Clinical Psychology Review 31 (2011) 499–513

Contents lists available at ScienceDirect

Clinical Psychology Review

A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants

Jay P. Singh [a], Martin Grann [b], Seena Fazel [a,*]

[a] Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, OX3 7JX, UK
[b] Swedish Prison and Probation Service, and Centre for Violence Prevention, Karolinska Institute, Sweden

ARTICLE INFO

Article history:
Received 25 August 2010
Received in revised form 26 October 2010
Accepted 29 November 2010
Available online 13 December 2010

Keywords:
Violence
Risk assessment
Meta-analysis
Forensic
Crime
Area under curve

ABSTRACT

There are a large number of structured instruments that assist in the assessment of antisocial, violent and sexual risk, and their use appears to be increasing in mental health and criminal justice settings. However, little is known about which commonly used instruments produce the highest rates of predictive validity, and whether overall rates of predictive validity differ by gender, ethnicity, outcome, and other study characteristics. We undertook a systematic review and meta-analysis of nine commonly used risk assessment instruments following PRISMA guidelines. We collected data from 68 studies based on 25,980 participants in 88 independent samples. For 54 of the samples, new tabular data was provided directly by authors. We used four outcome statistics to assess rates of predictive validity, and analyzed sources of heterogeneity using subgroup analysis and metaregression. A tool designed to detect violence risk in juveniles, the Structured Assessment of Violence Risk in Youth (SAVRY), produced the highest rates of predictive validity, while an instrument used to identify adults at risk for general offending, the Level of Service Inventory – Revised (LSI-R), and a personality scale commonly used for the purposes of risk assessment, the Psychopathy Checklist – Revised (PCL-R), produced the lowest. Instruments produced higher rates of predictive validity in older and in predominantly White samples. Risk assessment procedures and guidelines by mental health services and criminal justice systems may need review in light of these findings.

Contents

* Corresponding author. Tel.: +44 8452191166; fax: +44 1865793101.
E-mail address: seena.fazel@psych.ox.ac.uk (S. Fazel).

0272-7358/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cpr.2010.11.009

1. Introduction

Risk assessment tools assist in the identification and management of individuals at risk of harmful behaviour. Due to the potential utility of such tools, researchers have developed many risk assessment instruments, the manuals for which promise high rates of construct and predictive validity (Bonta, 2002). Recent meta-analyses have identified over 120 different risk assessment tools currently used in general and psychiatric settings (for a metareview, see Singh & Fazel, 2010). These measures range from internationally utilized tools such as the Historical, Clinical, Risk Management – 20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) to locally developed and implemented risk measures such as the North Carolina Assessment of Risk (NCAR; Schwalbe, Fraser, Day, & Arnold, 2004). Given the large selection of tools available to general and secure hospitals and clinics, prisons, the courts, and other criminal justice settings, a central question is which measures have the highest rates of predictive accuracy. To date, no single risk assessment tool has been consistently shown to have superior ability to predict offending (Campbell, French, & Gendreau, 2007; Gendreau, Goggin, & Smith, 2002; SBU, 2005; Walters, 2003), and several major uncertainties remain regarding the populations and settings in which risk assessments may be accurately used (Leistico, Salekin, DeCoster, & Rogers, 2008; Guy, Edens, Anthony, & Douglas, 2005; Schwalbe, 2008; Smith, Cullen, & Latessa, 2009).

Such uncertainties are important given that risk assessment tools have been increasingly used to influence decisions regarding accessibility of inpatient and outpatient resources, civil commitment or preventative detention, parole and probation, and length of community supervision in many Western countries including the US (Cottle, Lee, & Heilbrun, 2001; Schwalbe, 2008), Canada (Gendreau, Goggin, & Little, 1996; Hanson & Morton-Bourgon, 2007), UK (Kemshall, 2001; Khiroya, Weaver & Maden, 2009), Sweden (SBU, 2005), Australia (Mercado & Ogloff, 2007), and New Zealand (Vess, 2008). Recent work has suggested that the influence of risk assessment tools appears to be growing in both general and forensic settings. For example, violence risk assessment is now recommended in clinical guidelines for the treatment of schizophrenia in the US and the UK (APA, 2004; NICE, 2009). In the US, risk assessment tools are used routinely in the mental health care systems of the majority of the 17 states that have civil commitment laws (Mercado & Ogloff, 2007). Recent studies in England have found that two thirds of mental health clinicians in general settings are using structured risk assessment forms (Higgins, Watts, Bindman, Slade & Thornicroft, 2005), as are clinicians working in over 70% of forensic psychiatric units (Khiroya et al., 2009). Risk measures are also being used with increasing regularity in both criminal and civil court cases in the US and the UK (DeMatteo & Edens, 2006; Young, 2009). The widespread, often legally required use of risk measures (Seto, 2005) necessitates the regular and high-quality review of the evidence base.

1.1. Uncertainties in risk assessment

The research base on the predictive validity of risk assessment tools has expanded considerably; however, policymakers and clinicians continue to be faced with conflicting findings of primary and review literature on a number of central issues (Gendreau, Goggin & Smith, 2000; Singh & Fazel, 2010). Key uncertainties include:

(1) Are there differences between the predictive validity of risk assessment instruments?
(2) Do risk assessment tools predict the likelihood of violence and offending with similar validity across demographic backgrounds?
(3) Do actuarial instruments or tools which employ structured clinical judgment produce higher rates of predictive validity?

1.1.1. Demographic factors

There is contrasting evidence whether risk assessment tools are equally valid in men and women. Several recent reviews have found no difference in tool performance between the genders (e.g., Schwalbe, 2008; Smith et al., 2009). Schwalbe (2008) conducted a meta-analysis on the validity literature of risk assessment instruments adapted for use in juvenile justice systems, and found no differences in predictive validity based on gender. This finding was supported by a meta-analysis conducted by Smith, Cullen, and Latessa (2009), who found that the Level of Service Inventory – Revised (LSI-R) produced non-significantly different rates of predictive validity in men and women. In contrast, recent meta-analyses have found that the predictive validity of certain risk assessment tools is higher in juvenile men (Edens, Campbell & Weir, 2007) or in women (Leistico et al., 2008).

Another uncertainty is whether risk measures' predictive validity differs across ethnic backgrounds. There is evidence from primary studies and meta-analyses that risk assessment tools provide more accurate risk predictions for White participants than other ethnic backgrounds (Bhui, 1999; Edens et al., 2007; Långström, 2004; Leistico et al., 2008). This variation may be due to differences in the base rate of offending among individuals of different ethnicities (Federal Bureau of Investigation, 2002). These differences are seen in inpatient settings (Fujii, Tokioka, Lichton & Hishinuma, 2005; Hoptman, Yates, Patalinjug, Wack & Convit, 1999; Lawson, Yesavage & Werner, 1984; McNiel & Binder, 1995; Wang & Diamon, 1999) and on discharge into the community (Lidz, Mulvey, & Gardner, 1993). Contrary evidence has been provided by reviews which have assessed the moderating influence of ethnicity on predictive validity rates in White, Black, Hispanic, Asian, and Aboriginal participants and have found no differences (Guy et al., 2005; Edens & Campbell, 2007; Schwalbe, 2007; Skeem, Edens, Camp, & Colwell, 2004).

Previous meta-analyses (e.g., Blair, Marcus & Boccaccini, 2008; Guy, 2008; Leistico et al., 2008) have found that participant age does not affect the predictive validity of risk assessment tools. However, epidemiological investigations and reviews (e.g., Gendreau et al., 1996) have found that younger age is a significant risk factor for offending. Therefore, we investigated the influence of age on predictive validity in the present meta-analysis.

1.1.2. Actuarial instruments vs. tools employing structured clinical judgment

Actuarial risk assessment tools estimate the likelihood of misconduct by assigning numerical values to risk factors associated with offending. These numbers are then combined using a statistical algorithm to translate an individual's total score into a probabilistic estimate of offending. The actuarial approach is an attempt to ensure that each individual is appraised using the same criteria and, in doing so, can be directly compared to others who have had the same tool administered regardless of who conducted the assessment.

The individual administering the assessment is thought to play a critical role when clinically based instruments are used. Broadly, clinical approaches to risk assessment can be dichotomized into unstructured clinical judgment and structured clinical judgment. Unstructured clinical judgment refers to a clinician's subjective prediction of whether an individual is likely to offend or not. No prespecified set of risk factors is used to guide the clinician's analysis, which relies on professional experience for accuracy in assessing the likelihood of offending (Hanson, 1998). Recent reviews have suggested that this form of risk assessment has poor predictive validity (Daniels, 2005; Hanson & Morton-Bourgon, 2009). The poor performance of unstructured clinical judgment is thought to be a consequence of its reliance on subjective risk estimates that lack inter-rater and test–retest reliability (Hanson & Morton-Bourgon, 2009).

To increase construct validity and reliability, the authors of risk assessment tools developed new measures that adopt an approach known as structured clinical judgment (SCJ). In this approach, clinicians use empirically-based risk factors to guide their predictions of an individual's risk for offending (Douglas, Cox & Webster, 1999). Advocates of this approach believe that SCJ does more than simply assess risk, it provides information that can be used for treatment planning and risk management (Douglas et al., 1999; Heilbrun, 1997). While past reviews have provided evidence that actuarial tools produce higher rates of predictive validity than instruments which rely on structured clinical judgment (Hanson & Morton-Bourgon, 2009; Hanson & Morton-Bourgon, 2007), other researchers have presented evidence that both forms of risk assessment produce equally valid predictions (Guy, 2008). Due to this uncertainty, we investigated this issue in the present meta-analysis.

1.1.3. Study design characteristics

Secondary uncertainties in the field of risk assessment include whether study design characteristics such as study setting, prospective vs. retrospective design, or length of follow-up influence predictive validity (Singh & Fazel, 2010).

1.1.3.1. Study setting. Prisons, psychiatric hospitals, courts, and the community are typical settings in most research in the forensic risk assessment literature (Bjørkly, 1995; DeMatteo & Edens, 2006; Edens, 2001; Bauer, Rosca, Khawalled, Gruzniewski & Grinshpoon, 2003). Meta-analytic evidence regarding the moderating role of study setting on effect size varies (Leistico et al., 2008; Skeem et al., 2004). Some experts (e.g., Edens, Skeem, Cruise & Cauffman, 2001) suggest that differences in the accuracy of risk assessments may be attributed to the contextual differences in these study settings. Due to these differences, one measure may be superior to others in one study setting but not in another (Hanson & Morton-Bourgon, 2007).

1.1.3.2. Prospective vs. retrospective methodology. Whether a study has a prospective or retrospective design may influence predictive validity findings. Being that the primary goal of risk assessment is to predict future offending, some researchers have stated that prospective research is not just appropriate, but necessary to establish a tool's predictive validity (Caldwell, Bogat & Davidson, 1988). However, strengths of retrospective study designs are that researchers do not have to wait for time to elapse before they investigate whether the studied individuals reoffended or not. This methodology is particularly useful with low base rate outcomes such as violent crime (Maden, 2001). Both actuarial and clinically based instruments can be used retrospectively. The latter tools can be scored using file information from sources such as psychological reports, institutional files, and/or court reports (de Vogel, de Ruiter, Hildebrand, Bos & van de Ven, 2004).

1.1.3.3. Length of follow-up. The evidence regarding the long-term efficacy of risk assessment tools is mixed. Using ROC analysis, Mossman (2000) concluded that accurate long-term predictions of violence were possible. In support, several recent meta-analyses have found that length of follow-up does not moderate effect size (Blair, Marcus, & Boccaccini, 2008; Edens & Campbell, 2007; Edens et al., 2007; Schwalbe, 2007). Contrasting evidence has been found by reviews which have reported higher rates of predictive validity for studies with longer follow-up periods (Leistico et al., 2008; McCann, 2006; Smith et al., 2009). Other researchers have found tools to be valid only in the short-term and, even then, only at modest levels (Bauer et al., 2003; Sreenivasan, Kirkish, Garrick, Weinberger & Phenix, 2000). Given that studies often follow participants for different lengths of time and given that effect size may vary with time at risk, the role of this variable needs to be examined (Cottle, Lee, & Heilbrun, 2001). Very few tools have been validated for short follow-up time, such as hours, days or one to two weeks, which typically is a most relevant timeframe in clinical real-world decisions-making situations (SBU, 2005).

