The predictive validity of risk assessment tools for female offenders

Geraghty, Kate; Woodhams, Jessica

The predictive validity of risk assessment tools for female offenders: a systematic review.

Kate Anya Geraghty, Jessica Woodhams

PII: S1359-1789(15)00003-8
DOI: doi: 10.1016/j.avb.2015.01.002
Reference: AVB 874

To appear in: *Aggression and Violent Behavior*

Received date: 11 January 2014
Revised date: 15 December 2014
Accepted date: 6 January 2015

Please cite this article as: Geraghty, K.A. & Woodhams, J., The predictive validity of risk assessment tools for female offenders: a systematic review., *Aggression and Violent Behavior* (2015), doi: 10.1016/j.avb.2015.01.002

Article Submitted: 11th January 2014

Title: The predictive validity of risk assessment tools for female offenders: a systematic review.

Article Type: Literature Review

Key Words: risk assessment, female offenders, predictive validity

Corresponding Author: Ms Kate Anya Geraghty, M.A.

Corresponding Author's Institution: University of Birmingham, UK

Order of Authors: Kate Anya Geraghty, M.A.; Jessica Woodhams, Ph.D.

Abstract

Assessing an offender's risk level is important given the impact of criminal behavior on victims, the consequences for the offender, and for society more generally. A wide range of assessment tools have been developed to assess risk in offenders. However, the validity of such tools for female offenders has been questioned. We present a systematic literature review of studies examining the accuracy with which risk assessment tools can predict violence and recidivism in female offenders. Five databases were searched, reference lists of relevant publications were hand searched, and an online search engine was used to identify studies. Fifteen studies were subject to review which evaluated nine risk assessment instruments (COMPAS, CAT-SR, HCR-20, LSI, PLC-R, OGRS, RISc, RM2000V, VRAG). The quality of these studies was systematically examined using a detailed quality assessment. The review findings indicate that the most effective tool for assessing both violence and recidivism in women was the LSI. There was variability in the quality scores obtained, with studies limited by measurement issues and standards of reporting results. Future research should aim to improve the quality of studies in this area, assess predictive accuracy across *subtypes* of female offenders, and compare correctional and psychiatric samples independently.

Author Inquiry: *Kate Anya Geraghty. KAG211@bham.ac.uk, +447450260438, School of Psychology, University of Birmingham, Edgbaston, Birmingham UK, B15 2TT.*

1. Introduction

Women comprise a minority of the offending population. Less than 5% of the prison population are female while women comprise 15% of offenders within the community (Ministry of Justice [MoJ], 2014; 2012a). Lower rates of violence and recidivism are also evident in female offenders. In terms of recidivism, the reoffending rate among offenders within one year following release is 18.3% for females while 28.3% for men (MoJ, 2012b). Rates of general violence in female offenders can vary from 14% to 27% (Greenfield & Snell, 1999; MoJ, 2012a), and it is widely acknowledged that female offenders are less likely to perpetrate violence than males (de Vogel & de Vries Robbé, 2013). However, rates for particular types of violence, such as intimate partner violence and violence committed by psychiatric offenders, are comparable between male and female offenders (de Vogel & de Vries Robbé, 2013; de Vogel, de Vries Robbé, van Kalmthout & Place, 2012). Furthermore, Logan (2004) suggested that violence against partners and children is more likely to lead to death when perpetrated by a woman (as cited in de Vogel, 2005). Criminal behavior is a significant problem that cannot be ignored, and adequately assessing risk of reoffending and violence in females is crucial.

