & Drake, E. K. (2007). *Washington's Offender Accountability Act: Department of Corrections' static risk instrument.* Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://www.wsipp.wa.gov/rptfiles/07-03-1201R.pdf

Wisconsin Risk and Needs Scales

Description

The Wisconsin Risk and Needs scales (WRN) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations. A revised version (WRN-R) was designed specifically for use with probationers and parolees (Eisenberg, Bryl, & Fabelo, 2009). Both the WRN and WRN-R are 4th generation risk assessment instruments.

The WRN contains 53 static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at work or school, substance use problems, and mental health problems. Individual item scores are scored and summed for a total risk score ranging from 0 to 52. The total score is used to place the offender in a risk category based on predetermined cut-offs: Low = 0-7; Medium = 8-14; and High = 15+.

The WRN-R retained 52 of the WRN's items and covers the same content areas. The weights of the different factors have been revised from the original WRN based on the results of a validation study, and the revised total risk score has a range of 0 to 25. The total score is used to estimate risk level based on new cut-offs: Low = 0-8; Medium = 9-14; and High = 15+.

WRN assessments are completed using information obtained through interview. The WRN is non-proprietary and available through Justice Systems Assessment & Training (http://www.j-satresources.com/Toolkit/Adult/adf6e846-f4dc-4b1e-b7b1-2ff28551ce85).

U.S. Research Evidence

Predictive validity of the WRN assessments have ranged from fair (Eisenberg et al., 2009) to excellent (Connolly, 2003). WRN assessments appear to perform better when predictive violations (excellent) than general offending (good). Our comparisons between predictive validity of assessments completed in research versus practice failed to identify any differences. As of December 2012, no U.S. studies compared predictive validity between WRN total scores and risk classifications, by offender sex, or race/ethnicity. We also did not identify any U.S. evaluations of inter-rater reliability that met our inclusion criteria.

As of December 2012, predictive validity of WRN-R assessments had been evaluated in one U.S. study; assessments demonstrated good predictive validity. To date, there have been no studies comparing predictive validity in U.S. samples between WRN-R total scores and risk classifications, assessments completed in research and practice contexts, by recidivism outcome, offender race/ethnicity, or sex that met our inclusion criteria. We also did not identify any U.S. evaluations of inter-rater reliability of WRN-R assessments.

Practical Issues and Considerations

A high percentage of offenders are classified as high risk using the WRN due to the heavy weight given to convictions for an assaultive offense in the past five years. There is concern that such over-classification is "counter to the goal of risk classification: to differentiate the population by risk and allocate resources accordingly" (Eisenberg et al., 2009, p. iv).

In 2004, a new, automated assessment and case management system called the Correctional Assessment and Intervention System (CAIS) was developed based upon the WRN and the Client Management Classification tools (Baird, Heinz, & Bemus, 1979). This CAIS is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations, as well as to be used in the development of case management plans. Its predictive validity has not yet been evaluated.

Selected References and Suggested Readings

Baird, C., Heinz, R., & Bemus, B. (1979). *The Wisconsin Case Classification/Staff Deployment Project.* Madison, WI: Wisconsin Department of Corrections.

Eisenberg, M., Bryl, J., & Fabelo, T. (2009). *Validation of the Wisconsin Department of Corrections risk assessment instrument.* New York: Council of State Governments Justice Center. Retrieved from http://www.wi-doc.com/PDF_Files/WIRiskValidation_August%202009.pdf

OTHER TYPES OF INSTRUMENTS USED TO ASSESS RECIDIVISM RISK

Violence Risk Assessment Instruments

Violence risk assessment instruments, such as the Historical-Clinical-Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) and Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 2006), are intended to assess risk of future violence specifically, but also are frequently used to assess risk of (non-violent) recidivism.

HCR-20

The HCR-20 is a structured professional judgment scheme comprised of 20 static and dynamic items that assess historical risk factors, clinical risk factors, and risk management factors. The individual item ratings are used to inform a final professional judgment of low, moderate, or high risk. Only one study has evaluated the validity of HCR-20 assessments in predicting recidivism in a U.S. sample (Barber-Rioja, Dewey, Kopelovich, & Kucharski, 2012). Overall, the assessment total score was found to have excellent validity in predicting both general offending and violations. The HCR-20 has been widely validated outside of the U.S. (see http://kdouglas.files.wordpress.com/2007/10/hcr-20-annotated-biblio-sept-2010.pdf).

VRAG

The VRAG is an actuarial instrument designed for use with previously violent, mentally disordered offenders. It consists of 12 items that gather information on static and dynamic risk factors. Individual item responses are weighted and summed for a total score, which is then used to estimate level of risk based on an actuarial table. The predictive validity of VRAG assessments for both general offending and violations also has been evaluated in only one U.S. sample (Hastings et al., 2011). Validity in predicting general offending ranged from good to excellent for male offenders, and fair to good for female offenders. Validity in predicting violations ranged from fair to good for male offender and poor to fair for female offenders. Like the HCR-20, much research completed outside of the U.S. has examined the validity of VRAG assessments. For more information, visit http://www.mhcp.on.ca/

References and Suggested Readings

Barber-Rioja, V., Dewey, L., Kopelovich, S., & Kucharski, L. T. (2012). The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Criminal Justice and Behavior, 39,* 475-492.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). The prediction of violence and antisocial behaviour: A systematic review and meta-analysis of the utility of risk assessment instruments in 73 samples involving 24,827 individuals. *British Medical Journal, 345,* e4692.

Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Appraisal Guide scores with male and female jail inmates. *Psychological Assessment, 23,* 174-183.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence* (version 2). Burnaby, BC: Simon Fraser University, Mental Health, Law, and Policy Institute.

Personality Assessment Instruments

Personality assessment instruments, such as the Psychopathy Checklist-Revised (PCL-R; Hare, 2003), the Psychopathy Checklist: Screening Version (PCL:SV; Hart, Cox, & Hare, 1995), and the Personality Assessment Instrument (PAI; Morey, 1991), evaluate personality constructs that correlate with criminal offending (for a meta-analytic review see Singh & Fazel, 2010).

PCL Instruments

The PCL-R is a 20-item actuarial assessment that can be used to diagnosis psychopathy, a form of antisocial personality disorder characterized by a persistent pattern of severe and refractory callous-unemotionality. Individual items are scored through file review and semi-structured interview, then summed for total score ranging from 0 to 40 (where 30+ indicates the presence of psychopathy). The PCL:SV is a shorter, 12-item version. Again, individual item ratings are scored and summed, with a cutoff score of 18 typically used for classification of psychopathy. Research demonstrates excellent correspondence between the two measures in correctional samples (Guy & Douglas, 2006). Validity of PCL-R and PCL:SV assessments in predicting recidivism has been evaluated extensively in the U.S., with performance ranging from poor to good (e.g., Gonsalves, Scalora, & Huss, 2009; Salekin, Rogers, Ustad, & Sewell, 1998; Walters & Duncan, 2005). For more information on the PLC-R and PCL:SV, see http://www.hare.org/scales/.

PAI

The PAI contains 344 self-report items that are divided into 22 validity, clinical, treatment consideration, and interpersonal scales. Individual item responses within the scales are hand scored and assessed in conjunction with interpretive guidelines included in the professional manual (Morey, 2007). In U.S. studies assessing the predictive validity of the PAI, the assessment scale scores had fair to good validity in predicting general offending (e.g., Barber-Rioja et al., 2012; Walters, 2009; Walters & Duncan, 2005). For an overview and bibliography, see http://www4.parinc.com/Products/Product.aspx?ProductID=PAI.

Other Personality Assessment Instruments

Other instruments including the California Psychological Inventory: Socialization Scale (CPI:SO), Lifestyle Criminality Screening Form (LCSF), Minnesota Multiphasic Personality Inventory (MMPI), Neuroticism, Openness to Exposure Personality Inventory-Revised (NEO-PI-R), and the Peterson, Quay, and Cameron Psychopathy Scale (PQC) can produce valid assessments of recidivism risk, though performance varies widely (see Walters, 2003, 2006).

References and Suggested Readings

Barber-Rioja, V., Dewey, L., Kopelovich, S., & Kucharski, L. T. (2012). The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Criminal Justice and Behavior, 39,* 475-492.

Hare, R. D. (2003). *Hare Psychopathy Checklist-Revised (PCL-R): Second edition, technical manual.* Toronto, ON, Canada: Multi-Health Systems.

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version* (1st ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Gonsalves, V. M., Scalora, M. J., & Huss, M. T. (2009). Prediction of recidivism using the Psychopathy Checklist-Revised and the Psychological Inventory of Criminal Thinking Styles within a forensic sample. *Criminal Justice and Behavior, 36,* 741-756.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Salekin, R. T., Rogers, R., Ustad, K. L., & Sewell, K. W. (1998). Psychopathy and recidivism among female inmates. *Law and Human Behavior, 22,* 109-128.

Walters, G. D. (2009). The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening Version as incrementally valid predictors of recidivism. *Law and Human Behavior, 33,* 497-505.

Walters, G. D. & Duncan, S. A. (2005). Use of the PCL-R and PAI to predict release outcome in inmates undergoing forensic evaluation. *Journal of Forensic Psychiatry and Psychology, 16,* 459-476.

Criminal Thinking Questionnaires

Criminal thinking questionnaires, such as the Psychological Inventory of Criminal Thinking Styles (PICTS; Walters, 1995) and the Texas Christian University Criminal Thinking Scales (TCU CTS; Knight, Simpson, & Morey, 2002), are designed to identify attitudes and thought patterns associated with criminal behavior.

PICTS

The PICTS is an 80-item, self-report measure composed of eight thinking pattern scales, two validity scales, four factor scales, two composite scales, and a General Criminal Thinking (GCT) scale. The validity of PICTS scores in predicting general offending has been evaluated in a number of U.S. studies with mixed findings. Performance of the GCT scale scores ranges from poor to good (e.g., Walters, 2009a, 2009b, 2011); however, other research suggests the eight thinking pattern scales have poor validity (Gonsalves, Scalora, & Huss, 2009).

TCU CTS

The TCU CTS is an actuarial, self-report instrument designed to measure criminal thinking. The instrument contains 37 items distributed across six thinking pattern scales: Entitlement, Justification, Power Orientation, Cold Heartedness, Criminal Rationalization, and Personal Irresponsibility. In one U.S. study, the six thinking pattern scale scores had poor validity in predicting both general offending and violations (Taxman, Rhodes & Dumenci, 2011). More information and a copy of the TCU CTS assessment materials are available from http://www.ibr.tcu.edu/pubs/datacoll/cjtrt.html.

References and Suggested Readings

Gonsalves, V. M., Scalora, M. J., & Huss, M. T. (2009). Prediction of recidivism using the Psychopathy Checklist-Revised and the Psychological Inventory of Criminal Thinking Styles within a forensic sample. *Criminal Justice and Behavior, 36,* 741-756.

Knight, K., Simpson, D. D., & Morey, J. T. (2002). *TCU-NIC Cooperative Agreement: Final report.* Fort Worth, TX: Texas Christian University, Institute of Behavioral Research.

Taxman, F. S., Rhodes, A. G., & Dumenci, L. (2011). Construct and predictive validity of Criminal Thinking Scales. *Criminal Justice and Behavior, 38,* 174-187.

Walters, G. D. (1995). The Psychological Inventory of Criminal Thinking Styles, Part I: Reliability and preliminary validity. *Criminal Justice and Behavior, 22,* 307-325.

Walters, G. D. (2009a). Effect of a longer versus shorter test-release interval on recidivism prediction with the Psychological Inventory of Criminal Thinking Styles (PICTS). *International Journal of Offender Therapy and Comparative Criminology, 53,* 665-678.

Walters, G. D. (2009b). The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening Version as incrementally valid predictors of recidivism. *Law and Human Behavior, 33,* 497-505.

Walters, G. D. (2011). Predicting recidivism with the Psychological Inventory of Criminal Thinking Styles and Level of Service Inventory-Revised: Screening Version. *Law and Human Behavior, 35,* 211-220.

CONCLUSION

Summary of Findings

Our review of validation studies conducted in the United States did not identify one instrument that systematically produced more accurate assessments than the others. However, performance within and between instruments varied considerably depending on the assessment sample, circumstances, and recidivism outcome.

Overall, there were very few U.S. evaluations examining the predictive validity of assessments completed using instruments commonly used in U.S. correctional agencies. In most cases, validity of assessments completed using any given instrument had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument. Moreover, only two of the 53 studies included in this review reported evaluations of inter-rater reliability. (We return to these two points later.)

Our selection criteria and, specifically, our focus on studies of predictive validity conducted in the United States resulted in the exclusion of some prominent and promising instruments, such as the LS/CMI or the Women's Risk/Need Assessment. Similarly, none of the reviewed studies examined the predictive validity of structured professional judgment, as opposed to actuarial, instruments, though we know of at least a few that are being used for the purposes of assessing recidivism risk (e.g., the Short-Term Assessment of Risk and Treatability, START, see Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012). Importantly, findings of the current review are not intended to suggest that these instruments do not produce reliable and valid assessments of recidivism risk and should not necessarily preclude their use in practice. Instead, we are simply asserting that they have yet to be evaluated as such in the United States. Indeed, decision makers interested in any risk assessment instrument should balance considerations of the empirical evidence, but also the practical issues we review in the following section.

Finally, risk classifications (e.g., identification of offenders as low, moderate, or high risk) generally outperformed total scores, yet total scores were evaluated much more frequently. This finding is consistent with prior research (e.g., Desmarais et al., 2012) and emphasizes the importance of using the instruments as they were designed to be used.

Selecting a Recidivism Risk Assessment Instrument

When deciding which recidivism risk assessment instrument to implement in practice, we recommend reviewing the empirical evidence, as well as answering the following questions:

What is your outcome of interest?

Our review revealed that some instruments performed better in predicting particular recidivism outcomes than others. Specifically, the SFS instruments performed particularly well in predicting general offending *including* violations, whereas the ORAS-CST, ORAS-CSST, PCRA, and

STRONG were excellent predictors of offenses *excluding* violations. WRN assessments stood out as the best predictors of violations alone.

What is your population?

Some instruments were developed to assess for specific populations; for example, the SFS instruments are specifically designed for use with parolees. Also, some instruments appear to perform better for some subgroups of offenders than others. The LSI instruments, for instance, produced assessments with only fair validity for female offenders, though predictive validity was generally good for male offenders. Other instruments, such as the COMPAS, ORAS and STRONG, produced assessments with good validity for both male and female offenders.

What resources are required to complete the assessment?

Answering this question includes considering characteristics of both the risk assessment tool as well as the setting; for instance, the information necessary to complete the assessment and whether this information is available. Some instruments, such as the IORNS, are completed based solely on offender self-report; other instruments, such as the PCRA and COMPAS, combine information derived from a variety of sources, including self-report, interview, and review of official records. Similarly, the time required to complete a risk assessment will depend not only on the nature and amount of information required, but also the number of items included. We found that the number of items varied broadly across instruments from four items (ORAS-CSST) to 130 items (IORNS). Decision makers should consider whether staff have the time and information required to complete the assessments. Other resource considerations include staff training and backgrounds. Some instruments, such as the PCRA, require that assessors complete training courses and are certified prior to implementation. Others, such as the LSI family of instruments, require that assessors be supervised by professionals with specific degrees and/or credentials. Last, but certainly not least, decision makers should consider the costs associated with implementing any given risk assessment tool. Costs may include those associated with purchasing materials and staff training, among others, and they may be fixed, one-time costs or costs that will continue to be incurred over time. Long-term sustainability of implementation will hinge, in part, on a *realistic* appraisal of the match between the available and required resources.

Additional Considerations

In addition to identifying the instrument best-suited to an agency's specific needs and constraints, there are additional issues to consider during the process of selecting and implementing a recidivism risk assessment tool.

First, caution is warranted when attempting to generalize the findings of research studies to the use of risk assessment instruments in practice. In research contexts, risk assessments are routinely conducted by graduate students, who may have more or less training than those who will be conducting the risk assessments in practice. Assessors in research studies also may be given more time and resources to complete risk assessments and may receive ongoing

supervision in the specific risk assessment protocol; these luxuries typically are not afforded to professionals in practice settings.

Second, there have been very few evaluations of predictive validity within specific offender subgroups. Indeed, only a handful of studies included in this review compared validity depending on offender sex or race/ethnicity and none examined predictive validity across psychiatric diagnostic categories. As such, there is insufficient evidence to conclude that assessments perform comparably or are equally applicable to specific offender subgroups. As described earlier, actuarial instruments estimate risk of recidivism through comparison of a given offender's total score against the recidivism rates of offenders with the same (or a similar) score in the construction sample. Race/ethnicity and sex are important factors associated with recidivism that may not be accounted for in these actuarial models. There is considerable evidence to suggest that race/ethnicity and sex are potentially important sources of assessment bias (Holtfreter & Cupp, 2007; Leistico, Salekin, DeCoster, & Rogers, 2008).

Third, allegiance, which occurs when at least one developer of the risk assessment instrument is an author on a study investigating that instrument's predictive validity, was present for many of the articles included in this review. Strong effects of allegiance on evaluations of assessment and treatment approaches, including risk assessment, have been found in many fields. In the violence risk assessment literature, a recent meta-analysis demonstrated the impact of allegiance on the predictive validity of three commonly used actuarial instruments (Blair, Marcus, & Boccaccini, 2008). Performance of the instruments was significantly better in studies conducted by the tool authors than in studies conducted by independent researchers. We were unable to test for allegiance effects due to the relatively small number of studies per instrument. Though the reasons for allegiance effects are unclear (e.g., bias, fidelity, see Harris & Rice, 2010), there is a critical need for independent evaluation of the predictive validity of risk assessments completed using the instruments included in this review.

Fourth, most studies included in this review reported statistics that speak to whether recidivists generally received higher risk estimates than did non-recidivists (known as *discrimination*). Very few studies reported statistics that speak to whether those offenders who were identified as high risk for recidivism went on to recidivate during follow-up and whether those offenders who were identified as low risk did not (known as *calibration*). This is not unique to the studies included in the current review; a recent review found that calibration estimates were reported in less a fourth of violence risk assessment studies (see Singh, Desmarais & Van Dorn, 2013). Discrimination and calibration are two sides of the same coin – both representing important qualities of an instrument's predictive validity – but address different issues (Singh, 2013).

Fifth, there was an almost complete lack of information regarding the inter-rater reliability of available recidivism risk assessment instruments. With the exception of LSI-R and LSI-R:SV, we do not have any information regarding whether assessments completed using the instruments reviewed in this report are consistent across assessors. This is not trivial; reliability has been referred to as "the most basic requirement for a risk assessment instrument" (Douglas, Nicholson, & Skeem, 2011, p. 333). Indeed, an assessment *must* be reliable in order for it to be valid (though the reverse is not true). Inter-rater reliability is relevant to any assessment in which

an assessor must rate or code items as part of the process; thus, inter-rater reliability should be examined for all instruments except those completed exclusively through offender self-report.

Sixth and finally, there have been few evaluations of the impact of implementing a risk assessment tool on recidivism rates. Though many of the instruments included in the present review have acceptable levels of predictive validity, the goal of risk assessment is not simply to predict, but, ultimately, to *reduce* recidivism. Achieving this goal will necessitate the following:

1. The risk assessment tool *must* be implemented in a sustainable fashion with fidelity. It is not as simple as deciding on a tool and applying it in practice. Successful implementation of a risk assessment tool involves completing a series of steps, from preparation to training and pilot testing to full implementation. This multi-step process requires ongoing supervision to ensure sustainability, including regular evaluations of fidelity and booster training for staff on a semi-annual basis (see Vincent, Guy & Grisso, 2012 for a guide to implementation).
2. Findings of the risk assessment *must* be communicated accurately and completely. Indeed, "Improper risk communication can render a risk assessment that was otherwise well-conducted completely useless or even worse, if it gives consumers the wrong impression." (Heilbrun, Dvoskin, Hart & McNiel, 1999, p. 94).
3. Information derived during the risk assessment process *must* be used to guide risk management and rehabilitation efforts, with particular attention to the steps described by the RNR model; specifically, assess offenders' risk of recidivism, with more restrictive and intensive efforts focused on high-risk offenders; match treatment and rehabilitation efforts to offenders' individual criminogenic needs (as identified in the risk assessment process) and deliver them in a way that is responsive to their individual learning style, motivation, personality and strengths. This will require regular review of staff performance. How performance, as well as fidelity, will be measured should be detailed in a comprehensive program evaluation plan established *prior to* implementation.

BIBLIOGRAPHY

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgement project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34,* 341-382.

Anderson, D. A. (1999). The aggregate burden of crime. *Journal of Law and Economics, 42,* 611-642.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*, 7-27.

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15,* 346–360.

Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation* (User Report 2007–06). Ottawa, Ontario: Public Safety Canada.

Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123*, 123-142.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation, 29,* 860-864.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical And Experimental Neuropsychology, 23,* 695-700.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict in patient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24,* 685-700.

Desmarais, S. L., Van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences & the Law, 30,* 448–469.

Douglas, K. S., Otto, R., Desmarais, S. L., & Borum, R. (in press). Clinical forensic psychology. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, volume 2: Research methods in psychology.* Hoboken, NJ: John Wiley & Sons.

Douglas, K. S., Skeem, J. L., & Nicholson, E. (2011). Research methods in violence risk assessment. In B. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 325-346). Hoboken, NJ: John Wiley & Sons, Inc.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). The prediction of violence and antisocial behaviour: A systematic review and meta-analysis of the utility of risk assessment instruments in 73 samples involving 24,827 individuals. *British Medical Journal*, *345*, e4692.

Federal Bureau of Investigation (FBI). (2012). *Crime in the United States, 2011.* Washington, D.C.: Authors.

Gendreau, P., Goggin, C., & Little, T. (1996). *Predicting adult offender recidivism: What works!* (Cat. No. JS4-1/1996-7E). Ottawa, ON: Public Works and Government Services Canada.

Glaze, L. E. (2011). *Correctional population in the United States, 2011.* Washington, D.C.: Bureau of Justice Statistics.

Hanson, R. K., & Harris, A. J. R. (2000). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 13,* 105-122.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (2010). Allegiance or fidelity? A clarifying reply. *Clinical Psychology: Science and Practice, 17,* 82-89.

Hart, S. D., Michie, C., & Cooke, D. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *The British Journal Of Psychiatry, 190*(Suppl 49), s60-s65.

Hart, S. D., Webster, C. D., & Douglas, K. S. (2001). Risk management using the HCR-20: A general overview of focusing on historical factors. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves, & J. R. P. Ogloff (Eds.), *HCR-20 violence risk management companion guide* (pp. 27-40). Burnaby, Canada/Tampa, FL: Simon Fraser University, Mental Health, Law & Policy Institute/University of South Florida, Dept. of Mental Health Law & Policy.

Heilbrun, K., Dvoskin, J., Hart, S., & McNiel, D. (1999). Violence risk communication: Implications for research, policy, and practice. *Health, Risk & Society, 1,* 91-105,

Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal Of Contemporary Criminal Justice, 23*, 363-382.

Kyckelhahn, T. (2012). *Justice Expenditure And Employment Extracts, 2007 - Revised.* Washington, D.C.: Bureau of Justice Statistics.

Langan, P. A. & Levin, D. J. (2002). *Recidivism of prisoners released in 1994* (NCJ 193427). Washington, D.C.: Bureau of Justice Statistics.

Leistico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28-45.

Liptak, A. (2008, April 23). Inmate count in U.S. dwarfs other nations. *The New York Times.* Retrieved from http://www.nytimes.com.

Lowenkamp, C. T., Pealer, J., Smith, P., & Latessa, E. J. (2006). Adhering to the risk and need principles: Does it matter for supervision-based programs? *Federal Probation, 70,* 3-8.

Mamalian, C. A. (2011). *State of the science of pretrial risk assessment.* Washington, D.C.: Bureau of Justice Assistance.

Pew Center on the States (2009). *One in 31: The long reach of American corrections.* Washington, DC: The Pew Charitable Trusts.

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*, 615-620.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences and the Law, 31,* 8-22.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in studies of risk assessment instruments: A second-order systematic review. *Behavioral Sciences & the Law, 31,* 55–73.

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice & Behavior*, *37*, 965-988.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20,* 38-42.

Smith, P., Cullen, F., & Latessa, E. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183-208.

Vincent, G. M., Guy, L. M., & Grisso, T. (2012). *Risk assessment in juvenile justice: A guidebook for implementation.* John D. And Catherine T. MacArthur Foundation. Available at: http://modelsforchange.net/publications/346

Walmsley, R. (2010). *World prison population list, 9th edition.* London: International Centre for Prison Studies.

Walters, G. D. (2003). Outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences & the Law, 21*, 89-102.

Walters, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice & Behavior, 33*, 279-304.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (in press). Incremental validity of dynamic factors in the assessment of violence risk. *Law and Human Behavior.*

APPENDIX A

List of Jurisdiction-Specific Risk Assessment Instruments

1. Alabama Risk and Needs Assessment
2. Allegheny County Risk Assessment
3. Arizona Risk Assessment Suite
4. Arkansas Post-Prison Board Transfer Risk Assessment
5. California Parole Violation Decision Making Instrument
6. California Static Risk Assessment
7. Colorado Actuarial Risk Assessment Scale
8. Connecticut Salient Factor Score
9. Delaware Parole Board Risk Assessment
10. Georgia Board of Pardons and Parole's Field Log of Interaction Data
11. Georgia Parole Behavior Response and Adjustment Guide
12. Georgia Parole Decisions Guidelines Grid System
13. Georgia Department of Corrections Offender Tracking Information System
14. Hawaii Risk and Needs Assessment
15. Illinois Risk Assessment Instrument
16. Illinois Risks, Assets and Needs Assessment Tool
17. Indiana Risk Assessment System
18. Kentucky Pretrial Risk Assessment Instrument
19. Kentucky Parole Guidelines Risk Assessment Instrument
20. Iowa Board of Parole Risk Assessment
21. Louisiana Risk Needs Assessment
22. Maryland Public Safety Risk Assessment
23. Michigan Parole Guidelines Score Sheet
24. Mississippi Parole Risk Instrument
25. Missouri Sentencing Assessment Risk Instrument
26. Missouri Parole Board Salient Factor Guidelines
27. Montana Risk Assessment Instrument
28. Nebraska Criminal History Assessment instrument
29. Nevada Parole Risk Assessment

30. New Mexico Risk and Needs Assessment
31. North Carolina Risk Needs Assessment
32. Oregon Criminal History/Risk Assessment
33. Public Safety Checklist for Oregon
34. Orange County Pretrial Risk Assessment
35. Rhode Island Parole Risk Assessment
36. South Carolina Parole Risk Assessment Instrument
37. South Dakota Initial Community Risk/Needs Assessment
38. State of Hawaii LSI-R Proxy
39. Tennessee Offender Risk Assessment/Needs Assessment
40. Tennessee Parole Grant Prediction Scale and Guidelines
41. Texas Parole Risk Assessment Instrument
42. Utah Criminal History Assessment
43. Vermont Parole Board Risk Assessment
44. Virginia Pretrial Risk Assessment Instrument
45. Virginia Risk Assessment Tool
46. Washington Risk Level Classification
47. West Virginia Parole Board Assessment

APPENDIX B

Glossary of Terms

Actuarial Risk Assessment

Mechanical approach to risk assessment in which offenders are scored on a series of items statistically associated with recidivism risk in the sample of offenders upon whom the instrument was developed. The total score is cross-referenced with a statistical table that translates the score into an estimate of recidivism risk during a specified timeframe.

Area Under the Curve (AUC)

Performance indicator measuring the probability that a randomly selected offender who recidivated during follow-up would have received a higher risk classification using a given risk assessment approach than a randomly selected offender who did not recidivate during follow-up.

Cohen's d

Performance indicator measuring the standardized mean difference between the estimated level of risk or total score of offenders who did and did not recidivate during follow-up.

Dynamic Factor

Changeable characteristics (e.g., substance abuse) that establish a relative level of risk and help inform intervention; they can be either relatively *stable*, changing relatively slowly over time (e.g., antisocial cognition) or *acute,* changing more quickly over time (e.g., mood state).

Kappa (k)

Measure of inter-rater reliability representing the percentage of categorizations (e.g., low, moderate or high risk) upon which multiple assessors agreed, statistically corrected for chance.

Intra-Class Correlation Coefficient (ICC)

Measure of inter-rater reliability representing the strength of agreement between multiple assessors on *continuous* variables (e.g., total scores), statistically corrected for chance.

Meta-analysis

Systematic review that includes a quantitative synthesis of the findings of *primary research*.

Observed Agreement

Measure of inter-rater reliability representing the percentage of categorizations (e.g., low, moderate or high risk) upon which multiple assessors agreed.

Odds ratio (OR)

Performance indicator measuring the odds of the risk estimate in an offender who recidivates during follow-up being one higher than the risk estimate of an offender who does not recidivate.

Parole

Conditional release of a prisoner before the expiration of his or her sentence subject to conditions supervised by a designated parole officer.

Performance Indicator

Statistical measure of predictive validity.

Point-Biserial Correlation Coefficient (r_{pb})

Performance indicator measuring the direction and strength of the association between a *continuous predictor* (e.g., total score) and a *dichotomous outcome* (e.g., recidivating vs. not).

Primary Research

Collection of new data that does not already exist.

Probation

Release of an offender from detention or sentence served in the community in lieu of detention, subject to conditions supervised by a probation officer.

Protective Factor

Characteristic of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or situation (e.g., living situation) that is associated with a decrease in the likelihood of offending.

Recidivism

Relapse into criminal behavior by an individual who has previously been convicted of one or more offenses.

Risk Assessment

Process of estimating the likelihood an offender will recidivate to identify those at higher risk and in greater need of intervention. Also may assist in the identification of treatment targets and the development of risk management and treatment plans.

Risk Assessment Instrument

Instrument composed of empirically- or theoretically-based risk and/or protective factors used to aid in the assessment of recidivism risk.

Risk Factor

Characteristic of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or situation (e.g., living situation) that is associated with an increase in the likelihood of offending.

Somer's d

Performance indicator measuring the direction and strength of the association between an *ordinal predictor* (e.g., estimate of risk as low, moderate or high) and a *dichotomous outcome* (e.g., recidivating vs. not).

Structured Professional Judgment

Structured approach to risk assessment focused on creating individualized and coherent risk formulations and comprehensive risk management plans. Assessors estimate risk through consideration of a set number of factors that are empirically and theoretically associated with the outcome of interest. Total scores are not used to make the final judgments of risk. Instead, assessors consider the relevance of each item to the individual offender, as well as whether there are any case specific factors not explicitly included in the list.

Static Factor

A historical or otherwise unchangeable characteristics (e.g., history of antisocial behavior) that help establish absolute level of risk.

Systematic Review

A process in which the empirical literature from multiple primary studies on a particular topic meeting pre-determined inclusion and exclusion criteria is descriptively analyzed.

Technical Violation

A breach of the conditions of parole or probation.

Unstructured Risk Assessment

A subjective assessment of recidivism risk based on the assessor's intuition, knowledge of theory, and professional experience.