In summary, despite the increasing use and potential importance of risk assessment instruments, it is unclear which instruments have the highest rates of predictive validity, and whether these rates differ by important demographic and study design characteristics. We have therefore undertaken a systematic review and meta-analysis to explore rates of predictive validity in commonly used instruments and to assess the potential sources of heterogeneity outlined above.

2. Method

2.1. Review protocol

The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff & Altman, 2009), a 27-item checklist of review characteristics designed to enable a transparent and consistent reporting of results, was followed.

2.2. Tool selection

Our goal was to analyze the predictive validity of the most commonly used risk assessment tools in the field today. Based on reviews of the literature (e.g., Bonta, 2002; Doren, 2002; Kemshall, 2001; Singh & Fazel, 2010), we identified the following 9 instruments that appeared to be used most often in the context of forensic risk assessment: the *Historical, Clinical, Risk Management – 20* (HCR-20; Webster et al., 1997; Webster, Eaves, Douglas & Wintrup, 1995), the *Level of Service Inventory – Revised* (LSI-R; Andrews & Bonta, 1995), the *Psychopathy Checklist – Revised* (PCL-R; Hare, 1991, 2003), the *Sex Offender Risk Appraisal Guide* (SORAG; Quinsey, Harris, Rice & Cormier, 1998, 2006), the *Sexual Violence Risk – 20* (SVR-20; Boer, Hart, Kropp & Webster, 1997), the *Spousal Assault Risk Assessment* (SARA; Kropp, Hart, Webster & Eaves, 1994, 1995, 1999), the *Static-99* (Harris, Phenix, Hanson & Thornton, 2003; Hanson & Thornton, 1999), the *Structured Assessment of Violence Risk in Youth* (SAVRY; Borum, Bartel & Forth, 2002, 2003), and the *Violence Risk Appraisal Guide* (VRAG; Quinsey, Harris, Rice & Cormier, 1998, 2006). For a description of each tool and a summary of their psychometric properties, see Table 1.

2.3. Search strategy

A systematic search was conducted in PsycINFO, EMBASE, MEDLINE, and National Criminal Justice Reference Service Abstracts. The search criteria consisted of the acronyms and full names of the nine risk assessment tools. The search was restricted to articles which had been published between January 1, 1995 and November 30, 2008 because we wished the review to summarize the contemporary literature. Additional articles were located through the reference lists of previous reviews, through annotated bibliographies of identified risk assessment tools, and through discussion with researchers in the field. Studies in all languages from all countries were considered for inclusion as were studies not published in peer-reviewed journals (also known as “gray literature”; i.e., doctoral dissertations, Master's theses, conference presentations, and government reports). Studies were included in the meta-analysis if their titles and/or abstracts

Table 1
Characteristics of the nine risk assessment tools included in the meta-analysis.

Tool	Act vs. SCJ	Items	Description	Point system	Type of offending predicted	Authors	Domains
LSI-R	Act	54	Designed to use psychosocial status to predict the likelihood of general recidivism in adult offenders. The tool is designed to assist professionals make decisions regarding level of supervision and treatment	0 = item present 1 = item absent	General	Andrews and Bonta (1995)	(1) Criminal history (2) Leisure/Recreation (3) Education/Employment (4) Companions (5) Financial (6) Alcohol/Drug problem (7) Family/Marital (8) Emotional/Personal (9) Accommodation (10) Attitudes/Orientation
PCL-R	Act	20	Designed to diagnose psychopathy as operationally defined in Cleckley's (1941) *The Mask of Sanity*	0 = item does not apply 1 = item applies to a certain extent 2 = item applies	n/a (Tool not developed for the purpose of forensic risk assessment)	Hare (1991, 2003)	(1) Selfish, callous, and remorseless use of others (2) Chronically unstable and antisocial lifestyle
SORAG	Act	14	Designed to assess the likelihood of violent (including sexual) recidivism specifically in previously convicted sex offenders	n/a (Different point values awarded for different items)	Violent	Quinsey, Harris, Rice, and Cormier (1998, 2006)	n/a
Static-99	Act	10	Designed to predict the long-term probability of sexual recidivism amongst adult male offenders who have committed a sexual offense	n/a (Different point values awarded for different items)	Sexual	Hanson and Thornton (1999) Harris, Phenix, Hanson, and Thornton (2003)	n/a
VRAG	Act	12	Designed to be used to predict risk of violence in previously violent mentally disordered offenders	n/a (Different point values awarded for different items)	Violent	Quinsey, Harris, Rice, and Cormier (1998, 2006)	n/a
HCR-20	SCJ	20	Designed to assess violence risk in forensic, criminal justice, and civil psychiatric settings	0 = item not present 1 = item possibly present 2 = item definitely present	Violent	Webster, Eaves, Douglas, and Wintrup (1995) Webster, Douglas, Eaves, and Hart (1997)	(1) Historical risk factors (2) Clinical risk factors (3) Risk management factors
SVR-20	SCJ	20	Designed to predict the risk of violence (including sexual violence) in sex offenders	0 = item does not apply 1 = item possibly applies 2 = item definitely applies	Violent	Boer, Hart, Kropp, and Webster (1997)	(1) Historical risk factors (2) Social/Contextual risk factors (3) Individual/Clinical risk factors (4) Protective factors
SARA	SCJ	20	Designed to predict future violence in men arrested for spousal assault	0 = item not present 1 = item possibly present 2 = item definitely present	Violent	Kropp & Hart (1994, 1995, 1999)	(1) General violence (2) Spousal violence
SAVRY	SCJ	24	Designed to assess the risk of violence in adolescents	0 = item presents a low risk of reoffending 1 = item presents a moderate risk of reoffending 2 = item presents a high risk of reoffending	Violent	Borum, Bartel, & Forth (2002, 2003)	(1) Historical risk factors (2) Social/Contextual risk factors (3) Individual/Clinical risk factors (4) Protective factors

Note. Act = actuarial instrument; SCJ = structured clinical judgment instrument.

revealed evidence of the work having measured the predictive validity of one of the nine primary risk assessment tools. Articles were excluded if they only included select scales of a tool (e.g., Douglas & Webster, 1999) or if they were the original calibration study conducted by the authors of the tool (e.g., Hanson & Thornton, 1999). Such studies were excluded as we wished to compare the predictive validity of complete tools rather than their individual scales and to control for the potential inflation of effect size found in development samples (Blair, Marcus, & Boccaccini, 2008). In cases where the same participants were used to investigate the predictive validity of a tool in several studies, the study with the most participants was included to avoid double-counting.

The initial search identified a total of 1743 records. When the records' abstracts were scrutinized to see whether they showed evidence of the study having investigated the predictive validity of one of the nine tools of interest, the number of records was reduced to 401. Due to the use of select scales of one of the tools, being a review, or the use of overlapping samples, an additional 198 studies were excluded, leaving a total of 203 studies of interest. To be included in the meta-analysis, raw data needed to be available for a 2×2 contingency table, which presents the outcomes of risk predictions (Fig. 1) and contains the data used to calculate relevant outcome statistics (see below).

2.3.1. Binning strategies

Tools for diagnostic and prognostic clinical decision-making are used differently depending on the context: Either for screening purposes, whereby high sensitivity is strived for, or for specific case identification, whereby high specificity is required. In the literature these situations are referred to as “rule-out decisions” and “rule-in decisions”, respectively (Ransohoff & Feinstein, 1978).

Therefore, two sets of analyses were of the predictive accuracy of risk assessment instruments were conducted: The first grouped participants who were classified as low or moderate risk and compared them with those classified by the instrument as high risk. This “rule out” approach will be referred to as the high risk vs. low/moderate risk binning strategy. The second set of analyses grouped participants who were classified as moderate or high risk and compared their scores with participants who were low risk. This “rule in” approach will be referred to as the moderate/high risk vs. low risk binning strategy.

Six of the nine risk assessment tools covered by this review allow for placing scores into one of three classifications: low, moderate, or high risk of offending. The Static-99 uses four risk bins: low risk, moderate-low risk, moderate-high risk, and high risk. For the purposes of analysis, the moderate-low risk and moderate-high risk bins were combined and considered the moderate risk bin. The PCL-R dichotomously classifies individuals as being either non-psychopathic or psychopathic, and for the purposes of this study, psychopathic individuals (i.e., with scores of +30 and above) were classified as high risk and non-psychopathic individuals (i.e., with scores of +29 and below) as low risk. There is no moderate risk bin for this tool. The LSI-R uses five risk bins: low, low-moderate, moderate, moderate-high, and high. For the purposes of this study, the low and low-moderate risk bins were combined and considered the low risk bin. The moderate-high and high risk bins were also combined and considered the high risk bin.

		Outcome	
		Positive	**Negative**
Test Result	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Fig. 1. 2×2 contingency table comparing risk assessment tool predictions and outcomes.

2.3.2. Data collection

Preliminary inspection of the 203 studies revealed different score thresholds on the risk assessment tools had been used to place participants into risk bins (i.e., low, moderate, or high risk of offending). In these cases, authors were contacted and asked to complete a standardized form into which raw outcome data (i.e., true positive [TP], false positive [FP], true negative [TN], and false negative [FN] rates) could be entered at the standard threshold scores for these tools.[1] When multiple datasets were available for each sample because different tools were administered to the same participants, all datasets were included and counted as different samples. In cases where multiple indices of offending (e.g., general offending, violent offending, non-violent offending) had been used, authors were asked for outcome data using the most conservative definition of offending (i.e., the definition that would produce the highest sensitivity). We considered general offending to be the most conservative outcome followed by (in order) violent (including sexual) offending, violent (non-sexual) offending, and sexual offending.[2]

Raw data for 2×2 tables were extracted from studies which either placed each individual in low, moderate, or high risk bins according to the SCJ approach, or, for actuarial instruments, used manual-suggested score thresholds.[3] Such data was available in the manuscripts of 27 eligible studies ($k=34$). Additional data was requested from the authors of 133 studies ($k=268$) and obtained for 41 studies ($k=54$; $n=15{,}775$). For 8 studies ($k=10$), raw data was available in the manuscript for the high risk vs. low/moderate risk binning strategy but not the moderate/high risk vs. low risk binning strategy. Efforts to contact the authors of these investigations to obtain data for the second binning strategy were unsuccessful. Thus, data on 88 independent samples from 68 studies were included in the meta-analysis for the first binning strategy and data on 78 independent samples from 60 studies were included in the second binning strategy (Fig. 2).

Using Cohen's *d* values, we tested whether there were differences between the effect sizes of the included studies and studies from which we were unable to obtain tabular data. We were able to calculate *d* values for 195 of the 214 samples from which data was not available. We converted effect sizes to *d* using formulas published by Cohen (1988), Rosenthal (1994), and Ruscio (2008).

As variance parameters were not commonly reported alongside the effect sizes used to calculate *d*, pooling was deemed inappropriate. The median *d* value produced by those samples which we had data on (*Median* = 0.75; *Interquartile Range* [IQR] = 0.54–0.94) and those which we did not have data on (0.70; IQR = 0.47–0.88) were similar. Using the “cendif” command in the STATA/IC statistical software package, version 10 (StataCorp, 2007), we calculated the Hodges–Lehmann percentile difference between the median effect sizes to be 0.01 (95% CI = 0.00–0.07). The evidence of similar median effect sizes and overlapping interquartile ranges suggests that the 88 samples included in the present meta-analysis were not different (in terms of effect sizes) to the studies that were eligible but not included.

As no individual participant data was obtained, ethics approval was not required.

[1] Study authors were asked to provide TP, FP, TN, and FN rates for the overall sample as well as for male and female participants, separately. Raw data on female participants was available either in the study manuscript or obtained from authors for 7 samples and for male participants in 71 samples.

[2] Sexual offenses were considered violent offenses for the purposes of this review.

[3] When zero counts in the 2×2 table of a sample were found, a constant of +1.00 was added to each cell. Zero counts result in the inability to extract odds ratio data due to division by zero. Adding a constant allows those samples' data to be included in meta-analytic, subgroup, and metaregression analyses (Higgins, Deeks, & Altman, 2008).

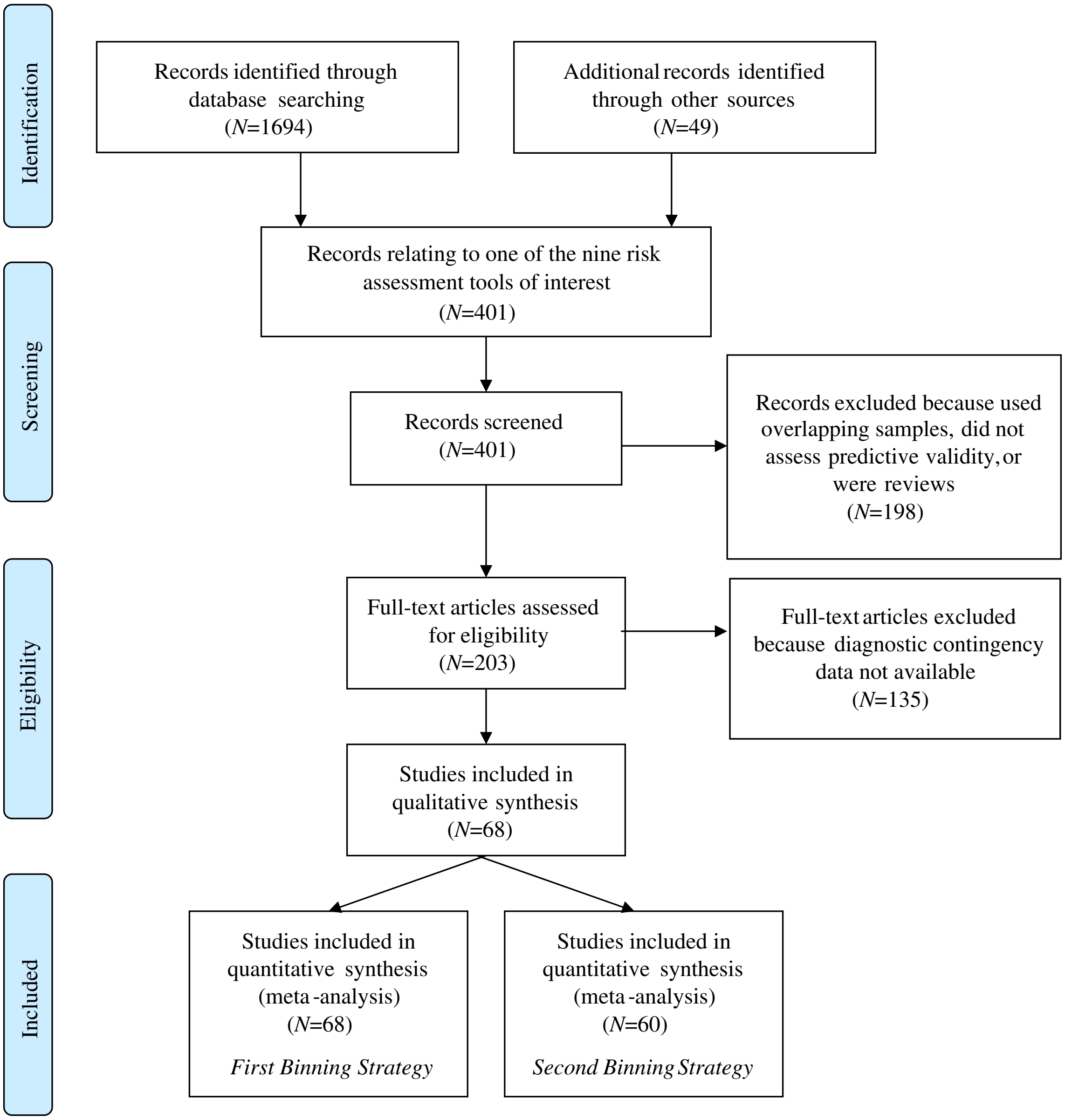
Identification
Screening
Eligibility
Included
Records identified through database searching (N=1694)
Additional records identified through other sources (N=49)
Records relating to one of the nine risk assessment tools of interest (N=401)
Records screened (N=401)
Records excluded because used overlapping samples, did not assess predictive validity, or were reviews (N=198)
Full-text articles assessed for eligibility (N=203)
Full-text articles excluded because diagnostic contingency data not available (N=135)
Studies included in qualitative synthesis (N=68)
Studies included in quantitative synthesis (meta-analysis) (N=68) First Binning Strategy
Studies included in quantitative synthesis (meta-analysis) (N=60) Second Binning Strategy

Fig. 2. Results of a systematic search conducted to identify replication studies of nine commonly used risk assessment tools.

2.4. Data extraction

We extracted 23 descriptive and demographic characteristics of the predictive validity studies. JS extracted the data. When information was unclear or seemingly conflicting, SF was consulted. When the data was missing or no consensus reached, it was coded as such.

2.5. Inter-rater reliability and quality assessment

As a measure of quality control, 10 of the included articles were randomly selected and the TP, FP, TN, and FN rates for both binning strategies calculated by a research assistant working independently of JS. The research assistant was provided with a coding manual, a coding sheet, the 10 study manuscripts, and, where appropriate, the data sheets provided by the study's authors. Cohen's (1960) kappa was measured at 1.00.

The quality of the included studies was measured using the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement (Bossuyt et al., 2003), a 25-item checklist of reporting characteristics.

2.6. Data analyses

2.6.1. Statistical appraisal of the literature

Reviews of the risk assessment literature often use a single outcome statistic to summarise their findings, commonly the area under the receiver operating characteristic curve (AUC). In addition to the AUC, commonly used primary outcome statistics in the medical diagnostic literature include the positive predictive value (PPV) and negative predictive value (NPV), and the diagnostic odds ratio (DOR) (Honest & Khan, 2002). The DOR has not, to our knowledge, been used in previous meta-analyses of the risk assessment literature (Singh & Fazel, 2010).

2.6.1.1. Area under the curve. The ROC curve plots a risk assessment tool's false positive rate against its true positive rate across score thresholds. The AUC is thus an index of how well a risk assessment tool discriminates between offenders and non-offenders across all cut-offs, and is considered by some experts to be the preferred measure of predictive accuracy (Swets, Dawes & Monahan, 2000), and a key methodological development (Mossman, 1994). However, the AUC has several major limitations. For the purposes of meta-analysis, pooling AUCs is problematic in that the statistic is not sample size dependent and, therefore, weights all effect sizes equally during statistical synthesis. Others have expressed concern that the AUC is being (mis-)used in the field such that findings are interpreted with far too much optimism (Sjöstedt & Grann, 2002). Another limitation of the AUC is that it does not allow researchers to explore sources of heterogeneity by metaregression, an increasingly important statistical approach that allows for the systematic exploration of the influence of continuous variables on effect size and adjust for the influence of one variable on another (Thompson & Higgins, 2002).

2.6.1.2. Positive predictive and negative predictive values. The PPV is the proportion of individuals who were predicted to offend who actually offended. The NPV represents the proportion of individuals who are predicted by a tool to not offend who go on to not offend. As with the AUC, using PPV or NPV does not allow researchers to use metaregression to examine sources of heterogeneity. Both statistics are limited in that they vary depending on the base rate of the outcome being predicted. What constitutes a “strong” or a “weak” PPV or NPV of a diagnostic test may differ by outcome, and hence no general guidelines exist for their interpretation.

2.6.1.3. Diagnostic odds ratio. An alternative effect size used to compare predictive validity between tests is the DOR. The DOR is the ratio of the odds of a positive test result in an offender (true positive) relative to the odds of a positive result in a non-offender (false positive). As it is a single risk estimate that is not base rate dependent, the DOR has been recommended by various methodologists for meta-analyses of diagnostic studies (Deville et al., 2002; Deeks, 2001; Glas, Lijmer, Prins, Bonsel & Bossuyt, 2003). Furthermore, the DOR allows for the systematic investigation of heterogeneity using metaregression. In addition, as researchers and many clinicians are familiar with the concept of an odds ratio, the DOR may be easier to comprehend to non-specialists than AUC.

2.6.2. Choice of primary outcome statistic

We tested assumptions for pooling data using these four outcome statistics. The DOR met the necessary assumptions for meta-analysis.[4] AUC, PPV and NPV did not meet pooling criteria. Instead of pooling these estimates, medians and interquartile ranges were calculated. DORs were pooled using the standard Mantel–Haenszel method of meta-analysis of risk ratios. Meta-analysis was performed using the DOR data and the “metan” command in the STATA/IC statistical software package, version 10 (StataCorp, 2007).

2.6.3. Risk assessment tool ranking

To assess which risk assessment tool produced the highest level of predictive validity across outcome statistics, we devised a ranking system combining these estimates. First, the tools were ordered from worst to best with regard to their DORs, AUCs, PPVs, and NPVs. The measures were then given a score of one (poorest performance) through nine (strongest performance) based on their ranking within an outcome statistic. This procedure was repeated for both binning strategies.[5] The scores for each instrument across outcome statistics were then summed and used as a summary performance score to determine the strongest and weakest tools.

2.6.4. Investigation of heterogeneity

Heterogeneity among the samples was estimated using the I^2 statistic (Higgins, Thompson, Deeks & Altman, 2003). I^2 describes the percentage of variation across samples due to heterogeneity rather than chance alone (Higgins & Thompson, 2002; Higgins et al., 2003). As such, a resulting I^2 of 0 implies that all variability in effect estimates is produced by sampling error rather than between-study heterogeneity. As the I^2 percentage increases, so too does the proportion of effect size variability that is due to between-study heterogeneity. To investigate sources of heterogeneity, subgroup and metaregression analyses were performed using the DORs of the included samples.

2.6.4.1. Subgroup analyses and metaregression. Random effects subgroup analyses were conducted for all variables of interest as we wished to make inferences extending beyond the included samples. When $I^2<75\%$ for the effect estimates of both variable subgroups, fixed effects analyses were also conducted (Higgins et al., 2003). Random effects metaregression was used to investigate whether different sample or study characteristics were associated with the predictive validity of the risk assessment tools. In addition, a test of independence on those variables which were found to be significant at the $p<0.05$ level was used to assess whether these variables influenced effect size independently of one another. Analyses were performed using the DOR data and the “metan” and “metareg” commands in STATA/IC, version 10 (StataCorp, 2007).

2.6.4.2. Variables of primary interest. Variables of primary interest included gender, ethnicity, age, and type of risk assessment tool. Gender, ethnicity, and mean participant age were investigated using both categorical and continuous data. The influence of the gender composition of samples was analyzed as both a categorical variable (i.e., male sample data vs. female sample data for the subgroup analysis and 100% female vs. <100% female for metaregression) and continuously (i.e., percentage of study sample which was male). The ethnic composition of samples was also analyzed categorically (i.e., <90% White vs. ≥90% White) and continuously (i.e., percentage of study sample which was White). The mean age of participants in a sample was analyzed both as a dichotomous variable (i.e., less than 25 years of age vs. 25 years of age and older)[6] and as a continuous variable (in years). Finally, type of risk assessment tool was analyzed as a dichotomous variable (i.e., actuarial vs. SCJ).

2.6.4.3. Variables of secondary interest. Secondary variables of interest included study design, length of follow-up, study setting, sample size, country of origin, and type of offending. Study design was included as a dichotomous variable (i.e., prospective vs. retrospective). Mean length of follow-up was investigated in both categorical (i.e., ≤2 years vs. >2 years) and continuous (in months) forms. Study setting was explored in two ways: institutional setting (i.e., prison or psychiatric unit) vs. community setting, and prison setting vs. psychiatric setting. Sample size was included as both a dichotomous variable (i.e., <500 participants vs. ≥500 participants) and as a continuous variable. Also included in the analyses were the dichotomous variables of country of origin (i.e., study conducted in North America vs. Europe) and type of offending being predicted (i.e., general vs. violent offending). We intend to report our findings on publication bias separately.

[4] To pool samples' DORs, individual DORs must follow a symmetrical ROC curve (Deeks, 2001). Using Moses, Littenberg and Shapiro's (1993) regression test, which plots a measure of threshold against the natural log of the DOR for each sample, we found a non-significant variation in diagnostic performance with threshold for both the high risk versus low/moderate risk binning strategy ($\beta=0.01$, $p=0.98$) as well as the moderate/high risk versus low risk binning strategy ($\beta=$-0.01, $p=0.28$). This non-significant finding meant that a symmetrical ROC curve was generated by the data for each binning strategy and, therefore, sample DORs could be pooled.

[5] As the PCL-R classifies participants into two risk bins (i.e., psychopathic versus non-psychopathic) rather than three (i.e., low risk, moderate risk, and high risk), it did not produce different effect estimates for the two binning strategies. Therefore, the pooled DOR and median AUC, PPV, and NPV for the PCL-R were identical in both binning strategies.

[6] The traditional age cut-off of 18 years was not used as only SAVRY samples had a mean participant age below 18.

3. Results

3.1. Descriptive characteristics

Information was collected on 25,980 participants in 88 independent samples from 68 studies. Information from 54 ($n = 15,775$; 60.7%) of the samples was specifically obtained from study authors for the purposes of this synthesis. Included studies were conducted in 13 countries: Argentina ($n = 199$; 0.8%), Austria ($n = 799$; 3.1%), Belgium ($n = 732$; 2.8%), Canada ($n = 9,112$; 35.1%), Denmark ($n = 304$; 1.2%), Finland ($n = 208$; 0.8%), Germany ($n = 1,337$; 5.1%), The Netherlands ($n = 622$; 2.4%), New Zealand ($n = 220$; 0.8%), Spain ($n = 276$; 1.1%), Sweden ($n = 2,204$; 8.5%), the UK ($n = 3,929$; 15.1%), and the USA ($n = 6,038$; 23.2%). Of the 25,980 participants in the included samples, 8,155 (31.4%) went on to offend.

The tools with the most samples included the PCL-R ($k = 20$; 22.7%), the Static-99 ($k = 14$; 15.9%), and the VRAG ($k = 12$; 13.6%; Table 2). The majority of the samples ($k = 61$; 69.3%) were assessed using an actuarial tool. Of the 88 included samples, 64 (72.7%) were in studies published in peer-reviewed journals. Prospective research methodology was used with 40 (45.5%) samples, whereas 43 (48.9%) samples were investigated retrospectively. Regarding study setting, 4 (4.5%) samples consisted of community persons, 37 (42.0%) of prisoners, 33 (37.5%) of psychiatric patients, and 14 (15.9%) of participants from a mixture of different settings. Of those 33 samples in psychiatric settings, 3 (3.4%) consisted of civil patients and 30 (34.1%) of forensic patients. Study participants were selected using a specified sampling method in 52 (59.1%) samples. The mean length participants were followed up was 56.3 ($SD = 41.3$) months. Regarding type of offending, general offending was the outcome criteria in 46 (52.3%) samples as opposed to violent offending in 40 (45.5%) samples, or non-violent offending in 1 sample (1.1%).

Table 2
Descriptive and demographic characteristics of 88 samples investigating the predictive validity of risk assessment tools.

Category	Subcategory	Group	Number of $k = 88$ (%)
Study information	Source of study	Journal article	64 (72.7)
		Conference	7 (8.0)
		Dissertation	13 (14.8)
		Government report	4 (4.5)
	Study location	United States of America	14 (15.9)
		Canada	31 (35.2)
		United Kingdom	9 (10.2)
		Other European Union	31 (35.2)
		Latin America	2 (2.3)
		Australia/New Zealand	1 (1.1)
Tool information	Type of tool	Actuarial	61 (69.3)
		SCJ	27 (30.7)
	Tool used	HCR-20	9 (10.2)
		LSI-R	8 (9.1)
		PCL-R	20 (22.7)
		SARA	4 (4.5)
		SAVRY	9 (10.2)
		SORAG	7 (8.0)
		SVR-20	5 (5.7)
		Static-99	14 (15.9)
		VRAG	12 (13.6)
Study methodology	Study design	Prospective	40 (45.5)
		Retrospective	43 (48.9)
		Unstated/unclear	5 (5.7)
	Study setting	Community	4 (4.5)
		Prison	37 (42.0)
		Psychiatric unit	33 (37.5)
		Mixed	14 (15.9)
Outcome characteristics	Length of follow-up (months)	Mean (SD)	56.3 (41.3)
	Type of outcome	Recharge/rearrest/reconviction	56 (63.6)
		Institutional incident	11 (12.5)
		Mixed	16 (18.2)
		Unstated/unclear	5 (5.7)
	Type of offending	General (violent and non-violent)	46 (52.3)
		Violent only	40 (45.5)
		Non-violent only	1 (1.1)
		Unstated/unclear	1 (1.1)
Sample size	Mean (SD)		296 (422)
Age	Mean age in years (SD)		31.6 (7.6)
White ethnic background	Mean number of participants per sample (SD)		138 (118)
Male sex	Mean number of participants per sample (SD)		295 (429)

Note. SCJ = structured clinical judgment instrument.

The average study included in the meta-analysis fulfilled 17 ($SD = 3$) of the 25 STARD criteria. In the 68 studies, the criteria which were least commonly fulfilled included specifying sample selection criteria ($N = 28$; 41.2%), reporting the number of and training of coders ($N = 28$; 41.2%), detailing why some participants were excluded or did not complete follow-up ($N = 28$; 41.2%), how indeterminate results, missing responses, and outliers were handled ($N = 25$; 36.8%), and whether there were adverse effects from the testing procedures ($N = 0$; 0.0%). There was no evidence of an association between STARD score and sample DOR in either binning strategy ($\beta = -0.03$; $p = 0.36$; $\beta = -0.01$; $p = 0.87$, respectively).

3.2. Demographic characteristics

The mean sample size was 296 ($SD = 422$) participants (Table 2). The mean age of participants was 31.6 ($SD = 7.6$) years. Data on participants' ethnic backgrounds was available for 38.6% ($k = 34$) of samples and showed a trend towards predominantly White samples. Of the included samples, 28 (31.8%) were composed of over 50% White participants. An average of 138 ($SD = 118$) White participants were included in each sample. Data on gender composition was available for 93.2% ($k = 82$) of samples, with a trend towards predominantly male samples. Of the included samples, 80 (90.9%) were composed of over 50% male participants. On average, 295 ($SD = 429$) male participants were included in each sample. Participant psychiatric diagnosis data was available for 23 (26.1%) samples. Personality disorders were the most commonly reported diagnoses ($n = 1,333$; 5.1%) followed by psychotic disorders ($n = 697$; 2.7%), drug or alcohol abuse or dependency ($n = 629$; 2.4%), mood disorders ($n = 127$; 0.5%), conduct disorder ($n = 26$; 0.1%), and anxiety disorders ($n = 5$; 0.1%).

3.3. Median area under the curve

The risk assessment tools with the highest median AUCs were the SVR-20 (0.78; $IQR = 0.71–0.83$), the SORAG (0.75, $IQR = 0.69–0.79$), and the VRAG (0.74. $IQR = 0.74–0.81$). See Table 3 for a summary of these results.

Table 3
Median area under the curve produced by nine risk assessment tools ranked in order of strength.

Tool	*n*	*k*	Median AUC	*IQR*
SVR-20	380	3	0.78	0.71–0.83
SORAG	1599	6	0.75	0.69–0.79
VRAG	2445	10	0.74	0.74–0.81
SAVRY	915	8	0.71	0.69–0.73
HCR-20	1320	8	0.70	0.64–0.76
SARA	102	1	0.70	–
Static-99	8246	12	0.70	0.62–0.72
LSI-R	856	3	0.67	0.55–0.73
PCL-R	2645	10	0.66	0.54–0.68

Note. n = sample size; k = number of samples; AUC = area under the curve; IQR = interquartile range.

3.4. Median positive predictive and negative predictive values

When we analyzed the data using the high vs. low/moderate risk binning strategy, the risk assessment tools with the highest median PPVs were the SAVRY (0.76; *IQR* = 0.42–0.85), the HCR-20 (0.71; *IQR* = 0.55–0.85), and the VRAG (0.66; *IQR* = 0.37–0.79) (Table 4, upper panel). The three tools with the highest median NPVs were the Static-99 (0.82; *IQR* = 0.71–0.94), the SARA (0.79; *IQR* = 0.67–0.92), and the SAVRY (0.76; *IQR* = 0.49–0.91).

The moderate/high vs. low risk binning strategy found that the risk assessment tools with the highest median PPVs were the HCR-20 (0.64; *IQR* = 0.45–0.70), the SAVRY (0.60; *IQR* = 0.27–0.73), and the PCL-R (0.52; *IQR* = 0.45–0.75) (Table 4, lower panel). The risk assessment tools with the highest median NPVs were the SARA (0.96; *IQR* = 0.82–0.98), the Static-99 (0.95; *IQR* = 0.76–0.99), and the SAVRY (0.90; *IQR* = 0.74–0.95).

3.5. Pooled diagnostic odds ratios

When the data from the high vs. low/moderate risk binning strategy was analyzed, the three tools with the highest pooled DORs were the SAVRY (6.93; 95% CI = 4.93–9.73), the VRAG (3.84; 95% CI = 2.85–5.16), and the HCR-20 (3.48; 95% CI = 2.62–4.62) (Table 5, upper panel). When the moderate/high vs. low risk binning strategy was used, the three tools with the highest pooled DORs were the SARA (7.87; 95% CI = 3.12–19.87), the SAVRY (6.40; 95% CI = 4.40–9.32), and the SORAG (5.54; 95% CI = 4.09–7.50) (Table 5, lower panel).

3.6. Risk assessment tool comparison

Using our ranking system which collated results from median AUC, PPV, NPV and pooled DOR, the SAVRY produced the highest rates of overall predictive validity (Table 6). The VRAG and the SARA also produced high rates. Ranked lowest were the LSI-R and the PCL-R. In both binning strategies, the pooled DOR, alone, accurately identified the three most and least accurate risk assessment tools.

3.7. Heterogeneity

We conducted subgroup and metaregression analyses for both binning strategies using sample DORs (Tables 7 and 8, respectively). For a summary of these results, see Table 9.

Table 4
Median positive predictive and negative predictive values produced by risk assessment tools ranked in order of strength.

Binning	Tool	*n*	*k*	Median (PPV)	*IQR* (PPV)	Median (NPV)	*IQR* (NPV)
First	SAVRY	1026	9	0.76	0.42–0.85	0.76	0.49–0.91
First	HCR-20	1374	9	0.71	0.55–0.85	0.67	0.51–0.70
First	VRAG	2703	12	0.66	0.37–0.79	0.74	0.62–0.84
First	LSI-R	4005	8	0.57	0.44–0.70	0.53	0.44–0.65
First	SARA	2305	4	0.53	0.31–0.62	0.79	0.67–0.92
First	PCL-R	3854	20	0.52	0.45–0.75	0.68	0.39–0.82
First	SORAG	1637	7	0.38	0.33–0.86	0.64	0.59–0.90
First	Static-99	8555	14	0.33	0.18–0.56	0.82	0.71–0.94
First	SVR-20	521	5	0.33	0.16–0.60	0.65	0.56–0.87
Second	HCR-20	1320	8	0.64	0.45–0.70	0.80	0.71–0.86
Second	SAVRY	1026	9	0.60	0.27–0.73	0.90	0.74–0.95
Second	PCL-R	3854	20	0.52	0.45–0.75	0.68	0.39–0.82
Second	LSI-R	4005	8	0.48	0.24–0.67	0.57	0.54–0.75
Second	SORAG	1599	6	0.40	0.19–0.71	0.88	0.77–0.96
Second	VRAG	2602	11	0.39	0.23–0.54	0.89	0.83–0.92
Second	SARA	465	3	0.37	0.10–0.60	0.96	0.82–0.98
Second	SVR-20	268	3	0.23	0.18–0.72	0.71	0.53–0.78
Second	Static-99	8097	10	0.18	0.08–0.31	0.95	0.76–0.99

Note. First = high risk vs. low/moderate risk binning strategy; Second = moderate/high risk vs. low risk binning strategy; *n* = sample size; *k* = number of samples; *IQR* = interquartile range; PPV = positive predictive value; NPV = negative predictive value.

Table 5
Pooled diagnostic odds ratios for risk assessment tools ranked in order of strength.

Binning	Model	Tool	*n*	*k*	*DOR* (95% CI)
First	FE	SAVRY	1026	9	6.93 (4.93–9.73)
First	FE	VRAG	2703	12	3.84 (2.85–5.16)
First	FE	HCR-20	1374	9	3.48 (2.62–4.62)
First	FE	SARA	2305	4	3.42 (2.72–4.29)
First	RE	Static-99	8555	14	3.12 (1.94–5.02)
First	FE	SORAG	1637	7	2.52 (2.15–2.96)
First	RE	PCL-R	3854	20	2.08 (1.14–3.81)
First	RE	LSI-R	4005	8	1.75 (0.96–3.22)
First	RE	SVR-20	521	5	1.56 (0.36–6.84)
Second	FE	SARA	465	3	7.87 (3.12–19.87)
Second	FE	SAVRY	1026	9	6.40 (4.40–9.32)
Second	FE	SORAG	1599	6	5.54 (4.09–7.50)
Second	FE	VRAG	2602	11	5.21 (3.61–7.53)
Second	FE	HCR-20	1320	8	4.90 (3.65–6.56)
Second	FE	Static-99	8097	10	2.95 (2.38–3.66)
Second	RE	PCL-R	3854	20	2.08 (1.14–3.81)
Second	RE	LSI-R	4005	8	1.26 (0.77–2.06)
Second	RE	SVR-20	268	3	1.21 (0.18–8.32)

Note. First = high risk vs. low/moderate risk binning strategy; Second = moderate/high risk vs. low risk binning strategy; RE = random effects model; FE = fixed effects model (where I^2 < 75%); *n* = sample size; *k* = number of samples; *DOR* = diagnostic odds ratio.

3.7.1. High risk vs. low/moderate risk binning strategy

3.7.1.1. Non-significant findings. Subgroup and metaregression analyses of the sample DOR data using the high risk vs. low/moderate risk binning strategy data resulted in non-significant findings for the following variables: gender composition, ethnic composition, type of risk assessment tool, prospective vs. retrospective methodology, mean length of follow-up,[7] study setting, sample size, and country of origin.

3.7.1.2. Gender composition. A sensitivity analysis was conducted to further explore gender effects. As women-only data was available for the HCR-20, SAVRY, VRAG, and LSI-R, we excluded data on all other instruments for the men. Hence, outcomes were investigated in male and female data using the same set of instruments. The pooled DOR of these risk assessment tools for men was 3.83 (95% CI = 2.13–6.87), and for women was 4.66 (95% CI = 2.87–7.55). The I^2 for the men-only data was 82.1% and for the women-only data was 12.6%.

3.7.1.3. Mean age. Neither subgroup nor metaregression analyses found that predictive validity estimates varied with mean participant age. As all SAVRY samples (k = 9) had mean participant ages below 25 and as this instrument produced the highest predictive validity estimates across outcome measures, a sensitivity analysis was conducted in which SAVRY samples were excluded. When SAVRY samples were excluded, metaregression analysis of mean age as a continuous variable found a significant trend: the higher the mean age of a sample, the higher the DOR (β = 0.09, p = 0.02). To further investigate this finding, subgroup analyses were conducted dividing mean participant age into three groups (<25 years, 25–40 years, and >40 years). The summary random effects DORs were 0.79 (95% CI = 0.16–3.95), 2.86 (95% CI = 2.12–3.86), and 4.01 (95% CI = 3.00–5.35), respectively.

3.7.1.4. Type of offending. Random effects subgroup analysis revealed that samples reporting violent offending produced significantly higher DORs than samples investigating general offending. Metaregression analysis confirmed this finding (β = 0.81, p = 0.01).

[7] Ancillary subgroup and metaregression analyses were conducted on the dichotomous form of this variable in both binning strategies using cut-off periods of 1, 3, 4, 5, and 10 years. None of these comparisons yielded a significant difference.

Table 6
Summary performance scores of risk assessment tools across four outcome statistics.

	Summary performance score		
Tool	First binning strategy	Second binning strategy	Total
SAVRY	31	29	60
VRAG	28	23	51
SARA	22	25	47
HCR-20	23	23	46
SORAG	20	25	45
Static-99	18	16	34
SVR-20	15	15	30
PCL-R	13	13	26
LSI-R	11	11	22

Note: First binning strategy = high risk vs. low/moderate risk binning strategy; Second binning strategy = moderate/high risk vs. low risk binning strategy. The four outcome statistics included diagnostic odds ratio, positive predictive and negative predictive values, and area under the curve. Summary performance scores were calculated by ordering tools from poorest to strongest performance on each effect estimate. Each tool was assigned a score of +1 (poorest performance) through +9 (strongest performance) on each outcome statistic. These values were then summed for each tool, yielding a composite performance score.

Table 7
Subgroup analysis investigating sources of heterogeneity in replication samples of nine risk assessment tools.

	Diagnostic odds ratio (95% CI)	
Sample or study characteristic	First binning strategy	Second binning strategy
Gender composition		
Male sample data	3.15 (2.48–4.00)	3.35 (2.44–4.59)
Female sample data	4.66 (2.87–7.55) [b]	5.30 (3.08–9.11) [b]
Ethnic composition		
<90% White	3.59 (2.38–5.42)	3.86 (2.50–5.96)
≥90% White	2.02 (1.03–3.94)	3.20 (1.08–9.48)
Mean age of participants [c]		
<25 years	4.40 (1.99–9.70)	4.51 (2.13–9.54)
≥25 years	3.26 (2.50–4.24)	3.21 (2.39–4.30)
Type of risk assessment tool		
Actuarial	2.88 (2.21–3.75)	2.77 (2.08–3.69)
Structured clinical judgment	4.01 (2.81–5.71)	4.15 (2.42–6.75)
Study setting		
Institution [d]	3.36 (2.58–4.39)	3.47 (2.56–4.70)
Community	3.13 (1.39–7.06)	3.00 (1.21–7.41)
Prison	3.65 (2.56–5.22)	3.20 (2.19–4.67)
Psychiatric unit	2.96 (1.96–4.45)	4.05 (2.59–6.32)
Study design		
Prospective	3.14 (2.31–4.28)	3.16 (2.19–4.57)
Retrospective	3.40 (2.49–4.65)	3.58 (2.63–4.90)
Mean length of follow-up		
≤2 years	2.80 (1.68–4.65)	2.16 (1.32–3.54)
>2 years	3.56 (2.73–4.65)	4.01 (3.10–5.18)
Sample size		
<500 participants	3.34 (2.67–4.18)	3.52 (2.72–4.56)
≥500 participants	2.35 (1.30–4.25)	2.21 (1.26–3.87)
Country of origin		
North America	2.78 (2.11–3.67)	3.27 (2.38–4.50)
Europe	3.85 (2.76–5.38)	3.49 (2.35–5.18)
Type of offending		
General	2.52 (1.90–3.36) [a]	2.54 (1.81–3.56) [a]
Violent	4.52 (3.59–5.70) [a]	4.78 (3.80–6.01) [a]

Note. First binning strategy = high risk vs. low/moderate risk binning strategy; Second binning strategy = moderate/high risk vs. low risk binning strategy.
[a] Non-overlapping confidence intervals.
[b] Results of fixed effects analysis.
[c] SAVRY samples included.
[d] Prison or psychiatric unit.

Table 8
Metaregression analysis investigating sources of heterogeneity in replication samples of nine risk assessment tools.

	First binning strategy		Second binning strategy	
Sample or study characteristic[b]	β (SE)	p	β (SE)	p
Gender composition				
Dichotomous	0.10 (0.40)	0.81	−0.05 (0.38)	0.89
Continuous	−0.02 (0.01)	0.10	−0.02 (0.01)	0.19
Ethnic composition				
Dichotomous	−0.10 (0.63)	0.87	0.11 (0.60)	0.85
Continuous	0.01 (0.21)	0.21	0.02 (0.01)	**0.04** [a]
Mean age of participants [c]				
Dichotomous	−0.49 (0.38)	0.20	−0.32 (0.40)	0.43
Continuous	−0.02 (0.02)	0.26	−0.01 (0.02)	0.55
Type of risk assessment tool				
Actuarial vs. SCJ	0.40 (0.30)	0.18	0.35 (0.27)	0.13
Study design				
Prospective vs. retrospective	0.10 (0.29)	0.74	0.07 (0.28)	0.81
Mean length of follow-up				
Dichotomous	0.13 (0.31)	0.69	0.48 (0.29)	0.14
Continuous	−0.01 (0.01)	0.61	−0.01 (0.01)	0.85
Study setting				
Institution [d] vs. community	−0.45 (0.50)	0.38	−0.35 (0.56)	0.54
Prison vs. psychiatric unit	0.02 (0.36)	0.95	0.22 (0.34)	0.52
Sample size				
Dichotomous	−0.48 (0.40)	0.24	−0.40 (0.40)	0.32
Continuous	−0.01 (0.01)	0.28	−0.01 (0.01)	0.32
Country of origin				
North America vs. Europe	0.38 (0.28)	0.17	0.15 (0.28)	0.59
Type of offending				
General vs. violent	0.81 (0.21)	**0.01** [a]	0.57 (0.21)	**0.01** [a]

Note. First binning strategy = high risk vs. low/moderate risk binning strategy; Second binning strategy = moderate/high risk vs. low risk binning strategy; SCJ = structured clinical judgment instrument.
[a] $p<0.05$.
[b] Dichotomous variables same as subgroup analyses for ethnicity, age, follow-up, and sample size.
[c] SAVRY samples included.
[d] Prison or psychiatric unit.

3.7.2. Moderate/High risk vs. low risk binning strategy

3.7.2.1. Non-significant findings. Subgroup and metaregression analyses of the moderate/high risk vs. low risk binning strategy data found that the following variables did not significantly influence effect size: gender composition, type of risk assessment tool, prospective vs. retrospective methodology, mean length of follow-up,[7] study setting, sample size, and country of origin.

3.7.2.2. Gender composition. A sensitivity analysis was conducted using only the HCR-20, SAVRY, VRAG, and LSI-R sample data. Non-significant increases in DOR were found for women as opposed to men using subgroup analysis (DOR for men = 3.21 [95% CI = 1.61–6.39] vs. DOR for women = 5.30 [95% CI = 3.08–9.11]). The I^2 for the men-only data was 90.7% and for the women-only data was 0.0%.

3.7.2.3. Ethnic composition. Subgroup analysis revealed that samples with less than 90% White individuals did not produce different DORs than samples of 90% or more White individuals. Using metaregression, the higher the proportion of White individuals in a sample, the higher the resulting DOR ($\beta = 0.02$, $p = 0.04$).

3.7.2.4. Mean age of participants. As in the first binning strategy, neither subgroup analysis nor metaregression revealed that DORs varied with mean participant age. When SAVRY samples were excluded, a significant trend was found such that the higher the mean age of participants in a sample, the higher the DOR ($\beta = 0.08$, $p = 0.04$). In addition, post hoc subgroup analyses were conducted on three mean participant age bands (<25 years, 25–40 years, and>40 years).

The summary random effects DORs were 0.85 (95% CI = 0.16–4.54), 2.73 (95% CI = 1.99–3.74), and 4.85 (95% CI = 2.86–8.21), respectively.

3.7.2.5. Type of offending. Random effects subgroup analysis found that samples which used violent offending as their outcome produced higher DORs than samples which used general offending. Metaregression analysis confirmed this finding ($\beta = 0.57$, $p = 0.01$).

3.7.3. Multivariate metaregression

Metaregression was carried out on all variables which were significant at the $p < 0.05$ level to determine which, if any, produced effects independently of one another. For the high vs. low/moderate risk binning strategy, only type of offending satisfied these conditions, so no multivariate metaregression analyses were conducted. When SAVRY data was excluded and both mean age of participants (continuous) and type of offending were regressed together, multivariate metaregression revealed that mean participant age ($\beta = -0.96$, $p = 0.03$) remained a significant predictor of sample DOR.

For the moderate/high vs. low risk binning strategy, variables at the $p < 0.05$ level included: ethnic composition (continuous) and type of offending. Using multivariate metaregression, both ethnic composition ($\beta = 0.02$; $p = 0.03$) and type of offending ($\beta = 0.02$; $p = 0.04$) remained significant predictors of DOR. When SAVRY data was excluded and mean age (continuous) was added to the metaregression model, only ethnic composition ($\beta = 0.02$; $p = 0.04$) remained significant.

4. Discussion

This systematic review and meta-analysis investigated the predictive validity of nine commonly used risk assessment tools: HCR-20, LSI-R, PCL-R, SARA, SAVRY, SORAG, Static-99, SVR-20, and VRAG. We collected data from 68 studies constituting 88 independent samples. These samples included a total of 25,980 participants from 13 countries. Information on 61% of the participants was specifically obtained from study authors for the purposes of this synthesis. We investigated three main research questions: (1) are there differences between the predictive validity of risk assessment tools, (2) what demographic factors are associated with higher and lower rates of predictive validity in risk assessment, and (3) do actuarial and clinically based risk measures produced different rates of predictive validity.

A central finding of the present meta-analysis was that there are substantial differences between the predictive validity of these tools. These differences were found in three of the four outcome statistics we used. For example, pooled DORs varied from 1.2 to 7.9. This suggests that it may in fact matter what instrument is used for violence risk assessment.

4.1. Comparative performance of risk assessment tools

The risk assessment tool that produced the highest rate of predictive validity varied slightly depending on which outcome statistic was used. Therefore, we developed a ranking system that collated performance on four outcome statistics to identify the most and least accurate measures. Overall, we found that instruments designed to assess violence risk in specific populations produced higher rates of predictive validity than tools designed for more general populations. Hence, the SAVRY, an instrument designed to assess the risk of violence in adolescents, produced the highest rates of predictive validity across outcome statistics in both binning strategies. The LSI-R, a tool which was designed to predict the likelihood of general offending in adult offenders, and the PCL-R, a clinical rating scale that was not designed for the purposes of forensic risk assessment, produced the lowest rates of predictive validity. While these two measures may be administered to a broad range of participants, this appears to come at the cost of predictive accuracy. The present meta-analysis would therefore argue against the view of some experts that the PCL-R is unparalleled in its ability to predict future offending (e.g., Salekin, Rogers & Sewell, 1996).

The finding that risk assessment tools designed for more specific populations produce higher rates of predictive validity is supported by another main finding in the current report: instruments were better at detecting risk of violent offending than general offending. Future research could examine whether tools investigating more specific forms of violent offending produce even higher rates of predictive validity. The results of the present review suggest that the future development of risk assessment tools could take the direction of designing measures for specific populations or specific forms of offending.

The finding that the SAVRY has the highest rates of predictive validity may be partly due to replication samples for the SAVRY all having been conducted on adolescent offenders, the population with which the tool was validated, unlike other tools where replication samples were more varied. In addition, as the SAVRY is amongst the most specific (designed for use with juveniles) and thorough of the included tools (containing 24 items), researchers may have been more attentive to using the instrument according to the protocol set forth by its authors. Despite these reservations, we would contend that, currently, the SAVRY should be routinely used when assessing violence risk in adolescents.

4.1.1. Demographic factors

Our review found some potentially interesting differences in the predictive validity of risk assessment tools according to age, ethnicity, and gender (Table 9). The most consistent finding was that older age was associated with higher rates of predictive validity. This finding is not surprising in view of the fact that the mean age of the samples was 32 years, and many of these instruments were developed in released prisoners who would have been in their late-twenties and thirties. A second finding was that there was some evidence that validity was better in those samples comprising mainly White participants. Again, most of these risk assessment tools were calibrated on samples of White participants, so this is not unexpected.

One of the implications of the ethnicity and age findings is that caution is warranted when using these tools to predict offending in samples dissimilar to their validation samples. A recent study by Dernevick, Beck, Grann, Hogue and McGuire (2010) is consistent with this view. The researchers found that instruments designed to detect recidivism risk in general offender populations performed poorly

Table 9
Summary of results from examining sources of heterogeneity.

Significant difference [a]	Evidence of trends [b]	No significant difference [c]
Mean age of participants [d] Type of offending	Ethnic composition [d]	Gender composition Actuarial vs. SCJ Study design Mean length of follow-up Study setting Sample size Country of origin

Note. SCJ = structured clinical judgment. Unless otherwise noted, the classification of a variable which was analyzed in both dichotomous and continuous forms concerned both its variations. Significance level set at p = 0.05"
[a] Variables significant in both binning strategies using subgroup or metaregression analyses.
[b] Variables significant in one binning strategy using subgroup or metaregression analyses.
[c] Variables not significant in either binning strategy using subgroup or metaregression analyses.
[d] Significant only in continuous form of variable.

when used to predict misconduct in terrorists (Dernevick et al., 2010). In addition, an investigation based on all sex offenders leaving prisons in Sweden found important differences in factors associated with reoffending by 10 year age bands, particularly in those aged over 60 (Fazel, Sjöstedt, Långström & Grann, 2006).

Although we found some evidence that risk assessment tools have higher rates of predictive validity in women than in men, the data on women was based on 7 (8.0%) samples and hence should be interpreted with considerable caution. Future research should present predictive validity estimates for men and women separately (in addition to overall effect sizes).

4.1.2. Type of risk assessment

Our study found no evidence that, compared with SCJ tools, actuarial instruments produced better levels of predictive validity. This finding suggests that clinicians and researchers could focus on identifying which measure, actuarial or not, produces the highest rate of predictive validity for their population and setting of interest. Additional considerations when choosing a risk measure may include the additional costs of training and materials, ease of use, and whether a tool is useful in making decisions regarding effective treatment and risk management. The latter has been considered a primary strength of the SCJ approach (Douglas, Cox, & Webster, 1999; Heilbrun, 1997). The relative utility of actuarial and clinically based tools may, however, be different for certain subgroups, such as sexual offenders. Meta-analyses on the accuracy of risk assessment tools for sexual offenders (Hanson & Morton-Bourgon, 2004, 2007, 2009) have found that actuarial instruments outperform measures which employ structured clinical judgment. To be included in these three reviews, studies had to include only sexual offenders. The present meta-analysis included samples of offenders regardless of their index offense, making it more representative of the criminal population. Future meta-analyses could investigate predictive validity estimates for more specific forms of offending in more specific populations. Finally, in 5 (18.5%) of the 27 samples which were administered an SCJ tool, the instrument was used in an actuarial manner (i.e., cut-off scores used to place an individual into a low, moderate, or high risk bin rather than relying on clinical judgment). This finding suggests that researchers should aim to use SCJ tools as they were designed.

4.1.3. Re-evaluating the single effect estimate of choice

One implication of the present meta-analysis concerns the utility of the AUC, the effect estimate currently preferred when measuring predictive accuracy (Swets et al., 2000). As the assumptions of pooling were not met, it was not possible to statistically pool study-specific AUCs. We found that comparing tools by median AUC was not useful as the interquartile ranges of all nine instruments overlapped, and thus, the statistic was unable to discriminate between the tools. Furthermore, the utility of the AUC is limited as it only allows for subgroup analysis to investigate sources of heterogeneity (if pooling is possible) rather than metaregression. The latter method enables researchers to explore the moderating role of continuous variables on effect size and adjust for the effects of one variable on another.

When our ranking system was used to compare tools across all four outcome statistics, the DOR data identified the risk assessments with the highest and lowest rates of predictive validity, suggesting that DOR can be used effectively to discriminate between the diagnostic accuracy of different measures. Further, DOR allows researchers to investigate sources of heterogeneity using both subgroup and metaregression analyses. Given this advantage and the statistic's ease of use, researchers should consider including DORs in future studies concerning risk assessment.

4.2. Limitations

One limitation of the present meta-analysis is that we were unable to obtain sample data from all eligible studies. However, we found evidence that the effect sizes produced by the included studies were similar to those produced by the studies which we were unable to include. As a consequence of not being able to include all eligible studies, there was insufficient statistical power to examine sources of heterogeneity by individual instrument. Future of reviews could attempt to improve the inclusion of more primary data. Initiatives to promote the registering of primary observational data might assist in this (Editorial, 2010).

A second limitation of the meta-analysis is that we grouped some outcomes together as we thought this reflected clinical practice. Clinical assessment is more likely to be interested in risk of any serious offending (rather than separating out risks for sexual and violent offending). Nevertheless, some of the instruments included were designed specifically for sexual offenders, and future work should investigate whether they have different rates of predictive validity when the outcome is purely sexual offending.

Another possible limitation is that we included studies regardless of their methodological quality in order to analyze a representative sample of the literature. To address this limitation, we conducted a thorough investigation of sources of within-study heterogeneity that included traditional markers of methodological quality such as prospective or retrospective design and length of follow-up. The study characteristics chosen for investigation were identified *a priori*. Nevertheless as we conducted a number of tests of heterogeneity and analyzed the data using several outcome statistics, the threshold for statistical significance needs to take this into account and caution is warranted if some of the significant findings are taken on their own without considering other potentially relevant clinical factors that risk assessment instruments do not measure. We attempted to address this limitation by summarizing the findings in relation to different thresholds of significance (Table 9).

The studies included in the present meta-analysis were conducted in 13 countries. As legal systems and rates of resolved criminality differ between countries, the predictive validity estimates produced by risk assessment tools may differ between nations. We have attempted to address this by comparing the rates of predictive validity produced by studies conducted in North America compared with those conducted in Europe. Our findings suggest that the included risk assessment tools perform comparably in different continents.

While we investigated a number of sources of within-study heterogeneity, it is possible that there are additional study characteristics that contribute to effect size variation that we did not explore. For example, we did not include the clinical background of the individuals who administered the risk assessment tools. Previous research has evidenced that this may be an important moderator of risk assessment tool validity (Ægisdóttir et al., 2006).

5. Conclusion

Violence risk assessment tools are increasingly used to make important decisions in clinical and criminal justice settings. The present meta-analysis found that the predictive validity of commonly used risk assessment measures varies widely. Our findings suggest that the closer the demographic characteristics of the tested sample are to the original validation sample of the tool, the higher the rate of predictive validity. We also found that tools designed for more specific populations were more accurate at detecting individuals' risk of future offending. Risk assessment tools were found to produce more valid risk predictions for older White individuals and possibly women. As this review identified substantial variations in the predictive accuracy of these instruments and heterogeneity in their validity according to different sample demographics, risk assessment procedures and guidelines by mental health services and criminal justice systems may need review.

Funding

There was no specific funding for this study.

Conflicts of Interest

None declared.

Acknowledgements

The authors thank Dr. Helen Doll for her assistance with statistical analysis, Professor Klaus Ebmeier for his assistance in the translation of several articles in German, and Sophie Westwood for her assistance with the inter-rater reliability check. The following authors are thanked for providing studies and/or tabular data for the analyses: April Beckmann, Sarah Beggs, Susanne Bengtson Pedersen, Klaus-Peter Dahle, Rebecca Dempster, Mairead Dolan, Kevin Douglas, Reinhard Eher, Jorge Folino, Monica Gammelgård, Robert Hare, Grant Harris, Leslie Helmus, Andreas Hill, Clive Hollin, Christopher Kelly, P. Randy Kropp, Michael Lacy, Calvin Langton, Henry Lodewijks, Karin Arbach Lucioni, Jeremy Mills, Catrin Morrissey, Thierry Pham, Martin Rettenberger, Marnie Rice, Michael Seto, David Simourd, Gabrielle Sjöstedt, Cornelis Stadtland, David Thornton, Vivienne de Vogel, Zoe Walkington, and Glenn Walters.

References

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 judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist*, *34*, 341—382.

American Psychiatric Association (2004). *Practice guideline for the treatment of patients with schizophrenia.* Arlington, VA: American Psychiatric Association.

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

Bauer, A., Rosca, P., Khawalled, R., Gruzniewski, A., & Grinshpoon, A. (2003). Dangerousness and risk assessment: The state of the art. *Israel Journal of Psychiatry and Related Sciences*, *40*, 182—190.

Bhui, H. S. (1999). Race, racism and risk assessment: Linking theory to practice with Black mentally disordered offenders. *Probation Journal*, *46*, 171—181.

Bjørkly, S. (1995). Prediction of aggression in psychiatric patients: A review of prospective prediction studies. *Clinical Psychology Review*, *15*, 475—502.

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*, *15*, 346—360.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20. Professional guidelines for assessing risk of sexual violence*. Burnaby, BC: Simon Fraser University, Mental Health, Law, and Policy Institute.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, *29*, 355—379.

Borum, R., Bartel, P., & Forth, A. (2002). *Manual for the structured assessment of violence risk in youth (SAVRY).* Tampa: University of South Florida.

Borum, R., Bartel, P., & Forth, A. (2003). *Manual for the structured assessment of violence risk in youth (SAVRY): Version 1.1*. Tampa: University of South Florida.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal*, *326*, 41—44.

Caldwell, R. A., Bogat, G. A., & Davidson, W. S. (1988). The assessment of child abuse potential and the prevention of child abuse and neglect: A policy analysis. *American Journal of Community Psychology*, *16*, 609—624.

Campbell, M., French, S., & Gendreau, P. (2007). *Assessing the utility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders (Cat. No. PS3-1/2007-4E-PDF)*. Ottawa, ON: Department of Safety and Emergency Preparedness.

Cleckley, H. (1941). *The mask of sanity.* St. Louis: C.V. Mosby.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37—46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (pp. 281—285). (2nd ed.). Hillsdale, NJ: Erlbaum.

Cottle, C. C., Lee, R. J., & Heilbrun, K. (2001). The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior*, *28*, 367—394.

Daniels, B. A. (2005). *Sex offender risk assessment: Evaluation and innovation*. Unpublished doctoral dissertation, Widener University, Chester, PA.

de Vogel, V., de Ruiter, C., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health*, *3*, 149—165.

Deeks, J. J. (2001). Systematic reviews of evaluation of diagnostic and screening tests. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context*. London: BMJ Publishing Groups.

DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist – Revised in court: A case law survey of U.S. courts (1991-2004). *Psychology, Public Policy, and Law*, *12*, 214—241.

Dernevick, M., Beck, A., Grann, M., Hogue, T., & McGuire, J. (2010). The use of psychiatric and psychological evidence in the assessment of terrorist offenders. *Journal of Forensic Psychiatry and Psychology*, *20*, 508—515.

Deville, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., de Vet, H. C. W., van der Windt, D. A. W. N., et al. (2002). Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Medical Research Methodology*, *2*, 1—13.

Doren, D. M. (2002). *Evaluating sex offenders: A manual for civil commitments and beyond.* Thousand Oaks, CA: Sage.

Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence risk assessment: Science and practice. *Legal and Criminological Psychology*, *4*, 149—184.

Douglas, K. S., & Webster, C. D. (1999). The HCR-20 violence risk assessment scheme: Concurrent validity in a sample of incarcerated offenders. *Criminal Justice and Behavior*, *26*, 3—19.

Edens, J. F. (2001). Misuses of the Hare Psychopathy Checklist – Revised in court. *Journal of Interpersonal Violence*, *16*, 1082—1093.

Edens, J., & Campbell, J. (2007). Identifying youths at risk for institutional misconduct: A meta-analytic investigation of the Psychopathy Checklist measures. *Psychological Services*, *4*, 13—27.

Edens, J., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior*, *31*, 53—75.

Edens, J., Skeem, J., Cruise, K., & Cauffman, E. (2001). The assessment of juvenile psychopathy and its association with violence: A critical review. *Behavioral Sciences and the Law*, *19*, 53—80.

Editorial (2010). Should protocols for observational research be registered? *Lancet*, *375*, 348.

Fazel, S., Sjöstedt, G., Långström, N., & Grann, M. (2006). Risk factors for criminal recidivism in older sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 159—167.

Federal Bureau of Investigation (2002). *Uniform crime reports for the United States.* Washington, DC: US Government Printing Office.

Fujii, D., Tokioka, A., Lichton, A., & Hishinuma, E. (2005). Ethnic differences in violence risk prediction of psychiatric inpatients using the Historical Clinical Risk Management – 20. *Psychiatric Services*, *56*, 711—716.

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.

Gendreau, P., Goggin, C., & Smith, P. (2000). *Cumulating knowledge: How meta-analysis can serve the needs of correctional clinicians and policy-makers.* Ottawa, ON: Correctional Service of Canada.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the "unparalleled" measure of offender risk?: A lesson in knowledge cumulation. *Criminal Justice and Behavior*, *29*, 397—426.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, *56*, 1129—1135.

Guy, L. (2008). *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Guy, L., 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.

Hanson, R. K. (1998). What do we know about sex offender risk assessment? *Psychology, Public Policy, and Law*, *4*, 50—72.

Hanson, R. K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis (Cat. No. PS3-1/2004-2E-PDF)*. Ottawa, ON: Public Works and Government Services Canada.

Hanson, R. K., & Morton-Bourgon, K. (2007). *The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis (Cat. No. PS4-36/2007E)*. Ottawa, ON: Public Safety and Emergency Preparedness.

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

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders (User Report 99-02)*. Ottawa, ON: Department of the Solicitor General of Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised (PCL-R)*. North Tonawanda, NY: Multi-Health Systems.

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

Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003.* Ottawa, ON: Solicitor General Canada.

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. *Law and Human Behavior*, *21*, 347—359.

Higgins, J. P. T., Deeks, J. J., & Altman, D. G. (2008). Special topics in statistics. In J. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 521). London: Wiley.

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539−1558.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557−560.

Higgins, N., Watts, D., Bindman, J., Slade, M., & Thornicroft, G. (2005). Assessing violence risk in general adult psychiatry. *Psychiatric Bulletin*, *29*, 131−133.

Honest, H., & Khan, K. S. (2002). Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Services Research*, *2*, 1−4.

Hoptman, M. J., Yates, K. F., Patalinjug, M. B., Wack, R. C., & Convit, A. (1999). Clinical prediction of assaultive behavior among male psychiatric patients at a maximum-security forensic facility. *Psychiatric Services*, *50*, 1461−1466.

Kemshall, H. (2001). *Risk assessment and management of known sexual and violent offenders: A review of current issues.* London: Home Office.

Khiroya, R., Weaver, T., & Maden, T. (2009). Use and perceived utility of structured violence risk assessments in English medium secure forensic units. *Psychiatrist*, *33*, 129−132.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1994). *Manual for the Spousal Assault Risk Assessment guide.* Vancouver, BC: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1995). *Manual for the Spousal Assault Risk Assessment guide* (2nd ed.). Vancouver, BC: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1999). *Spousal Assault Risk Assessment guide (SARA).* Toronto: Multi-Health Systems.

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 107−120.

Lawson, W. B., Yesavage, J. A., & Werner, P. A. (1984). Race, violence, and psychopathology. *Journal of Clinical Psychiatry*, *45*, 294−297.

Leistico, A., Salekin, R., 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.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of the American Medical Association*, *269*, 1007−1011.

Maden, A. (2001). Practical application of structured risk assessment. *British Journal of Psychiatry*, *178*, 479.

McCann, K. (2006). *A meta-analysis of the predictors of sexual recidivism in juvenile sexual offenders*. Unpublished master's thesis. Simon Fraser University, Burnaby, BC.

McNiel, D. E., & Binder, R. L. (1995). Correlates of accuracy in the assessment of psychiatric inpatients' risk of violence. *American Journal of Psychiatry*, *152*, 901−906.

Mercado, C. C., & Ogloff, J. R. P. (2007). Risk and the preventive detention of sex offenders in Australia and the United States. *International Journal of Law and Psychiatry*, *30*, 49−59.

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.

Moses, L. E., Littenberg, B., & Shapiro, D. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytical approaches and some additional considerations. *Statistics in Medicine*, *12*, 1293−1316.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783−792.

Mossman, D. (2000). Commentary: Assessing the risk of violence – Are "accurate" predictions useful? *Journal of the American Academy of Psychiatry and the Law*, *28*, 272−281.

National Institute for Health and Clinical Excellence (2009). *Schizophrenia: Core interventions in the treatment and management of schizophrenia in primary and secondary care.* London: National Institute for Health and Clinical Excellence.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk.* Washington, DC: American Psychological Association.

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.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*, 926−930.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York, NY: Sage.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*, 19−30.

Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist-Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice*, *3*, 203−215.

SBU (2005). *Riskbedömningar inom psykiatrin. Kan våld i samhället förutsägas? [Risk assessments in psychiatry. Is it possible to predict community violence?]. Stockholm: Swedish Council on Health Technology Assessment (SBU).*

Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior*, *31*, 449−462.

Schwalbe, C. S. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice and Behavior*, *35*, 1367−1381.

Schwalbe, C. S., Fraser, M. W., Day, S. H., & Arnold, E. M. (2004). North Carolina Assessment of Risk (NCAR): Reliability and predictive validity with juvenile offenders. *Journal of Offender Rehabilitation*, *40*, 1−22.

Seto, M. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment*, *17*, 156−167.

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

Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health*, *1*, 179−183.

Skeem, J., Edens, J. F., Camp, J., & Colwell, L. H. (2004). Are there ethnic differences in levels of psychopathy? A meta-analysis. *Law and Human Behavior*, *28*, 505−527.

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

Sreenivasan, S., Kirkish, P., Garrick, T., Weinberger, L. E., & Phenix, A. (2000). Actuarial risk assessment models: A review of critical issues related to violence and sex-offender recidivism assessments. *Journal of American Academy of Psychiatry and the Law*, *28*, 438−448.

StataCorp (2007). *Stata statistical software: Release 10.* College Station, TX: StataCorp LP.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *1*, 1−26.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*, 1559−1573.

Vess, J. (2008). Sex offender risk assessment: Consideration of human rights in community protection legislation. *Legal and Criminological Psychology*, *13*, 245−256.

Walters, G. (2003). Predicting institutional adjustment and recidivism with the Psychopathy Checklist factor scores: A meta-analysis. *Law and Human Behavior*, *27*, 541−558.

Wang, E. W., & Diamon, P. M. (1999). Empirically identifying factors related to violence risk in corrections. *Behavioral Sciences and the Law*, *17*, 377−389.

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.

Webster, C. D., Eaves, D., Douglas, K. S., & Wintrup, A. (1995). *The HCR-20 scheme: The assessment of dangerousness and risk.* Vancouver, BC: Mental Health, Law, and Policy Institute, and Forensic Psychiatric Services Commission of British Columbia.

Young, S. (2009). *Risk assessment tools for children in conflict with the law*. Dublin: Irish Youth Justice Service Retrieved December 2, 2008 from http://www.iyjs.ie/en/IYJS/Literature %20Review%20-%Risk%20Assessment.pdf/Files/Literature%20Review%20%20Risk%20Assessment.pdf.

Further reading[*]

Arbach, K., & Pueyo, A. A. (2007). Violence risk assessment in mental disorders with the HCR-20. *Papeles del Psicologo*, *28*, 174−186.

Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument.* Washington, DC: Pennsylvania Board of Probation and Parole.

Beggs, S. M., & Grace, R. C. (2008). Psychopathy, intelligence, and recidivism in child molesters: Evidence of an Interaction Effect. *Criminal Justice and Behavior*, *35*, 683−695.

Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime and Law*, *14*, 85−106.

Caperton, J. D. (2005). Predicting recidivism among sex offenders: Utility of the STATIC-99, Minnesota Sex Offender Screening Tool – Revised, and Psychopathy Checklist – Revised. Unpublished doctoral dissertation, Sam Houston State University, Huntsville, TX.

Dahle, K. P. (2006). Strengths and limitations of actuarial prediction of criminal reoffense in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. *International Journal of Law and Psychiatry*, *29*, 431−442.

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

de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical psychology and Psychotherapy*, *12*, 226−240.

de Vogel, V., de Ruiter, C., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health*, *3*, 149−165.

Dempster, R. J. (1998). Prediction of sexually violent recidivism: A comparison of risk assessment instruments. Unpublished Master's thesis, Simon Fraser University, Burnaby, BC.

Dempster, R. J. (2001). Understanding errors in risk assessment: The application of differential prediction methodology. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Dolan, M. C., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorder. *Psychological Assessment*, *20*, 35−46.

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services*, *54*, 1372−1379.

[*] These references are studies included in the meta-analysis.

Douglas, K. S., Yeomans, M., & Boer, D. P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior*, *32*, 479–510.

Dowdy, E. R., Lacy, M. G., & Unnithan, N. P. (2002). Correctional prediction and the Level of Supervision Inventory. *Journal of Criminal Justice*, *30*, 29–39.

Ducro, C., & Pham, T. (2006). Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 15–26.

Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2008). Failure of Static-99 and SORAG to predict relevant reoffense categories in relevant sexual offender subtypes: A prospective study. *Sexual Offender Treatment*, *3*, 1–14.

Folino, J., Almiron, M., & Ricci, M. A. (2007). Violent recidivism risk factor in filicidal women. *Vertex: Revista Argentina de Psiquiatria*, *18*, 258–267.

Folino, J. O., & Castillo, J. L. (2006). Las facetas de la psicopatia segun la Hare Psychopathy Checklist – Revised y su confiabildad [The facets of psychopathy described by the Hare Psychopathy Checklist – Revised and their reliability]. *Vertex*, *69*, 325–330.

Friendship, C., Mann, R. E., & Beech, A. R. (2003). Evaluation of a national prison-based treatment program for sexual offenders in England and Wales. *Journal of Interpersonal Violence*, *18*, 744–759.

Gammelgård, M., Koivisto, A. M., Eronen, M., & Kaltiala-Heino, R. (2008). The predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) among institutionalised adolescents. *Journal of Forensic Psychiatry and Psychology*, *19*, 352–370.

Gibas, A. L., Kropp, P. R., Hart, S. D., & Stewart, L. (2008, Julyy). Validity of the SARA in a Canadian sample of incarcerated adult males. *Paper presented at the International Association of Forensic Mental Health Services, Vienna, Austria*.

Grann, M., Belfrage, H., & Tengström, A. (2000). Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice and Behavior*, *27*, 97–114.

Grann, M., Långström, N., Tengström, A., & Kullgren, G. (1999). Psychopathy (PCL-R) predicts violent recidivism among criminal offenders with personality disorders in Sweden. *Law and Human Behavior*, *23*, 205–217.

Gray, N. S., Snowden, R. J., MacCulloch, S., Phillips, H., Taylor, J., & MacCulloch, M. J. (2004). Relative efficacy of criminological, clinical, and personality measures of future risk of offending in mentally disordered offenders: A comparative study of HCR-20, PCL: SV, and OGRS. *Journal of Consulting and Clinical Psychology*, *72*, 523–530.

Gretton, H., & Abramowitz, C. (2002, March). SAVRY: Contribution of items and scales to clinical risk judgments and criminal outcomes. *Paper presented at the American Psychology and Law Society, Austin, TX*.

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment*, *15*, 413–425.

Helmus, L. M. D., & Hanson, R. K. (2007). Predictive validity of the Static-99 and Static - 2002 for sex offenders on community supervision. *Sexual Offender Treatment*, *2*, 1–14.

Hill, A., Habermann, N., Klusmann, D., Berner, W., & Briken, P. (2008). Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy and Comparative Criminology*, *52*, 5–20.

Hollin, C. R., & Palmer, E. J. (2006). The Level of Service Inventory – Revised profile of English prisoners: Risk and reconviction analysis. *Criminal Justice and Behavior*, *33*, 347–366.

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.

Kloezeman, K. C. (2004). Violent behaviour on inpatient psychiatric units: The HCR-20 violence risk assessment scheme. Unpublished Master's thesis, University of Hawaii, Manoa, HI.

Kroner, C., Stadtland, C., Eidt, M., & Nedopil, N. (2007). The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Criminal Behaviour and Mental Health*, *17*, 89–100.

Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior*, *24*, 101–118.

Langton, C. M. (2003). Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy. Unpublished doctoral dissertation, University of Toronto, Toronto, ON.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior*, *34*, 37–59.

Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2008a). Gender differences in violent outcome and risk assessment in adolescent offenders after residential treatment. *International Journal of Forensic Mental Health*, *7*, 105–141.

Lodewijks, H. P. B., Doreleijers, T. A. H., & de Ruiter, C. (2008b). SAVRY risk assessment in violent Dutch adolescents: Relation to sentencing and recidivism. *Criminal Justice and Behavior*, *35*, 696–709.

Lodewijks, H. P. B., Doreleijers, T. A. H., de Ruiter, C., & Borum, R. (2008c). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. *International Journal of Law and Psychiatry*, *31*, 263–271.

McEachran, A. (2001). The predictive validity of the PCL:YV and the SAVRY in a population of adolescent offenders. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Meyers, J. R., & Schmidt, F. (2008). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with juvenile offenders. *Criminal Justice and Behavior*, *35*, 344–355.

Mills, J. F., Jones, M. N., & Kroner, D. G. (2005). An examination of the generalizability of the LSI-R and VRAG probability bins. *Criminal Justice and Behavior*, *32*, 565–585.

Mills, J. F., & Kroner, D. G. (2006). The effect of discordance among violence and general recidivism risk estimates on predictive accuracy. *Criminal Behaviour and Mental Health*, *16*, 155–166.

Morrissey, C., Hogue, T., Mooney, P., Allen, C., Johnston, S., Hollin, C., et al. (2007). Predictive validity of the PCL-R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry and Psychology*, *18*, 1–15.

Nicholls, T. L., Ogloff, J. R. P., & Ledwidge, B. (2007). Is the profound distrust of unbridled clinical opinion in the violence risk assessment field unfounded? Poster presented at the 4th Annual Forensic Psychiatry Conference, Victoria, BC.

Pham, T. H., Chevrier, I., Nioche, A., Ducro, C., & Reveillere, C. (2005). Psychopathie, evaluation du risque, prise en charge. *Annales Médico-Psychologiques*, *163*, 878–881.

Pham, T. H., Ducro, C., Marghem, B., & Reveillere, C. (2005). Evaluation du risque de recidive au sein d'une population de delinquants incarceres ou internes en Belgique francophone. *Annales Médico-Psychologiques*, *163*, 842–845.

Polvi, N. H. (1999). The prediction of violence in pre-trial forensic patients: The relative efficacy of statistical versus clinical predictions of dangerousness. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Public Safety and Solicitor General (2004). *Spousal Assault Risk Assessment (SARA) and Community Risk/Needs Assessment (CRNA): Predictive efficacy and interrelationships.* Victoria, BC: Public Safety and Solicitor General.

Ramirez, M. P., Illescas, S. R., Garcia, M. M., Forero, C. G., & Pueyo, A. A. (2008). Prediccion de riesgo de reincidencia en agresores sexuales. *Psicothema*, *20*, 205–210.

Reeves, K. A., Kropp, R., & Cairns, K. (2008). An independent validation study of the SARA. *Paper presented at the Annual Conference of the International Association of Forensic Mental Health Services, Vienna, Austria*.

Rettenberger, M., & Eher, R. (2007). Predicting reoffense in sexual offender subtypes: A prospective validation study of the German version of the Sexual Offender Risk Appraisal Guide (SORAG). *Sexual Offender Treatment*, *2*, 1–12.

Rice, M. E., & Harris, G. T. (2002). Men who molest their sexually immature daughters: Is a special examination required? *Journal of Abnormal Psychology*, *111*, 329–339.

Serin, R. C., Mailloux, D. L., & Malcolm, P. B. (2001). Psychopathy, deviant sexual arousal, and recidivism among sexual offenders. *Journal of Interpersonal Violence*, *16*, 234–246.

Seto, M. C., & Barbaree, H. E. (1999). Psychopathy, treatment behavior, and sex offender recidivism. *Journal of Interpersonal Violence*, *14*, 1235–1248.

Simourd, D. (2006). *Validation of risk/needs assessments in the Pennsylvania Department of Corrections.* Lower Allen, PA: Pennsylvania Department of Corrections.

Sjöstedt, G., & Långström, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior*, *25*, 629–645.

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime and Law*, *8*, 25–40.

Snowden, R. J., Gray, N. S., Taylor, J., & MacCulloch, M. J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine*, *37*, 1539–1549.

Soothill, K., Harman, J., Francis, B., & Kirby, S. (2005). Identifying future repeat danger from sexual offenders against children: A focus on those convicted and those strongly suspected of such crime. *Journal of Forensic Psychiatry and Psychology*, *16*, 225–247.

Sreenivasan, S., Garrick, T., Norris, R., Cusworth-Walker, S., Weinberger, L. E., Essres, G., et al. (2007). Predicting the likelihood of future sexual recidivism: Pilot study findings from a California sex offender risk project and cross-validation of the Static-99. *Journal of the American Academy of Psychiatry and the Law*, *35*, 454–468.

Stadtland, C., Hollweg, M., Kleindienst, N., Dietl, J., Reich, U., & Nedopil, N. (2006). Rueckfallprognosen bei Sexualstraftaetern - Vergleich der praediktiven Validitaet von Prognoseinstrumenten [Predictions of recidivism in sexual offenders: Comparison of the predictive validity of assessment tools]. *Der Nervenarzt*, *77*, 587–595.

Thomas, D.J. (2001). Identifying the sexual serial killer: A comparative study of sexual serial killers, serial rapists, and sexual offenders. Unpublished doctoral dissertation, Cincinnati, OH: Union Institute Graduate College.

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment*, *14*, 139–153.

Walkington, Z., O'Keeffe, C., & Thomas, S. (2006). Predicting violence recidivism in violent juveniles: A UK trial. *London: HM Prison Service*.

Walters, G., Duncan, S., & Geyer, M. (2003). Predicting disciplinary adjustment in inmates undergoing forensic evaluation: A direct comparison of the PCL-R and the PAI. *Journal of Forensic Psychiatry and Psychology*, *14*, 382–393.

Walters, G. D., Knight, R. A., Grann, M., & Dahle, K. P. (2008). Incremental validity of the Psychopathy Checklist facet scores: Predicting release outcome in six samples. *Journal of Abnormal Psychology*, *117*, 396–405.

Wormith, J. S., Olver, M. E., Stevenson, H. E., & Girard, L. (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services*, *4*, 287–305.