Accurate assessment of future risk for violence and re-offending not only informs the management of offenders, but also ensures public safety (Craig, Browne, & Beech, 2008). It includes consideration of the: (a) nature, (b) frequency, (c) severity and (d) likelihood of harm (Craig et al., 2008). Risk assessment tools have been designed to enable the evaluation of the likely level of risk an offender holds for future violence and/or reoffending, and provide information on potential areas for management and planning. Although the criminal profile of male and female offenders is different (de Vogel & de Vries Robbé, 2013), few risk assessment tools exist that have been designed and validated on the female offending

population to assess risk for future violence or reoffending. This is in spite of the increasing literature recognizing that risk factors for future violence and offending in females may be different to males (Caulfield, 2010; Chesney-Lind & Pasco, 2013). The generalizability of risk assessment tools to female offenders has, therefore, been questioned. As such, it is important that researchers and practitioners are aware of the strengths and weaknesses of risk assessment tools currently used to predict violence and recidivism in female offenders. This review sought to synthesize what is currently known about the predictive validity of these tools with female offenders and subject these studies to quality assessment.

1.1. Evaluating Predictive Validity

In evaluating the accuracy of risk assessment, studies typically assess the predictive validity of a risk assessment tool. Predictive validity (or accuracy) refers to the ability of an instrument to correctly assess the likelihood of violence or recidivism (Singh, 2013). The most commonly used statistical analysis of predictive accuracy is Receiver Operating Characteristic (ROC) analysis which was introduced to violence risk assessments in the 1990s (Douglas, Cox, & Webster, 1999; Mossman, 1994). This analysis produces a statistic of predictive accuracy called the Area Under the Curve (AUC). An AUC can be interpreted as a global discrimination index, equal to the probability of a randomly selected recidivist scoring higher on a risk instrument than a randomly selected non-recidivist (Mossman, 1994). An AUC of 0.00 represents perfect negative prediction, an AUC of .50 indicates chance prediction, and an AUC of 1.0 indicates perfect positive prediction. AUC values > .70 are considered ‘moderate’ and values > .75 ‘good’ (Douglas, Guy, & Reeves, 2008). A particular advantage of AUC estimates is that they are largely independent of base rates and selection ratios (Rice & Harris, 1995).

Predictive accuracy can also be measured using the Correlation Coefficient (*r*). This measures the direction and strength of association between two variables (Field, 2009; Warner, 2008), which, in this context, is risk score and violence or recidivism. Values range from -1.00 (perfect negative association) to +1.00 (perfect positive association). A value > .30 is indicative of a moderate relationship, while values > .50 represent a strong relationship (Cohen, 1988).

1.2. Approaches to Risk Assessment

There are three main approaches to risk assessment (Bonta, 1996). The first generation of risk assessment was 'clinical judgement' and involved the use of unstructured professional judgement to determine an offender's risk level. Predicated on professional experience and knowledge of the area, the predictive accuracy of this type of risk assessment was found to be no better than chance (Hanson & Bussière, 1998). Findings such as this led to the development of second generation risk assessment tools; actuarial assessments. These are static instruments which are based on factors empirically associated with recidivism. Particular benefits of actuarial measures are that they are less open to interpretation and they are structured and replicable (Kemshall, 2002). Examples of actuarial risk instruments include the Psychopathy Checklist Revised (PCL-R; Hare, 1991) and the Violence Risk Appraisal Guide (VRAG; Harris, Rice, & Quinsey, 1993). Although the PCL-R was not designed to predict violence or recidivism, it is used regularly in forensic settings to assess risk of these outcomes (Grann, Långstrom, Tengström, & Gunnar, 1999; Hart, 1998a). Accuracy estimates for actuarial instruments are within the moderate range (Hart, Michie, & Cooke, 1997) and research still attests to their predictive validity (Hare, Clark, Grann, & Thornton, 2000). Nevertheless, a myriad of criticisms have been levelled at actuarial risk tools which include concerns regarding their predictive and content validity (Hannah-Moffit

& Shaw, 2001). Actuarial risk assessments have also been criticized due to their lack of accuracy in estimating risk at an individual level and also their minimal utility in the *management* of offenders' risk (Hart, Michie, & Cooke, 2007).

A third generation of risk assessment tools was, therefore, developed which integrated dynamic and static risk factors. These tools are referred to as Structured Professional Judgement (SPJ). They are empirically guided, in that they are based on factors empirically demonstrated to be associated with risk, but judgements are also clinically informed (Hart, 1998b). Examples of SPJ tools include the HCR-20 (Webster, Douglas, Eaves & Hart, 1997), Level of Service Inventory (LSI-R; Andrews & Bonta, 1995) and the Violence Risk Scale (VRS; McNeil & Binder, 1994). All of these instruments have demonstrated good predictive validity with AUC > .70 (Douglas, Ogloff, Nicholls, & Grant 1999; Gray, Taylor & Snowden, 2008) and correlation values > .50 (Gray, et al., 2003).

1.3. Assessing risk with female offenders

Empirical evidence attesting to the predictive validity of risk assessment instruments for female offenders is mixed at best and hotly debated (Caulfield, 2010; Chesney-Lind & Pasko, 2013; McKeown, 2010). Critics of risk assessment tools have asserted that they may not capture salient factors relevant to pathways that lead to the onset and maintenance of offending for women (Blanchette, 2002; Blanchette & Brown, 2006; Chesney-Lind & Pasko, 2013). These include scales/items such as: relationships, parental issues, self-esteem, self-efficacy, depression, victimization, and trauma (Blanchette, 2002; Blanchette & Brown, 2006; Reisig, Holtfreter, & Morash, 2006). It is also suggested that these factors are either not typically seen in men, may be seen in men but occur at a greater frequency in women, or can be seen in men and women but affect women in unique personal and social ways (Chesney-Lind & Shelden, 2004; Farr, 2000; Funk, 1999). As such, the ability of current risk tools to

accurately measure risk in female offenders has been questioned, which leads to the fundamental question of whether risk assessment tools are valid for the female offending population (Davidson & Chesney-Lind, 2009).

Policy makers are increasingly recognizing this debate as was reflected in the publication of the Corston Report (Home Office, 2007), and an English and Welsh Government Green Paper in 2010 which asserted that female offenders may have a different profile of risks (Ministry of Justice, 2011). Therefore, the adoption of gender-responsive strategies to the assessment of female offending is popular on the political and mental health agenda (Nedopil, 2009). However, an evaluation of whether gender responsive risk assessments are needed has not taken place and a central question remains: Are risk assessment tools valid for female offenders and which tools have the highest rates of predictive accuracy? Even with samples of male offenders, no single risk assessment tool has been demonstrated to have greater predictive accuracy than another (Campbell, French, & Gendreau, 2007; Gendreau, Goggin, & Smith, 2002). It is still largely unknown which tools are more accurate in particular settings and for certain populations, including female offenders (Caulfield, 2010; McKeown, 2010; Singh, Grann, & Fazel, 2011). A systematic evaluation of the accuracy of risk assessment tools in predicting recidivism and violence for female offenders is therefore warranted.

1.4. The current review

To date, there has been no systematic review of the predictive accuracy of risk assessment instruments for adult female offenders which has included a systematic appraisal of the quality of studies in the area. Additionally, no review has considered the prediction of either recidivism or violence. The following review aimed to fill these gaps in our knowledge by drawing together what is known about the accuracy with which instruments can predict

recidivism and violence in female offenders while also evaluating the quality of this empirical research. Given the negative consequences for both the offender and the public arising from false positive and false negative errors in risk assessment (Douglas et al., 2008) and the tendency for professionals to underestimate risk in females (Skeem et al., 2005), the review has added importance. Adopting a systematic approach the current review aimed to:

- Identify instruments that have been used with female offenders to assess risk of violence or recidivism
- Collate information on the predictive accuracy of these instruments
- Determine the accuracy with which these risk instruments have been shown to predict violence and recidivism in female offenders
- Determine which instruments are more effective at predicting recidivism
- Determine which instruments are more effective at predicting violence
- Appraise the methodology and quality of studies in the area

2. Method

2.1. Review Protocol

The current review was conducted in accordance with the Centre for Reviews and Dissemination (CRD) guidelines (2009). A protocol was constructed prior to the review which stated the search strategy, inclusion/exclusion criteria and forms of quality assessment to be used.

2.1.2. Scoping Search

An initial scoping search was conducted in May, 2013 to determine the need for the systematic review. Gateways used for the scoping search included Cochrane Central and the Centre for Reviews and Dissemination. The search identified six previous meta-analyses. Andrews et al. (2011) conducted a meta-analysis of the predictive accuracy of the LSI-R;

however, only the LSI-R instrument was evaluated and juvenile offenders were included in their review. Andrews et al. (2012) further examined the LSI in male and female offenders; however, the review had no clear description of female offenders and included juveniles within their sample and was therefore excluded from the present review. O'Shea, Mitchell, Picchioni, and Dickens (2012) undertook a meta-analysis of the predictive validity of the HCR-20 in predicting violence in psychiatric facilities. However, its focus was limited to an inpatient, psychiatric setting and, therefore, did not include correctional samples. Additionally, a gender breakdown of the effect sizes was not reported for each of the subscales. Singh, Grann, and Fazel (2011) conducted a systematic review of the predictive accuracy of violence risk assessment tools for males and females but did not present predictive validity estimates separately for female offenders in this review and included juvenile offenders within the sample. Also, no systematic quality assessment of studies was undertaken. The meta-analysis by Yang, Wong, and Coid (2010) evaluated the efficacy of nine risk assessment tools and included female offenders within the analyses. However, there was no quality assessment of the studies included and their search was limited to studies published between 1999 and 2008 thereby potentially missing studies published outside of these timescales. Smith, Cullen, and Latessa (2009) conducted a meta-analysis evaluating the merits of the LSI-R in predicting recidivism in female offenders. This was excluded from the review as it was unclear whether the analysis had used juvenile offenders within their sample.

2.1.2.Systematic Review Search Strategy

The search was limited to 1990 onwards given that the majority of risk assessment tools had been developed post-1990. Five electronic databases were searched from January 1st 1990 to May 18th 2013; OVID PsycINFO, OVID EMBASE, OVID MEDLINE, Applied Social Science Index and Abstracts, and ISI Web of Science. The search combined terms related to

(a) assessing risk, (b) recidivism or violence, and (c) female offenders. The subject headings specific to each database were determined using the thesaurus function. As such, these differed across each database. "Wild card" search characters were used to obtain all permutations of the search term. In order to increase the comprehensiveness of the search, the reference lists of key papers in the area were hand-searched for other relevant articles to include in the review. Potential grey literature was sought by contacting seven experts and professionals identified through the scoping search. Additionally, Google Search Engine was searched on May 25th 2013 using the same search terms as were used with the databases to identify publications and key meta-analyses for use as potential sources of relevant publications.

2.2. Inclusion/Exclusion Criteria

2.2.1. Population

The inclusion criteria for the review were; an eligible study must have examined female offenders aged 18 years or older. Juvenile offenders were not included in the review given that risk assessment tools developed for use with juveniles, such as the SAVRY (Borum, Bartel, & Forth, 2006), are intended for this population only and cannot be generalized to other populations (such as adults). As research on individual subtypes of female offenders is still in its infancy (Caulfield, 2010), the aim of the current review was to be inclusive; studies with any type of female offender were included within the review, including psychiatric offenders.

2.2.2. Intervention.

Included studies must have examined the predictive accuracy of a standardized risk assessment tool (actuarial or structured professional judgement) to predict the risk of future

violence or recidivism. Standardized tools were considered to be those which have been validated on the offending population.

2.2.3. Outcome.

Given the limited research in the area, two outcomes were included in the review; recidivism and violence. Violence was defined as per the HCR-20 and included any violence, including threatening behavior, and verbal threats used to induce fear and/or cause harm in another person (Webster et al., 1997). Recidivism was defined as reconviction and/or rearrest for any offense. This broad definition of recidivism was used given the low numbers of female offenders within the criminal justice system and the low rates of conviction for this sample (Ministry of Justice, 2011).

2.2.4.Study Design.

The systematic review was not limited to any particular study design due to the dearth of literature in this area. Both retrospective and prospective study designs were included. