# 1

# Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*

*Sarah L. Desmarais, Kiersten L. Johnson, and Jay P. Singh*



The rates of correctional supervision and incarceration in the United States are staggering. To demonstrate, almost seven million people—or one in 35 adults—were under the supervision of correctional systems in the United States at the end of 2013 (Glaze & Kaeble, 2014). This includes approximately one in 51 adults on probation or parole and one in 110 adults incarcerated in prison or jail. These are rates higher than seen anywhere else in the world. As a point of comparison, the rate of incarceration in the United States is more than four times the rate of incarceration found in the majority of the world's countries (Walmsley, 2013). In fact, even though the United States has less than 5% of the global population, it has close to one-quarter of the world's prisoners (Walmsley, 2013). Clearly, there is a pressing need for efforts to reduce mass incarceration in the United States, including treatment to reduce recidivism and diversion of lower risk offenders to alternative settings and punishments. Indeed, not all offenders are at equal risk of recidivating (Langan & Levin, 2002) and, accordingly, may not require the same levels of supervision and intervention (Monahan & Skeem, 2016). Additionally, in contrast with the traditional "one-size-fits-all" criminal justice approach, research shows that the most effective strategies for reducing recidivism are those delivered to offenders at greater risk of recidivism that target individual needs (Andrews & Bonta, 2010). Consequently, psychologists and other professionals working in U.S. correctional agencies face mounting pressures to differentiate between offenders at greater and lower risk of recidivism and to guide decisions regarding treatment and supervision (Jung, Brown, Ennis, & Ledi, 2015; Monahan & Skeem, 2016).

In recent years, risk assessment has come to be recognized as a key component of criminal justice reform and evidence-based corrections in the United States (Casey, Warren, & Elek, 2011). There is overwhelming evidence that risk assessments completed using structured approaches produce estimates that are more reliable and more accurate than unstructured risk assessments (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Risk assessments completed using structured approaches also have been shown to lead to better public safety outcomes (Mamalian, 2011). For these reasons, instruments designed to predict risk of general recidivism, including committing a new crime and violating conditions of probation or parole, are increasingly required and being implemented in correctional agencies in the United States (Miller & Maloney, 2013; Monahan & Skeem, 2014; Monahan & Skeem, 2016). Despite the many different risk assessment instruments available, relatively little is known regarding the performance, and inter-rater reliability and predictive validity specifically, of recidivism risk assessments completed on adult offenders in U.S. correctional settings. Though there have been several high-quality reviews of risk assessment instruments to date, the ability

*Handbook of Recidivism Risk/Needs Assessment Tools*, First Edition. Edited by Jay P. Singh, Daryl G. Kroner, J. Stephen Wormith, Sarah L. Desmarais, and Zachary Hamilton.

of their findings to inform decisions regarding which recidivism risk assessment instrument to implement in U.S. correctional settings is limited in three crucial ways.

First, these reviews have focused primarily on instruments designed to predict specific forms of recidivism, notably violent or sexually violent offending, rather than predicting general recidivism more broadly (e.g., Fazel, Singh, Doll, & Grann, 2012; Gendreau, Goggin, & Little, 1996; Hanson & Morton-Bourgon, 2009; Singh, Grann, & Fazel, 2011; Tully, Chou, & Browne, 2013; Yang, Wong, & Coid, 2010). Violent and sexually violent offenders, however, comprise a relatively small proportion of the U.S. inmate population overall (Carson & Sabol, 2012). As such, the assessment of general recidivism risk is a more common task. Second, prior reviews have included a relatively short list of risk assessment instruments, typically fewer than 10, rather than including a comprehensive list of risk assessment instruments that are being used in practice. Third, and finally, prior reviews have examined the performance of risk assessments in studies conducted in multiple countries, with the predominance of studies conducted in Canada and the United Kingdom. Assessment instruments—risk assessment or otherwise—do not have reliability and validity that are transportable across populations and settings (AERA, APA, & NCME, 2014). There may be meaningful differences between offenders, assessors, and services in U.S. correctional settings and those in other jurisdictions that affect the reliability and validity of risk assessments (Monahan & Skeem, 2016).

To advance knowledge, we conducted a systematic review of studies conducted in correctional settings in the United States that have examined the predictive validity of assessments completed on adult offenders using instruments designed to predict risk of general recidivism. Our goal was to synthesize findings of the American validation research to help policymakers, psychologists, and other professionals working in U.S. correctional settings choose from among the potentially overwhelming list of risk assessment instruments available. In the sections that follow, we discuss characteristics of risk assessment instruments, samples, and studies that may affect the performance of assessments, and that, as such, we will examine in our review.

### Characteristics of Risk Assessment Instruments

Recidivism risk assessment instruments may be distinguished in terms of their approach, item type, and item content. First, there are two broad categories that distinguish between approaches used by risk assessment instruments: actuarial and structured professional judgment. The actuarial approach represents a mechanical model of risk assessment in which offenders are scored on a series of items that were most strongly associated with recidivism in the development samples. Then, total scores are cross-referenced with actuarial risk tables (Hilton, Harris, & Rice, 2006). In contrast, the structured professional judgment approach guides assessors to consider a set number of factors that are empirically and theoretically associated with the outcome of interest. Though individual items are scored, assessors ultimately make a categorical judgment of risk level (e.g., low, moderate, or high) based on their professional judgment rather than using total scores (Guy, Packer, & Warnken, 2012). Risk assessment instruments also can be differentiated by the type and content of their items, such as risk, protective, static, and dynamic factors. Risk factors are characteristics that are associated with increases in the likelihood of recidivism, whereas protective factors are characteristics associated with decreases in the likelihood of recidivism (de Ruiter & Nicholls, 2011). Risk and protective factors can either be static or dynamic in nature. Static factors are historical (e.g., history of antisocial behavior) or otherwise unchangeable characteristics (e.g., sex, race/ethnicity), whereas dynamic factors are

characteristics that may change over time and/or when targeted in treatment (e.g., substance abuse) (Douglas & Skeem, 2005).

### Characteristics of Samples and Studies

Beyond the characteristics of the instruments, performance of risk assessments may differ as a function of the characteristics of the research samples and study designs. For instance, offender race/ethnicity and sex are potentially important sources of assessment bias (Scurich & Monahan, 2016). In fact, Former U.S. Attorney General Eric Holder recently spoke *against* the use of risk assessments to inform sentencing decisions due to concerns over bias against racial/ethnic and other minorities (Holder, 2014). Some reviews of personality assessment tools and violence risk assessment instruments support his perspective, finding that assessments may be more accurate for White offenders compared to those of other racial/ethnic backgrounds (Leistico, Salekin, DeCoster, & Rogers, 2008; Singh et al., 2011). However, other studies have failed to identify such racial/ethnic biases (e.g., Guy, Edens, Anthony, & Douglas, 2005). No reviews, to our knowledge, have explored racial/ethnic biases in assessments of risk for general recidivism. With respect to offender sex, meta-analytic research suggests that we may expect risk assessment instruments to differ in their predictive validity for male compared to female offenders (Leistico et al., 2008). Again, however, the research is mixed (e.g., Holtfreter & Cupp, 2007; Singh et al., 2011; Smith, Cullen, & Latessa, 2009). Given the higher proportion of female prisoners in the United States (8.8%) compared to other jurisdictions, such as the United Kingdom (5.4%) and Canada (5.1%) (Walmsley, 2012), and the overrepresentation of racial/ethnic minorities among U.S. inmates (Carson & Sabol, 2012), there is a need to examine findings across studies conducted in U.S. correctional settings.

In addition to the characteristics of the offenders, there are aspects of the design of the studies themselves that may impact the reliability and validity of assessments completed using the recidivism risk assessment tools under investigation. For instance, much of the extant knowledge stems from research-based studies, in which researchers can carefully train and monitor assessors. However, these conditions are not necessarily present or realistic in routine practice (Desmarais et al., 2012; Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012; Douglas, Otto, Desmarais, & Borum, 2012; Vincent, Guy, Fusco, & Gershenson, 2012). Though there has been considerable discussion regarding the reliability of risk assessments completed in routine practice compared to the reliability of those completed in research studies, we are not aware of any research reviews that have specifically examined the predictive validity of risk assessments—predicting general recidivism or other outcomes—completed in the context of research versus routine practice.

### The Current Review

Herein we report findings of a systematic review of the U.S. research examining the validity of assessments completed using instruments designed to predict general recidivism among adult offenders. We sought to provide a comprehensive summary of the state of science and practice in the United States, acknowledging that, by design, our review focused on the context of the American penal system. Differences in the characteristics of offenders in the United States compared to those of offenders in other countries, combined with the remarkably high rate of incarceration in the United States, suggested the need for a review of the American empirical evidence and reflected recent calls for such data from clinicians tasked with conducting

risk assessments in U.S. correctional settings and policymakers alike (Holder, 2014). Our specific aims were to: (1) identify and describe the characteristics and content of risk assessment instruments designed to predict general recidivism that have been validated in the United States; (2) summarize the characteristics of the studies that have been conducted in U.S. correctional settings; and (3) synthesize the findings regarding the inter-rater reliability and predictive validity of risk assessments completed using these instruments on adult offenders in the United States.

## Method

### Review Protocol

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, & Altman, 2009), a 27-item checklist of review characteristics, to enable a transparent and consistent reporting of results.

### Search Strategy

#### Identification of recidivism risk assessment instruments

We identified risk assessment instruments designed to predict the likelihood of general recidivism, including new offenses and the violation of probation or parole conditions, in adult offenders by searching PsycINFO, the U.S. National Criminal Justice Reference Service Abstracts, and Google using combinations of the following keywords: *risk assessment*, *instrument*, *tool*, *general*, *recidivism*, *offending*, *parole violation*, and *prediction*. We located additional instruments using references in related systematic reviews (e.g., Fazel et al., 2012; Gendreau et al., 1996), surveys of clinicians (e.g., Singh et al., 2014; Viljoen, McLachlan, & Vincent, 2010) and through discussion with risk assessment experts. We restricted the search to instruments whose calibration studies (for actuarial instruments) or manuals (for structured professional judgment instruments) had been published by December 31, 2012. We excluded instruments if they: (a) were designed to predict specific forms of recidivism or criminal behavior (e.g., violence, sexual violence, or domestic violence); (b) were intended for guiding the assessment of juvenile offenders; (c) had not been validated in the United States; or (d) were developed for use in a specific jurisdiction or institution and had not been implemented elsewhere.

We also excluded violence risk assessment instruments (e.g., Historical, Clinical, Risk Management-20, Webster, Douglas, Eaves, & Hart, 1997; Violence Risk Appraisal Guide, Quinsey, Harris, Rice, & Cormier, 2006); clinical and behavioral inventories (e.g., Beck Depression Inventory, Beck, Steer, & Carbin, 1988; Lifestyle Criminality Screening Form, Walters, White, & Denney, 1995; Novaco Anger Scale, Novaco, 1994); personality assessment tools (e.g., Personality Assessment Inventory, Morey, 1991; Psychopathy Checklist-Revised, Hare, 2003;); and criminal thinking scales (e.g., TCU Criminal Thinking Scales, Knight, Garner, Simpson, Morey, & Flynn, 2006; Psychological Inventory of Criminal Thinking, Walters, 1995). Though often used to aid in the risk assessment process, these instruments were not designed to assess risk of general recidivism per se. Moreover, systematic reviews of their predictive validity have been reported elsewhere (Gendreau et al., 1996; Singh et al., 2011; Walters, 2012).

Using these inclusion and exclusion criteria, we identified the 19 assessment instruments or systems designed for predicting risk of general recidivism that are listed in Table 1.1. We also identified 47 instruments designed for use in specific jurisdictions or institutions. Detailed review of the latter is beyond the scope of the current analysis.

**Table 1.1** Characteristics of 19 Recidivism Risk Assessment Instruments Implemented and Validated in the United States

| | Characteristics | | | | |
|---|---|---|---|---|---|
| **Instruments** | ***k*** | **Number of Items** | **Target Population** | **Target Outcome** | **Administration Time (minutes)** |
| Correctional Offender Management Profile for Alternative Sanctions (COMPAS; Brennan et al., 2009) | 3 | 70 | All Offenders | Any Recidivism | 10–60 |
| Inventory of Offender Risks, Needs, and Strengths (IORNS; Miller, 2006) | 1 | 130 | All Offenders | Any Recidivism | 15–20 |
| Level of Service Inventory–Revised (LSI-R; Andrews & Bonta, 1995) | 25 | 54 | All Offenders | Any Recidivism | 30–40 |
| Level of Service Inventory–Revised: Screening Version (LSI-R:SV; Andrews & Bonta, 1998) | 2 | 8 | All Offenders | Any Recidivism | 10–15 |
| Ohio Risk Assessment System-Pretrial Assessment Tool (ORAS-PAT; Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009) | 3 | 7 | All Offenders | New Offenses | 10–15 |
| Ohio Risk Assessment System-Community Supervision Tool (ORAS-CST; Latessa et al., 2009) | 1 | 35 | All Offenders | New Offenses | 30–45 |
| Ohio Risk Assessment System-Community Supervision Screening Tool (ORAS- CSST; Latessa et al., 2009) | 1 | 4 | All Offenders | New Offenses | 5–10 |
| Ohio Risk Assessment System-Prison Intake Tool (ORAS-PIT; Latessa et al., 2009) | 1 | 31 | All Offenders | New Offenses | — |
| Ohio Risk Assessment System-Reentry Tool (ORAS-RT; Latessa et al., 2009) | 1 | 20 | All Offenders | New Offenses | — |
| Federal Post Conviction Risk Assessment (PCRA; Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011) | 2 | 30 | All Offenders | Any Recidivism | 15–30 |
| Risk Management System (RMS; Dow, Jones, & Mott, 2005) | 2 | 65 | All Offenders | New Offenses | — |
| Self-Appraisal Questionnaire (SAQ; Loza, 2005) | 2 | 72 | All Offenders | New Offenses | 15 |
| Salient Factor Score: Salient Factor Score-1974 Version (SFS74; Hoffman & Beck, 1974) | 3 | 9 | Parolees | New Offenses | – |
| Salient Factor Score-1976 Version (SFS76; Hoffman & Beck, 1980) | 4 | 7 | Parolees | New Offenses | – |
| Salient Factor Score-1981 Version (SFS81; Hoffman, 1983) | 8 | 6 | Parolees | New Offenses | – |
| Service Planning Instrument-Women (SPIn-W; Millson, Robinson, & Van Dieten, 2010) | 2 | 100 | All Offenders | New Offenses | — |

*(Continued)*

**Table 1.1** (Continued)

| Instruments | Characteristics | | | | |
|---|---|---|---|---|---|
| | *k* | Number of Items | Target Population | Target Outcome | Administration Time (minutes) |
| Static Risk and Offender Needs Guide (STRONG; Barnoski & Drake, 2007)[a] | 1 | 26 | All Offenders | New Offenses | — |
| Wisconsin Risk and Needs (WRN; Baird, Heinz, & Bemus, 1979) | 9 | 53 | All Offenders | New Offenses | — |
| Wisconsin Risk and Needs-Revised (WRN-R; Eisenberg, Bryl, & Fabelo, 2009) | 1 | 52 | All Offenders | New Offenses | — |

*Notes: k* = number of samples; All Offenders = inmates, probationers, and/or parolees; New Offenses = new charge, arrest, conviction, or incarceration; Violations = violations of conditions of probation or parole; Any Recidivism = new offenses or violations; Time = as reported in the instrument manual; – Data not provided.

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

### Identification of U.S. validation studies

We identified studies conducted in the United States investigating the predictive validity of the 19 recidivism risk assessment instruments through the same search engines and secondary sources as above, using both the acronyms and full names of the instruments as search criteria (see Figure 1.1). Investigations published in peer-reviewed journals were considered for inclusion, as were government reports, Master's theses, and doctoral dissertations. We included studies if their titles, abstracts, or methods sections described evaluations of validity in predicting general recidivism, including new offenses and violations of probation or parole conditions, conducted in U.S. correctional settings. When multiple instruments were administered to the same participants, we extracted predictive validity estimates for each instrument separately. When samples overlapped, we included the predictive validity estimate from the sample with the most participants to avoid double-counting. When predictive validity estimates were reported for more than one outcome, we included the estimate from the most sensitive outcome; for example, if a study reported predictive validity estimates for both arrest and incarceration, we included the arrest estimate in our analyses. We excluded studies if they only examined predictive validity of select items or scales of an instrument.

Using this search strategy, we filtered an initial total of 173 records to a final count of 53 studies (*k* samples = 72), including 26 journal articles (*k* = 30), 16 government reports (*k* = 31), two master's theses (*k* = 2), and nine doctoral dissertations (*k* = 9). References for the 53 included studies are marked with an asterisk in the reference list. As no validation studies investigating the Correctional Assessment and Intervention System (National Council on Crime and Delinquency, 2006), Community Risk/Needs Management Scale (Motiuk & Porporino, 1989), Dynamic Factor Identification and Analysis (Brown & Motiuk, 2005), Level of Service/Case Management Inventory (Andrews, Bonta, & Wormith, 2004), Level of Service/Risk-Need-Responsivity (Andrews, Bonta, & Wormith, 2008), Level of Service Inventory (Andrews, 1982), Offender Group Reconviction Scale (Copas & Marshall, 1998), Offender Assessment System (HM Prison Service and National Probation Directorate, 2001), Recidivism Risk Assessment Scales (Van der Knaap & Alberda, 2009), Risk of Reconviction (Bakker, Riley, & O'Malley, 1999), Salient Factor Score–1998 (United States Parole Commission, 2003),

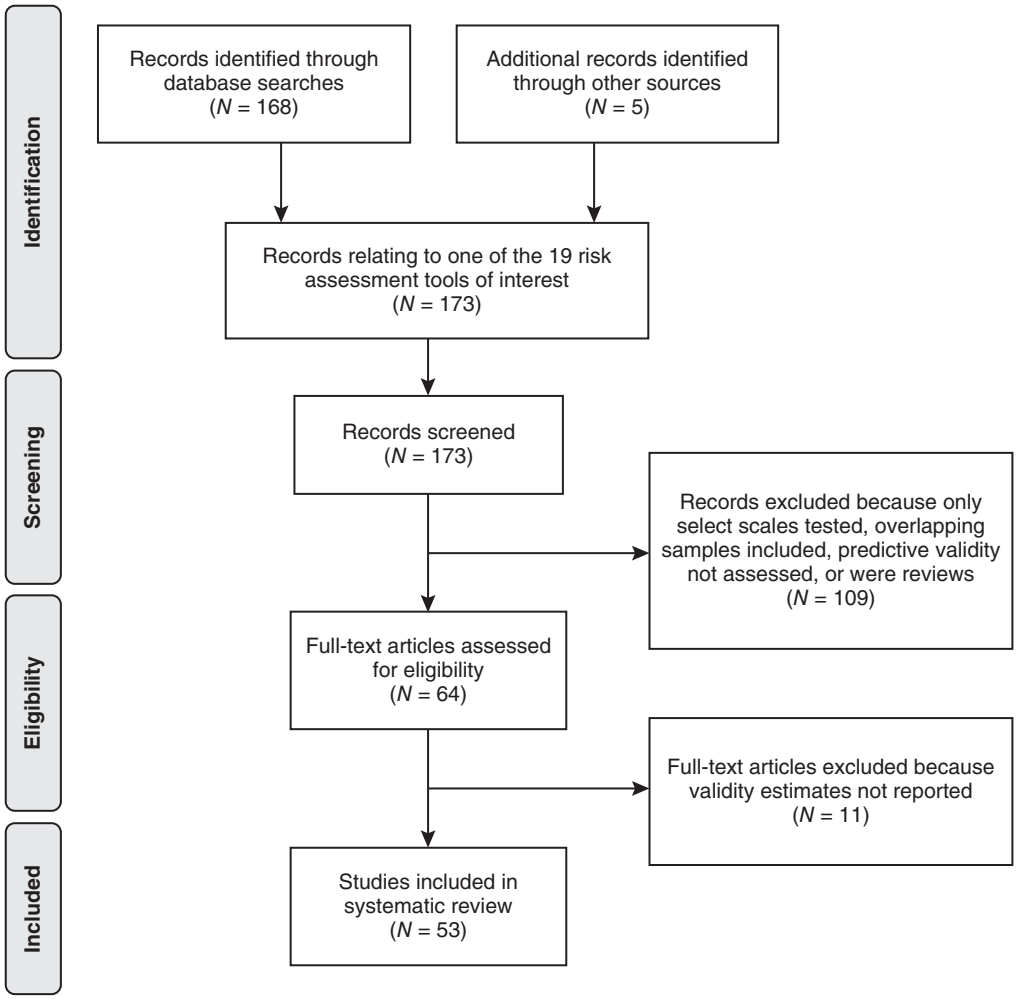


**Figure 1.1** Systematic search conducted to identify U.S. validation studies.

Statistical Information of Recidivism Scale (Nuffield, 1982), or Service Planning Instrument (Van Dieten & Robinson, 2007) met our inclusion criteria at the time of data analysis, we excluded these instruments from further review.

### Data extraction

Three research assistants enrolled in a doctoral program in psychology extracted the following information from each study: (1) demographics and design characteristics, including characteristics of the study samples (population, sample size, sex, race/ethnicity, age, psychiatric diagnoses); the assessment process (setting, timing, format, assessor, sources of information used to administer, amount of time needed to administer and score); and study designs and procedures (geographic location, research or practice context, temporal design, outcome, length of follow-up); (2) characteristics of the risk assessment instruments (assessment approach, number of items, types of items, domains measured, intended population, and predicted outcome); and (3) performance of the risk assessments (estimates of inter-rater reliability and

predictive validity). Where possible, we collected information on reliability and validity as a function of offender sex, race/ethnicity, and psychiatric diagnosis.

The research assistants were provided with a standardized extraction sheet and received training in its use by the first and second author. As a measure of quality control, 11 (20.8%) of the included studies were randomly selected and coded by all three assistants, establishing a high level of inter-rater reliability ($\kappa$ = .88; Landis & Koch, 1977). Disagreements were settled by consensus of the authors.

### Data analysis

First, sample, study design, and instrument characteristics were descriptively analyzed using measures of central tendency and dispersion parameters. Second, the item type and content of the instruments was summarized. Third, median inter-rater reliability and predictive validity estimates for total scores and risk classifications were calculated and compared across instruments, where possible. Predictive validity was assessed for any recidivism, new offenses only, and violations only as outcomes. Extracted predictive validity estimates included the area under the receiver operating characteristic curve (AUC), the point-biserial correlation coefficient ($r_{pb}$), the odds ratio (OR), and Somer's *d*, the indices most commonly reported in the field (Singh, Desmarais, & Van Dorn, 2013). Briefly, AUC represents the probability that a randomly selected recidivist would have received a higher risk rating than a randomly selected non-recidivist; $r_{pb}$ represents the direction and strength of association between risk rating and recidivism; OR represents the ratio of the odds of a lower risk rating in those who did not recidivate to the odds of a higher classification in those who did; and Somer's *d* represents 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) (see Singh, 2013). A supplemental table providing guidance regarding benchmarks and equivalency across indices of predictive validity is available online. Finally, subgroup analyses were conducted by offender sex, race/ethnicity, psychiatric diagnosis, and study context (research or practice), when possible.

## Results

### Characteristics and Content of Instruments

Table 1.1 describes the characteristics of instruments included in this review. The number of items ranged widely across instruments, from four for the ORAS-CSST to 130 for the IORNS, with an average of 41.00 (*SD* = 35.08, Range = 4–130). All instruments were intended for use across offender populations, with the exception of the SFS74, SFS76, and SFS81, which were intended for use with parolees, specifically. Most (*n* instruments = 14, 73.7%) were designed to assess risk of new offenses, excluding violations. Estimated administration time was reported in the manuals of about half of the instruments (*n* = 9, 47.4%) and when reported, ranged from 5–10 minutes for the ORAS-CSST up to 60 minutes for the COMPAS. All instruments used the actuarial approach to risk assessment.

The type and content of items included in the 19 recidivism risk assessment instruments are summarized in Table 1.2. Only two instruments, the IORNS and the SPIn-W, included protective factors; all others included risk factors exclusively. The majority of instruments (*n* = 15, 78.9%) included a combination of static and dynamic factors, with the exception of the SFS instruments and the STRONG, both of which only included static factors. None of the instruments were comprised uniquely of dynamic factors. All instruments included

**Table 1.2** Type and Content of Items Included in the 19 Recidivism Risk Assessment Instruments Implemented and Validated in the United States

| | Types | | | | Content Domains | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Instruments | Risk | Protective | Static | Dynamic | Attitudes | Associates/ Peers | Antisocial Behavior | Personality Problems | Relationships | Work/School | Leisure Activities | Substance Use Problems | Mental Health Problems | Housing Status |
| COMPAS | X | | X | X | X | X | X | X | X | X | X | X | X | X |
| IORNS | X | X | X | X | X | X | X | X | X | X | | X | X | |
| LSI-R | X | | X | X | X | X | X | X | X | X | X | X | X | X |
| LSI:SV | X | | X | X | X | X | X | | X | X | | X | X | |
| ORAS-PAT | X | | X | X | | | X | | | X | | X | | X |
| ORAS-CST | X | | X | X | X | X | X | X | X | X | X | X | | X |
| ORAS-CSST | X | | X | X | | X | X | | | X | | X | | |
| ORAS-PIT | X | | X | X | | X | X | X | X | X | X | X | X | X |
| ORAS-RT | X | | X | X | X | | X | X | X | X | | X | X | |
| PCRA | X | | X | X | X | X | X | | X | X | | X | | |
| RMS | X | | X | X | X | X | X | X | X | X | | X | X | X |
| SAQ | X | | X | X | X | X | X | X | | | | X | | |
| SFS74 | X | | X | | | | X | | | X | | X | | X |
| SFS76 | X | | X | | | | X | | | X | | X | | |
| SFS81 | X | | X | | | | X | | | | | X | | |
| SPIn-W | X | X | X | X | X | X | X | | X | X | X | X | X | X |
| STRONG | X | | X | | | | X | | | | | X | | |
| WRN | X | | X | X | X | X | X | | X | X | | X | X | |
| WRN-R | X | | X | X | X | X | X | | X | X | | X | X | |

items assessing history of antisocial behavior and substance use problems. Items assessing leisure activities, in contrast, were included relatively infrequently (*n* instruments = 5, 26.3%). Overall, the COMPAS and the LSI-R included items that captured the most content domains. The ORAS-CST, ORAS-PIT, RMS and SPIn-W evaluated all but one of the 10 domains; the exception varied for each instrument (see Table 1.2). The SFS81 and STRONG instruments considered the fewest of the domains, at two each.

### Sample and Study Characteristics

Characteristics of the 72 included samples can be found in Table 1.3. Risk assessments were completed by professionals in correctional settings for over three-quarters of the samples (81.9%); the remainder were conducted by the researchers (15.3%) or self-administered (2.8%). Assessments were most often completed in prison (27.8%) or community corrections (37.5%) settings, but also were conducted in jail (9.7%), a clinic or hospital (4.2%), or at other settings (5.6%). Setting was unstated or unclear for the remaining 11 samples (15.3%). In terms of timing, approximately one-third of samples (36.1%) included assessments conducted during community supervision, one-fourth of samples included assessments completed prior to release (26.4%), and the remaining reported assessments conducted either prior to incarceration (11.1%) or upon admission (9.7%). Timing was unstated or unclear for the remaining 12 samples (16.7%). File reviews were used to complete assessments in 24 samples (33.3%), interviews in 12 samples (16.7 %), and offender self-report in two samples (2.8%).

More than two-thirds of samples (69.4%) were investigated using a prospective study design. The average length of follow-up was almost two years (*M* = 23.5 months, *SD* = 6.3, Range = 6-138). Samples were most frequently drawn from Midwestern states (37.5%) followed by southwestern and northeastern states (11.1% each). For the majority of samples (69.4%) any recidivism as the outcome; roughly one-quarter (26.4%) reported on a variety of recidivism outcomes; and the remainder (18.1%) focused specifically on violations. The operational definition of recidivism varied, but arrest was used most frequently (30.6%), followed by conviction (12.5%), incarceration (9.7%), revocation of probation or parole (4.2%), and charge (2.8%). Assessments for the majority of samples (65.3%) were conducted in the context of routine practice rather than for the purposes of research. Nearly one-third of samples (30.6%) were from studies conducted by an author of the tool under investigation. For five instruments—the IORNS, the PCRA, the ORAS instruments, the STRONG, and the WRN-R—all studies included in our review were completed by an author of the instrument under investigation. For another three instruments—the RMS, the COMPAS, and the SFS family of instruments—at least half of the studies were completed by an author of the instrument under investigation.

More than one-third of samples (40.3%) comprised inmates and roughly one-quarter (22.2%) comprised probationers; the remainder included either parolees only (11.1%), inmates and parolees (6.9%), or probationers and parolees (11.1%). Offender legal status was not reported in six samples (8.3%). The average sample size after attrition was 5,032 (*SD* = 12599; Range = 49–51,648). The average offender age at the time of risk assessment was 33.5 years (*SD* = 10.0). In samples where sex was reported (83.3%), the majority of offenders (85.5%) were male. In samples where race/ethnicity was reported (76.4%), almost two-thirds of offenders (60.8%) were White and close to one-third (28.3%) were Black, with 13.6% identified as Hispanic. Psychiatric diagnoses were very rarely reported: Only five studies reported on the prevalence of major mental disorders, substance use disorders, or personality disorders in their samples. Each of these studies used different diagnostic categories, precluding comparisons of findings across subgroups.

**Table 1.3** Characteristics of 72 Samples Investigating the Predictive Validity of Recidivism Risk Assessment Instruments in the United States

| Category | Group | Number of *k* = 72 (%) |
|---|---|---|
| **Assessment process** | | |
| Risk assessor | Researcher | 11 (15.3) |
| | Professional | 59 (81.9)[a] |
| | Self-administered | 2 (2.8)[b] |
| Risk 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 risk assessment | Pre-incarceration | 8 (11.1) |
| | At admission | 7 (9.7) |
| | Pre-release | 19 (26.4) |
| | During community supervision | 26 (36.1) |
| | Unstated/Unclear | 12 (16.7) |
| Source 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) |
| **Study design** | | |
| 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) |
| Population | Inmates (pre-release) | 29 (40.3) |
| | Probationers | 16 (22.2) |
| | Parolees | 8 (11.1) |
| | Inmates + parolees | 5 (6.9) |
| | Probationers + parolees | 8 (11.1) |
| | Other | 6 (8.3) |

*(Continued)*

**Table 1.3** (Continued)

| Category | Group | Number of *k* = 72 (%) |
|---|---|---|
| Type of outcome | General recidivism | 50 (69.4) |
| | Violation/Breach of conditions | 13 (18.1) |
| | Mixed | 19 (26.4) |
| Length of follow-up (months) | Mean (*SD*)[d] | 23.5 (6.3) |
| Source of outcome detection | Arrest | 22 (30.6) |
| | Charge | 2 (2.8) |
| | Conviction | 9 (12.5) |
| | Incarceration | 7 (9.7) |
| | Violation of terms | 3 (4.2) |
| | Mixed | 29 (40.3) |
| **Sample demographics**[c] | | |
| Sample size after attrition | Mean (*SD*) | 5,032 (12,599) |
| Male participants (per sample) | Mean (*SD*) | 3,256 (8,965) |
| White participants (per sample) | Mean (*SD*) | 1,879 (6,148) |
| Black participants (per sample) | Mean (*SD*) | 906 (2,524) |
| Hispanic participants (per sample) | Mean (*SD*) | 685 (1,792) |
| Age at risk assessment (in years) | Mean (*SD*)[d] | 33.5 (10.0) |

*Notes: k* = number of samples; *SD* = standard deviation. Percentages may not sum to exactly 100% due to rounding.

a) Correctional officer (*k* = 35, 48.6%), parole service associate (*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, the only included instrument designed to be self-administered, was not be administered by either a researcher or professional.

c) Of those eight (11.1%) samples for which demographic characteristics were reported for samples before participant attrition, five (6.9%) had more than 25% attrition during follow-up (Dow et al., 2005; Fass, Heilbrun, DeMatteo, & Fretz, 2008; Holland, Holt, Levi, & Beckett, 1983; Miller, 2006; Millson et al., 2010).

d) Fixed-effects mean.

## Performance of Recidivism Risk Assessment Instruments

### Inter-rater reliability

Inter-rater reliability was evaluated in only two studies that met inclusion criteria, one examining the LSI-R (Simourd, 2006) and the other, the LSI-R:SV (Walters, 2011). In both studies, inter-rater reliability was excellent: 90% agreement and intra-class correlation coefficient = .80, respectively. Assessments in these studies were conducted by professionals rather than research assistants, providing strong evidence of inter-rater reliability in the field, specifically.

### Predictive validity

Table 1.4 presents the median validity estimates by instrument for the prediction of any recidivism (i.e., new offenses and/or violations), collapsed across total scores and risk classifications. Overall, no one instrument stood out as producing more accurate assessments than the others, with validity varying with the indicator reported. Specifically, the instruments that produced the risk assessments with the highest AUCs were the STRONG, SPIn-W, and PCRA.

**Table 1.4** Predictive Validity Estimates Produced by Total Scores or Risk Classifications for Any Recidivism

| | Predictive Validity Estimates | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | | $r_{pb}$ | | | OR | | | Somer's *d* | | |
| Instruments | *K* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* |
| COMPAS | 3 | 0.67 | 0.64–0.69 | 1 | 0.31 | — | 1 | 1.3 | — | — | — | — |
| LSI-R | 5 | 0.64 | 0.60–0.71 | 21 | 0.25 | 0.11–0.28 | 6 | 1.10 | 1.04–1.09 | 2 | 0.26 | 0.23–0.28 |
| LSI:SV | 1 | 0.57 | — | 1 | 0.27 | — | — | — | — | — | — | — |
| ORAS-PAT | — | — | — | 5 | 0.24 | 0.22–0.27 | — | — | — | — | — | — |
| ORAS-CST | — | — | — | 1 | 0.37 | — | — | — | — | — | — | — |
| ORAS-CSST | — | — | — | 1 | 0.38 | — | — | — | — | — | — | — |
| ORAS-PIT | — | — | — | 1 | 0.36 | — | — | — | — | — | — | — |
| ORAS-RT | — | — | — | 1 | 0.36 | — | — | — | — | — | — | — |
| PCRA | 2 | 0.71 | 0.71–0.71 | — | — | — | — | — | — | — | — | — |
| RMS | 3 | 0.67 | 0.64–0.94 | — | — | — | — | — | — | — | — | — |
| SFS74 | — | — | — | — | — | — | — | — | — | 2 | 0.34 | 0.32–0.36 |
| SFS76 | — | — | — | 1 | 0.40 | — | — | — | — | 2 | 0.36 | 0.34–0.37 |
| SFS81 | — | — | — | 4 | 0.44 | 0.39–0.46 | 2 | 3.00 | 0.76–5.23 | 5 | 0.41 | 0.38–0.52 |
| SPIn-W | 1 | 0.73 | — | — | — | — | 1 | 0.91 | — | — | — | — |
| STRONG | 1 | 0.74 | — | — | — | — | — | — | — | — | — | — |
| WRN | 3 | 0.67 | 0.61–0.74 | 6 | 0.19 | 0.10–0.21 | 1 | 0.98 | — | — | — | — |
| WRN-R | 1 | 0.66 | — | — | — | — | — | — | — | — | — | — |

*Notes: k* = number of samples; *Mdn* = median; *IQR* = inter-quartile range; AUC = area under the receiver operating characteristic curve; $r_{pb}$ = point-biserial correlation coefficient; OR = odds ratio. Estimates were calculated using either total scores or risk classifications.

The instruments that produced the risk assessments with the highest $r_{pb}$ values were the SFS81 and the SFS76. The instrument that produced the risk assessments with the highest OR was the SFS81. Finally, the instrument that produced the risk assessments with the highest Somer's *d* value was the SFS81 (see Table 1.4).

Table 1.5 presents the median validity estimates for total scores in predicting any recidivism, new offenses, and violations, in turn. Validity varied by outcome and indicator. For any recidivism (i.e., new offenses and/or violations), for example, the instrument that produced the risk assessments with the highest OR for any recidivism was the SFS81, whereas the instruments that produced the risk assessments with the highest $r_{pb}$ values were the SFS76 and SFS91 (see Table 1.5). For new offenses, the instrument that produced the risk assessments with the highest AUC value was the STRONG. The instruments that produced the risk assessments with the highest $r_{pb}$ values were the ORAS-CST and ORAS-CSST. While the $r_{pb}$ value and OR for risk assessments completed using the LSI-R were poor for the prediction of new offenses, the Somer's *d* and AUC values were stronger (see Table 1.5). For violations, the AUC value for risk assessments completed using the WRN was higher than for those completed using any other instrument (see Table 1.5).

**Table 1.5** Predictive Validity Estimates Produced by Total Scores for Any Recidivism, New Offenses, and Violations

| Outcomes and Instruments | Predictive Validity Estimates | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | | $r_{pb}$ | | | OR | | | Somer's *d* | | |
| | *K* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* |
| **Any recidivism** | | | | 2 | 0.28 | 0.27–0.28 | 1 | 1.09 | — | — | — | — |
| LSI-R | — | — | — | 1 | 0.23 | — | — | — | — | — | — | — |
| ORAS-PAT | — | — | — | 1 | 0.40 | — | — | — | — | — | — | — |
| SFS76 | — | — | — | 3 | 0.45 | 0.34–0.47 | 2 | 3.00 | 0.76–5.23 | 1 | 0.55 | — |
| SFS81 | — | — | — | — | — | — | 1 | 0.91 | — | — | — | — |
| SPIn-W | — | — | — | | | | | | | | | |
| **New offenses** | | | | | | | | | | | | |
| COMPAS | 3 | 0.67 | 0.66–0.70 | 1 | 0.31 | — | 1 | 1.30 | — | — | — | — |
| LSI-R | 4 | 0.66 | 0.61–0.71 | 17 | 0.16 | 0.11–0.26 | 5 | 1.08 | 1.04–1.10 | 1 | 0.28 | — |
| LSI:SV | 1 | 0.57 | — | 1 | 0.27 | — | — | — | — | — | — | — |
| ORAS-PAT | — | — | — | 2 | 0.23 | 0.21–0.24 | — | — | — | — | — | — |
| ORAS-CST | — | — | — | 1 | 0.37 | — | — | — | — | — | — | — |
| ORAS-CSST | — | — | — | 1 | 0.38 | — | — | — | — | — | — | — |
| ORAS-PIT | — | — | — | 1 | 0.36 | — | — | — | — | — | — | — |
| ORAS-RT | — | — | — | 1 | 0.36 | — | — | — | — | — | — | — |
| PCRA | 2 | 0.71 | 0.71–0.71 | — | — | — | — | — | — | — | — | — |
| RMS | 1 | 0.67 | — | — | — | — | — | — | — | — | — | — |
| STRONG | 1 | 0.74 | — | — | — | — | — | — | — | — | — | — |
| WRN | 2 | 0.64 | 0.61–0.67 | 5 | 0.19 | 0.10–0.21 | 1 | 0.98 | — | — | — | — |
| WRN-R | 1 | 0.66 | — | — | — | — | — | — | — | — | — | — |
| **Violations** | | | | | | | | | | | | |
| COMPAS | 1 | 0.61 | — | — | — | — | — | — | — | — | — | — |
| LSI-R | 1 | 0.62 | — | 4 | 0.24 | 0.16–0.30 | 1 | 1.09 | — | 1 | 0.23 | — |
| ORAS-PAT | — | — | — | 2 | 0.27 | 0.26–0.28 | — | — | — | — | — | — |
| RMS | 1 | 0.64 | — | — | — | — | — | — | — | — | — | — |
| WRN | 1 | 0.74 | — | — | — | — | — | — | — | — | — | — |

*Notes: k* = number of samples; *Mdn* = median; *IQR* = inter-quartile range; AUC = area under the receiver operating characteristic curve; $r_{pb}$ = point-biserial correlation coefficient; OR = odds ratio.

Table 1.6 presents the median validity estimates for risk classifications in predicting any recidivism. Overall, risk classifications were found to produce robust levels of predictive validity. Specifically, RMS and SPIn-W produced risk assessments with excellent AUC values, and strong Somer's *d* values were reported for risk assessments completed using the SFS74, SFS76, and SFS81 (see Table 1.6). The instrument that produced the risk assessments with the highest $r_{pb}$ value was the SFS81, while the $r_{pb}$ value for risk assessments completed using the WRN was much lower (see Table 1.6). There were too few studies to examine predictive validity of

**Table 1.6** Predictive Validity Estimates Produced by Risk Classifications for Any Recidivism (including New Offenses and Violations)

| | Predictive Validity Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC | | | $r_{pb}$ | | | Somer's *d* | | |
| **Instruments** | ***k*** | ***Mdn*** | ***IQR*** | ***k*** | ***Mdn*** | ***IQR*** | ***k*** | ***Mdn*** | ***IQR*** |
| RMS[a] | 1 | 0.94 | — | — | — | — | — | — | — |
| SFS74 | — | — | — | — | — | — | 2 | 0.34 | 0.32–0.36 |
| SFS76 | — | — | — | — | — | — | 2 | 0.36 | 0.34–0.37 |
| SFS81[a] | — | — | — | 1 | 0.43 | — | 4 | 0.40 | 0.38–0.45 |
| SPIn-W | 1 | 0.73 | — | — | — | — | — | — | — |
| WRN[a] | — | — | — | 1 | 0.18 | — | — | — | — |

*Notes: k* = number of samples; *Mdn* = median; IQR = inter-quartile range; AUC = area under the receiver operating characteristic curve; $r_{pb}$ = point-biserial correlation coefficient. Odds ratios were not calculated for any samples using risk classifications to measure predictive validity. The risk classifications evaluated were those recommended by instrument authors.

a) One or more estimates exclude technical violations as an outcome.

new offenses to the exclusion of violations. No studies that met our inclusion criteria reported on the validity of risk classifications in predicting violations to the exclusion of new offenses.

Table 1.7 presents the median validity estimates for risk assessments in predicting any general recidivism by offender sex. When validity estimates were reported by offender sex, instruments generally produced similar predictive validity estimates for men and women. However, predictive validity was slightly better for men than women for risk assessments completed using the LSI-R:SV and ORAS-CST, whereas the reverse was true for assessments completed using the ORAS-RT (see Table 1.7). Comparisons of predictive validity by offender race/ethnicity were possible only for assessments completed using the COMPAS and LSI-R. For COMPAS assessments, predictive validity was found in a single study to be identical for White and Black offenders (AUCs = .69; Brennan, Dieterich, & Ehret, 2009). For LSI-R assessments, predictive validity also was similar across offender race/ethnicity in the two studies reporting this data (White: $r_{pb}$ = .22; OR = 1.04, Cramer's *V* = .13; Black: OR = 1.03, Cramer's *V* = .09; Hispanic: OR = 1.03, Cramer's *V* = .10; Non-White: $r_{pb}$ = .24; Lowenkamp & Bechtel, 2007; Kim, 2010).

Finally, comparisons between the predictive validity of risk assessments completed in the context of research or routine practice were possible for the LSI-R, RMS, SPIn-W, and WRN. Table 1.8 presents the median validity estimates for risk assessments completed using these instruments in predicting any recidivism by study context. Whereas both LSI-R and WRN total scores performed comparably whether conducted in the context of research or practice, RMS risk classifications demonstrated better predictive validity when completed by researchers than by practitioners (see Table 1.8). In contrast, SPIn-W assessments performed better in the context of practice than research (see Table 1.8), though the former estimate was for total scores and the latter for risk classifications. No comparisons were possible for risk assessments completed using the COMPAS IORNS, SFS76, and SFS81 because they were only evaluated in the context of practice. Conversely, risk assessments completed using the LSI-R:SV, ORAS tools, PCRA, SAQ, SFS74, STRONG, and WRN-R were only evaluated in the context of research.

**Table 1.7** Predictive Validity Estimates Produced by Total Scores for Male and Female Offenders in U.S. Correctional Settings

| | Predictive Validity Estimates | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | | | | $r_{pb}$ | | | | | OR | | | | |
| Instruments | *k* | $Mdn_{Men}$ | $IQR_{Men}$ | $Mdn_{Women}$ | $IQR_{Women}$ | *K* | $Mdn_{Men}$ | $IQR_{Men}$ | $Mdn_{Women}$ | $IQR_{Women}$ | *k* | $Mdn_{Men}$ | $IQR_{Men}$ | $Mdn_{Women}$ | $IQR_{Women}$ |
| COMPAS | 1 | 0.67 | — | 0.69 | — | 1 | 0.32 | — | 0.32 | — | — | — | — | — | — |
| LSI-R | — | — | — | — | — | 7 | 0.25 | 0.18–0.32 | 0.23 | 0.26–0.30 | 2 | 1.06 | 1.03–1.09 | 1.03 | — |
| LSI:SV | 1 | 0.57 | — | — | — | 1 | 0.29 | — | 0.22 | — | — | — | — | — | — |
| ORAS-CST | — | — | — | — | — | 1 | 0.37 | — | 0.30 | — | — | — | — | — | — |
| ORAS-CSST | — | — | — | — | — | 1 | 0.36 | — | 0.37 | — | — | — | — | — | — |
| ORAS-PIT | — | — | — | — | — | 1 | 0.32 | — | 0.35 | — | — | — | — | — | — |
| ORAS-RT | — | — | — | — | — | 1 | 0.30 | — | 0.44 | — | — | — | — | — | — |
| SFS76 | — | — | — | — | — | 1 | 0.40 | — | — | — | — | — | — | — | — |
| SFS81 | — | — | — | — | — | 1 | 0.34 | — | — | — | 1 | 5.23 | — | — | — |
| SPIn-W | 1 | — | — | 0.73 | — | — | — | — | — | — | 1 | — | — | 0.91 | — |
| STRONG | 1 | 0.74 | — | 0.72 | — | — | — | — | — | — | — | — | — | — | — |
| WRN | — | — | — | — | — | 1 | 0.21 | — | — | — | — | — | — | — | — |

*Notes:* *k* = number of samples; *Mdn* = median; *IQR* = inter-quartile range; AUC = area under the receiver operating characteristic curve; $r_{pb}$ = point-biserial correlation coefficient; OR = odds ratio. Somer's *d* was not calculated for any samples reporting sex-specific estimates.

**Table 1.8** Predictive Validity Estimates Produced by Total Scores or Risk Classifications by Study Context

| Study Context and Instruments | Predictive Validity Estimates | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | | $r_{pb}$ | | | OR | | | Somer's *d* | | |
| | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* | *k* | *Mdn* | *IQR* |
| **Research** | | | | | | | | | | | | |
| LSI-R | 1 | .74 | — | 3 | .14 | .11–.16 | — | — | — | — | — | — |
| RMS | 1 | .94 | — | — | — | — | — | — | — | — | — | — |
| SPIn-W | — | — | — | — | — | — | 1 | .91 | — | — | — | — |
| WRN | — | — | — | 3 | .19 | .08–.21 | — | — | — | — | — | — |
| **Practice** | | | | | | | | | | | | |
| LSI-R | 4 | .63 | .60–.66 | 18 | .25 | .10–.28 | 6 | 1.09 | 1.04–1.09 | 2 | .26 | .23–.28 |
| RMS | 2 | .66 | .64–.67 | — | — | — | — | — | — | — | — | — |
| SPIn-W | 1 | .73 | — | — | — | — | — | — | — | — | — | — |
| WRN | 3 | .67 | .61–.74 | 3 | .18 | .10–.21 | 1 | .98 | — | — | — | — |

*Notes: k* = number of samples; *Mdn* = median; *IQR* = inter-quartile range; AUC = area under the receiver operating characteristic curve; $r_{pb}$ = point-biserial correlation coefficient; OR = odds ratio. Estimates were calculated using either total scores or risk classifications.

## Discussion

With staggering numbers of adults under correctional supervision in the United States, ending mass incarceration has been identified as a national priority (Obama, 2015). Efforts are underway across the United States to adopt evidence-based correctional approaches that will more appropriately and effectively incapacitate and rehabilitate offenders at greater risk of recidivism, while diverting lower risk offenders to alternative settings and punishments. Risk assessment figures prominently in many of these strategies (Casey et al., 2011). As a result, psychologists and other professionals working in U.S. correctional settings are increasingly being required to use risk assessments to inform decisions regarding incarceration, diversion, and release, and to guide the development of interventions to reduce recidivism risk (Monahan & Skeem, 2016). However, relatively little is known regarding the accuracy and reliability of recidivism risk assessments completed on adult offenders in U.S. correctional settings. Instead, prior research reviews have been characterized by relatively short lists of instruments designed to predict specific forms of recidivism, namely violent and sexually violent offending, or studies conducted in other jurisdictions, notably Canada and the United Kingdom. This review summarized the state of science and practice in the United States with respect to the performance of risk assessments completed using instruments designed to predict general recidivism, including committing a new crime and violating conditions of probation or parole, among adult offenders.

Our literature review identified 19 risk assessment instruments that had been evaluated in 53 studies published between January 1970 and December 2012 representing 72 unique samples of adult offenders in U.S. correctional settings. The risk assessment instruments varied widely in the number, type, and content of their items, but generally were characterized by

static risk factors to the exclusion of dynamic risk factors and protective factors. For most instruments, predictive validity had been evaluated in one or two studies that met our inclusion criteria. Those studies often were completed by the developers of the instrument under investigation. Perhaps one our most striking findings, only two of the 53 studies reported on the inter-rater reliability of the risk assessments. These two studies revealed very high rates of field reliability for the LSI-R and LSI-R:SV. Whether risk assessments completed using the other 17 instruments are consistent across assessors in U.S. correctional settings was not addressed in the reviewed literature, though findings of recent research are promising (e.g., Lowenkamp, Johnson, Holsinger, VanBenschoten, & Robinson, 2013). Inter-rater reliability is relevant to any forensic assessment involving the rating or coding of items (Douglas et al., 2012) and a necessary criterion for validity (Douglas, Skeem, & Nicholson, 2011; Gottfredson & Moriarty, 2006). Consequently, there is a critical need for data on the inter-rater reliability of recidivism risk assessments completed on adult offenders in U.S. correctional settings.

No one instrument emerged as producing the "most" accurate risk assessments in U.S. correctional settings; however, findings of our review suggest that some instruments may perform better in predicting particular outcomes compared to others. Risk assessments completed using the SFS instruments, for example, performed especially well in predicting any recidivism (i.e., new offenses and/or violations), whereas risk assessments completed using the ORAS-CST, ORAS-CSST, PCRA, and STRONG performed especially well in predicting new offenses. WRN assessments stood out in the prediction of violations. These findings suggest that certain risk assessment instruments may be appropriately used to inform at least some sentencing decisions, such as the allocation of probation conditions and frequency of contact. However, the more widespread use of risk assessment instruments in the sentencing process is a topic of continued debate (Monahan & Skeem, 2016).

Additionally, findings of our review suggest that some instruments produced risk assessments that perform better for male compared to female offenders in U.S. correctional settings. In particular, the LSI instruments produced assessments with only fair validity for female offenders, though predictive validity was generally good for male offenders. In contrast, a large meta-analytic review of LSI assessments reported similar effect sizes for male and female offenders (Smith et al., 2009), suggesting that our findings may be specific to the two studies that met our inclusions criteria. Other instruments, such as the COMPAS, ORAS, and STRONG, produced risk assessments with good validity for both male and female offenders. That said, our findings regarding parity or differences in the predictive validity of risk assessments for male and female offenders in U.S. correctional settings are limited by the small number of studies that provided sex-specific validity estimates. Given the rising numbers of women in U.S. jails (Glaze & Kaeble, 2014), continued efforts are needed to evaluate the validity of instruments being used to predict recidivism risk among female offenders in U.S. correctional settings.

Due to data restrictions, we were unable to systematically compare performance of risk assessments as a function of race/ethnicity, a topic receiving considerable attention in contemporary public, political, and academic discourse (Hamilton, 2015; Holder, 2014; Scurich & Monahan, 2016; Starr, 2014). We found some evidence suggesting comparable predictive validity of COMPAS and LSI-R for White and non-White offenders. However, only three studies conducted in the United States at the time of our review provided estimates of predictive validity by racial/ethnic group: one for risk assessments completed using the COMPAS and two for risk assessments completed using the LSI-R. Findings of prior reviews and studies of individual risk assessment instruments have found evidence of racial bias in the effectiveness of risk assessments (e.g., Chenane, Brennan, Steiner, & Ellison, 2015; Leistico et al., 2008; Singh et al., 2011). Conversely, other studies have failed to find differences in risk assessment accuracy as a function of offender race (e.g., Lowenkamp, Holsinger, & Cohen, 2015; Miller, 2006b).

As the use and consequences of risk assessments in the American penal system continue to grow, there is a pressing need for research that investigates potential racial disparities in assessments of risk for general recidivism.

Finally, no studies that met our inclusion criteria provided estimates of predictive validity as a function of mental disorders, substance use disorders, or personality disorders. Even when the diagnostic characteristics of the study samples were reported, predictive validity estimates were not provided by subgroup. Recent research suggests that mentally disordered and non-disordered offenders share many of the same predictors of recidivism (Skeem, Winter, Kennealy, Louden, & Tatar, 2014), suggesting that risk assessments also may perform comparably across these subgroups of offenders. Yet, this remains an empirical question to be answered through further research.

### Limitations

The methodology of our review limits its findings in three ways. First, our review focused on the state of risk assessment science and practice in United States. Our search strategy resulted in the exclusion of some recently revised versions of well-known risk assessment instruments that either had not been validated in the United States within our search time frame or that were not identifiable by their current name, but that show very positive results, such as the Level of Service/Case Management Inventory (Andrews, Bonta, & Wormith, 2004) and the Women's Risk/Need Assessment (Van Voorhis, Wright, Salisbury, & Bauman, 2010). Our interest in risk assessments completed on adult offenders in U.S. correctional settings also resulted in the exclusion of studies conducted in other jurisdictions. There have been several evaluations of the inter-rater reliability and predictive validity of assessments completed in other jurisdictions, most often Canada and the United Kingdom, or that have been published since the time of our literature review (e.g., Lowenkamp et al., 2015).

Second, our intent was to present a representative sample of all the U.S. validation research available on a comprehensive list of risk assessment instruments. Accordingly, we used an inclusive selection strategy and did not conduct a systematic assessment of study quality. For this reason, we did not undertake a formal meta-analysis and did not compute inferential statistics, but rather compared the effect sizes reported across studies descriptively. Third, although we strove to include all studies conducted in the United States published in both the peer-reviewed and grey literatures between January 1970 and December 2012, our review was still subject to publication bias. There also may be studies that met inclusion criteria but were inadvertently and unintentionally excluded by our search strategy.

### Conclusions

Despite these limitations, this review represents a comprehensive summary of the inter-rater reliability and validity of risk assessments in predicting general recidivism in adult offenders in the United States. With efforts underway across the United States to reduce mass incarceration through evidence-based criminal justice practices, our overarching goal was to provide information that would assist clinicians and policymakers alike in selecting from the many different risk assessment tools available for implementation in U.S. correctional settings. Instead of identifying one instrument that produced the "best" or "most accurate" risk assessments, our findings suggest that predictive validity may vary as a function of offender characteristics, settings, and recidivism outcomes. Our review also identified important gaps in the U.S. validation research, such as limited reporting of inter-rater reliability and few comparisons of predictive validity between offender subgroups. In light of these findings, decisions regarding which

recidivism risk assessment instrument to use should be guided by the empirical evidence—or lack thereof—supporting the instrument's use with a given population (e.g., inmates, probationers, parolees) and for the outcome of interest (e.g., new offenses, violations). Practical issues should be taken into consideration as well, such as the sources of information needed to complete the assessments (e.g., self-report, interview, review of official records), instrument length and administration time, instrument cost, and training requirements. These issues may impact the feasibility of implementing recidivism risk assessment instruments, even those that have been well-validated in a given jurisdiction and population, with fidelity (Monahan & Skeem, 2016).

## Author Note

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. We gratefully acknowledge the research assistance and contributions of Krystina Dillard, Rhonda Morelock, and Grace Seamon. We also thank David D'Amora, Fred Osher, and other Council of State Governments Justice Center staff for their support and feedback on this project.

* Sections of this chapter are reprinted with permission from Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services, 13,* 206–222. https://doi.org/10.1037/ser0000075

## References

References marked with an asterisk indicate studies included in the systematic review.

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. doi:10.1177/0011000005285875

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andrews, D. A. (1982). *Level of Service Inventory.* Toronto, ON: Ontario Ministry of Correctional Services.

Andrews, D. A., & Bonta, J. (1995). *Level of Service Inventory–Revised.* North Tonawanda, NY: Multi-Health Systems.

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

Andrews, D. A., Bonta, J., & Wormith, S. J. (2004). *The Level of Service/Case Management Inventory (LS/CMI).* Toronto, ON: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, S. J. (2008). *The Level of Service/Risk-Need-Responsivity (LS/RNR).* Toronto, ON: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16,* 39–55. doi:10.1037/a0018362

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

Bakker, L., Riley, D., & O'Malley, J. (1999). *Risk of Reconviction: Statistical models predicting four types of re-offending*. Wellington, New Zealand: Department of Corrections Psychological Service.

*Barnoski, R., & Aos, S. (2003). *Washington's Offender Accountability Act: An analysis of the Department of Corrections' risk assessment*. Olympia, WA: Washington State Institute for Public Policy.

*Barnoski, R., & Drake, E. K. (2007). *Washington's Offender Accountability Act: Department of Corrections' static risk assessment*. Olympia, WA: Washington State Institute for Public Policy.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100. doi:10.1016/0272-7358(88)90050-5

*Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *Validation of the COMPAS risk assessment classification instrument.* Tallahassee, FL: Center for Criminology and Public Policy Research.

*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. doi:10.1177/0093854808326545

Brown, S. L., & Motiuk, L. L. (2005). *The Dynamic Factor Identification and Analysis (DFIA) component of the Offender Intake Assessment (OIA) process: A meta-analytic, psychometric, and consultative review* (Research Report R-164). Ottawa, ON: Correctional Service Canada.

Carson, E. A., & Sabol, W. J. (2012, December). *Prisoners in 2011* (NCJ 239808). Washington, DC: Bureau of Justice Statistics.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011). *Using offender risk and needs assessment information at sentencing: Guidance for courts from a national working group*. Williamsburg, VA: National Center for State Courts. Retrieved from http://ncsc.contentdm.oclc.org/cgi-bin/showfile.exe?CISOROOT=/criminal&CISOPTR=196

*Castillo, E. D., & Alarid, L. F. (2011). Factors associated with recidivism among offenders with mental illness. *International Journal of Offender Therapy and Comparative Criminology*, *55*(1), 98-117. doi:10.1177/0306624X09359502

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. doi:10.1080/03610911003650383

Chenane, J. L., Brennan, P. K., Steiner, B., & Ellison, J. M. (2015). Racial and ethnic differences in the predictive validity of the Level of Service Inventory–Revised among prison inmates. *Criminal Justice and Behavior*, *42*, 286–303. doi:10.1177/0093854814548195

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

*Connolly, M. M. (2003). *A critical examination of actuarial offender-based prediction assessments: Guidance for the next generation of assessments*. Unpublished doctoral dissertation, University of Texas at Austin, Austin, TX.

Copas, J., & Marshall, P. (1998). The Offender Group Reconviction Scale: The statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society, Series C*, *47*, 159–171. doi:10.1111/1467-9876.00104

*Davidson, C. (2012). *2012 Iowa Board of Parole risk assessment validation*. Des Moines, IA: Department of Human Rights, Division of Criminal and Juvenile Justice Planning.

*Davidson, J. (2007). *Risky business: What standard assessments mean for female offenders.* Unpublished doctoral dissertation, University of Hawaii at Manoa, Manoa, HI.

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *The International Journal of Forensic Mental Health*, *10*, 160–170. doi:10.1080/14999013.2011.600602

Desmarais, S. L., Sellers, B. G., Viljoen, J. L., Cruise, K. R., Nicholls, T. L., & Dvoskin, J. A. (2012). Pilot implementation and preliminary evaluation of START:AV assessments in secure juvenile correctional facilities. *The International Journal of Forensic Mental Health*, *11*, 150–164. doi:10.1080/14999013.2012.737405

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 and the Law*, *30*, 448–469. doi:10.1002/bsl.2022

Douglas, K. S., Otto, R., Desmarais, S. L., & Borum, R. (2012). Clinical forensic psychology. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, volume 2: Research methods in psychology* (pp. 213–244). Hoboken, NJ: John Wiley & Sons.

Douglas, K. S., & Skeem, J. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law*, *11*, 347–383. doi:10.1037/1076-8971.11.3.347

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.

*Dow, E., Jones, C., & Mott, J. (2005). An empirical modeling approach to recidivism classification. *Criminal Justice and Behavior*, *32*, 223–247. doi:10.1177/0093854804272892

*Eaglin, J., & Lombard, P. (1982). *A validation and comparative evaluation of four predictive devices for classifying federal probation caseloads*. Washington, DC: Federal Judicial Center.

*Eisenberg, M., Bryl, J., & Fabelo, T. (2009). *Validation of the Wisconsin Department of Corrections risk assessment instrument*. New York, NY: Council of State Governments Justice Center.

*Evans, S. E. (2009). *Gender disparity in the prediction of recidivism: The accuracy of LSI-R modified*. Master's thesis, University of Alabama, Tuscaloosa, AL.

*Farabee, D., & Zhang, S. (2007). *COMPAS validation study: First annual report*. Los Angeles, CA: Department of Corrections and Rehabilitation.

*Farabee, D., Zhang, S., Roberts, R. E. L., & Yang, J. (2010). *COMPAS validation study: Final report*. Los Angeles, CA: Department of Corrections and Rehabilitation.

*Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, *35*, 1095–1108. doi:10.1177/0093854806298468

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.

*Flores, A. W., Lowenkamp, C. T., Smith, P., & Latessa, E. J. (2006). Validating the Level of Service Inventory–Revised on a sample of federal probationers. *Federal Probation*, *70*(2), 44–48.

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. & Kaeble, D. (2014, December). *Correctional population in the United States, 2013* (NCJ 248479). Washington, DC: Bureau of Justice Statistics.

Gottfredson, D. S., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*, 178–200. doi:10.1177/0011128705281748

*Gould, L. (1991). *A comparison of models of parole outcome*. Unpublished doctoral dissertation, Louisiana State University, Baton Rouge, LA.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. doi:10.1037/1040-3590.12.1.19

Guy, L. S., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting and Clinical Psychology*, *73*, 1056–1064. doi:10.1037/0022-006X.73.6.1056

Guy, L. S., Packer, I. K., & Warnken, W. (2012). Assessing risk of violence using structured professional judgment guidelines. *Journal of Forensic Psychology Practice*, *12*, 270–283. doi:10.1080/15228932.2012.674471

HM Prison Service and National Probation Directorate. (2001). *The Offender Assessment System: User manual*. London, England: Home Office.

Hamilton, M. (2015). Risk-needs assessment: Constitutional and ethical challenges. *American Criminal Law Review*, *52*, 231.

Hanson, R., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*, 1–21. doi:10.1037/a0014421

Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised* (2nd ed.). Toronto, ON: Multi-Health Systems.

*Harer, M. (1994). *Recidivism among federal prison releasees in 1987: A preliminary report*. Washington, DC: Federal Bureau of Prisons Office of Research and Evaluation.

*Henderson, H. M. (2006). *The predictive utility of the Wisconsin Risk Needs Assessment in a sample of Texas probationers*. Unpublished doctoral dissertation, Sam Houston State University, Huntsville, TX.

Hilton, N., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist*, *34*, 400–409. doi:10.1177/0011000005285877

*Hoffman, P. (1983). Screening for risk: A revised Salient Factor Score (SFS 81). *Journal of Criminal Justice*, *11*, 539–547. doi:10.1016/0047-2352(83)90006-5

*Hoffman, P. (1994). 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. doi:10.1016/0047-2352(94)90090-6

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

Hoffman, P., & Beck, J. (1974). Parole decision-making: A Salient Factor Score. *Journal of Criminal Justice*, *2*, 195–206. doi:0.1016/0047-2352(74)90031-2

*Hoffman, P., & Beck, J. (1980). Revalidating the Salient Factor Score: A research note. *Journal of Criminal Justice*, *8*, 185–188. doi:10.1016/0047-2352(80)90025-2

*Hoffman, P., & Beck, J. (1985). Recidivism among released federal prisoners: Salient Factor Score and five-year follow-up. *Criminal Justice and Behavior*, *12*, 501–507. doi:10.1177/0093854885012004007

Holder, E. H. (2014). *Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and the 13th State Criminal Justice Network conference*. August 1, 2014. Retrieved from http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

*Holland, T., Holt, N., Levi, M., & Beckett, G. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology*, *68*, 203–211. doi:10.1037/0021-9010.68.2.203

*Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2004). Validating the LSI-R on a sample of jail inmates. *Journal of Offender Monitoring*, *17*, 8–9.

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. doi:10.1177/1043986207309436

*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, Palo Alto University, Palo Alto, CA.

*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.

Jung, S., Brown, K., Ennis, L., & Ledi, D. (2015). The association between presentence risk evaluations and sentencing outcome. *Applied Psychology in Criminal Justice*, *11*, 111–125.

*Kelly, B. (2009). *A validation study of Risk Management Systems*. Master's thesis, University of Nevada, Las Vegas, NV.

*Kelly, C. E., & Welsh, W. N. (2008). The predictive validity of the Level of Service Inventory–Revised for drug-involved offenders. *Criminal Justice and Behavior*, *35*, 819–831. doi:10.1177/0093854808316642

*Kim, H. (2010). *Prisoner classification re-visited: A further test of the Level of Service Inventory–Revised (LSI-R) intake assessment*. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Indiana, PA.

Knight, K., Garner, B. R., Simpson, D. D., Morey, J. T., & Flynn, P. M. (2006). An assessment for criminal thinking. *Crime & Delinquency*, *52*, 159–177. doi:10.1177/0011128705281749

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

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

*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.

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. doi:10.1007/s10979-007-9096-6

*Listwan, S., Piquero, N., & Voorhis, P. (2010). Recidivism among a white-collar sample: Does personality matter? *Australian and New Zealand Journal of Criminology*, *43*, 156–174. doi:10.1375/acri.43.1.156

*Lowenkamp, C. T., & Bechtel, K. (2007). The predictive validity of the LSI-R on a sample of offenders drawn from the records of the Iowa Department of Corrections Management System. *Federal Probation*, *71*, 25–29.

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA revisited: Testing the validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services*, *12*, 149–157. doi:10.1037/ser0000024

*Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2001). Risk/need assessment, offender classification, and the role of childhood abuse. *Criminal Justice and Behavior*, *28*, 543–563. doi:10.1177/009385480102800501

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The Federal Post Conviction Risk Assessment (PCRA): A Construction and validation study. *Psychological Services*, *10*, 87–96. doi:10.1037/a0030343

*Lowenkamp, C. T., & Latessa, E. J. (2004). *Validating the Level of Service Inventory–Revised in Ohio's community based correctional facilities*. Cincinnati: Center for Criminal Justice Research.

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

*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. *Prison Journal*, *89*, 192–204.

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

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

*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*. Master's thesis, Central Connecticut State University, New Britain, CT.

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 and the Law*, *24*, 767–782. doi:10.1002/bsl.728

Miller, J., & Maloney, C. (2013). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice and Behavior*, *40*, 716–736. doi:10.1177/0093854812468883

*Millson, B., Robinson, D., & Van Dieten, M. (2010). *Women offender case management model*. Wethersfield, CT: Court Support Services Division.

*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. doi:10.1177/0093854806287421

*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. doi:10.1177/0306624X12436504

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-Analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. doi:10.1371/journal.pmed.1000097

Monahan, J., & Skeem, J. (2014). Risk redux: The resurgence of risk assessment in criminal sanctioning. *Federal Sentencing Reporter*, *26*, 158–166.

Monahan, J., & Skeem, J. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*, 489–513.

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Motiuk, L. L., & Porporino, F. J. (1989). *Field test of the Community Risk/Needs Management scale: A study of offenders on caseload*. Correctional Service Canada, Research Branch.

Novaco, R. W. (1994). Anger as a risk factor for violence among the mentally disordered. In J. Monahan & H. Steadman (Eds.), *Violence and mental disorder: Developments in risk assessment* (pp. 21–59). Chicago, IL: University of Chicago Press.

Nuffield, J. (1982). *Parole decision making in Canada: Research towards decision guidelines*. Ottawa, ON: Solicitor General of Canada.

Obama, B. (2015). *President Obama Remarks on the Criminal Justice System*. July 14, 2015. Retrieved from http://www.c-span.org/video/?327099-4/president-obama-remarks-naacp

*O'Keefe, M. L., Klebe, K., & Hromas, S. (1998). *Validation of the Level of Supervision Inventory (LSI) for community-based offenders in Colorado: Phase II*. Model, CO: Department of Corrections.

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

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.

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. doi:10.1007/s10979-005-6832-7

*Robuck, B. (1976). *A study of inmate outcome in Kentucky*. Unpublished doctoral dissertation, University of Kentucky, Lexington, KY.

*Rubin, M., Rocque, M., & Ethridge, W. (2010). *An analysis of probation violations and revocations in Maine probation entrants in 2005–2006*. Portland, ME: Justice Research and Statistics Association.

*Schlager, M. D. (2005). *Assessing the reliability and validity of the Level of Service Inventory–Revised (LSI-R) on a community corrections sample: Implications for corrections and parole policy*. Unpublished doctoral dissertation, Rutgers University, Newark, NJ.

Scurich, N., & Monahan, J. (2016). Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, *40*, 36–41. doi:10.1037/lhb0000161

*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. doi:10.1177/0887403410372989

*Simourd, D. (2006). *Validation of risk/needs assessments in the Pennsylvania Department of Corrections: Final report*. Hampden Township, PA: Department of Corrections.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, *31*, 8–22. doi:10.1002/bsl.2052

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., de Ruiter, C., . . . Otto, R. K. (2014). Use and perceived utility of structured violence risk assessment tools in 44 countries: Findings from the IRiS Project. *International Journal of Forensic Mental Health Services*, *13*, 193–206. doi:10.1080/14999013.2014.922141

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. doi:10.1002/bsl.2053

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499–513. doi:10.1016/j.cpr.2010.11.009

Skeem, J. L., Winter, E., Kennealy, P. J., Louden, J. E., & Tatar II, J. R. (2014). Offenders with mental illness have criminogenic needs, too: Toward recidivism reduction. *Law and Human Behavior*, *38*, 212–224. doi:10.1037/lhb0000054

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. doi:10.1111/j.1745-9133.2009.00551.x

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803–872.

*Tillyer, M. S., & Vose, B. (2011). Social ecology, individual risk, and recidivism: A multilevel examination of main and moderating influences. *Journal of Criminal Justice*, *39*, 452–459. doi:10.1016/j.jcrimjus.2011.08.003

Tully, R. J., Chou, S., & Browne, K. D. (2013). A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review*, *33*, 287–316. doi:10.1016/j.cpr.2012.12.002

United States Parole Commission. (2003). *Rules and procedures manual*. Chevy Chase, MD: U.S. Parole Commission.

Van der Knaap, L. M., & Alberda, D. L. (2009). *De predictieve validiteit van de Recidive Inschattingsschalen (RISc) [Predictive validity of the Recidivism Risk Assessment Scales (RISc)]*. Den Haag, the Netherlands: Ministerie van Justitie, WODC.

Van Dieten, M., & Robinson, D. (2007). *The Service Planning Instrument (SPIn)*. Ottawa, ON: Orbis Partners.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: The current status of gender responsive assessment. *Criminal Justice and Behavior*, *34*, 261–288. doi:10.1177/0093854809357442

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment*, *17*, 377–395. doi:10.1177/1073191109359587

Vincent, G. M., Guy, L. S., Fusco, S. L., & Gershenson, B. G. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior*, *36*, 225–236. doi:10.1037/h0093974

*Vose, B., Lowenkamp, C. T., Smith, P., & Cullen, F. T. (2009). Gender and the predictive validity of the LSI-R: A study of parolees and probationers. *Journal of Contemporary Criminal Justice*, *25*, 459–471. doi:10.1177/1043986209344797

Walmsley, R. (2013). *World prison population list* (10th ed.). London, England: International Centre for Prison Studies.

Walmsley, R. (2012). *World female imprisonment list* (2nd ed.). London, England: International Centre for Prison Studies.

Walters, G. D. (1995). The Psychological Inventory of Criminal Thinking Styles Part I: Reliability and preliminary validity. *Criminal Justice and Behavior*, *22*, 307–325. doi:10.1177/0093854895022003008

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. doi:10.1007/s10979-010-9231-7

Walters, G. D. (2012). Criminal thinking and recidivism: Meta-analytic evidence on the predictive and incremental validity of the Psychological Inventory of Criminal Thinking Styles (PICTS). *Aggression and Violent Behavior*, *17*, 272–278. doi:10.1016/j.avb.2012.02.010

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

*Wright, E. M., Van Voorhis, P., Bauman, A., & Salisbury, E. (2007). *Gender-responsive risk/needs assessment: Final report*. St Paul, MN: Department of Corrections.

*Yacus, G. M. (1998). *Validation of the risk and needs assessment used in the classification for parole and probation of Virginia's adult criminal offenders*. Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Yang, M., Wong, S. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*, 740–767. doi:10.1037/a0020473

## Supplemental Table

Benchmarks Across Indices of Predictive Validity

| | Indices of Predictive Validity | | | | |
|---|---|---|---|---|---|
| **Benchmarks** | **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:* Benchmarks 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.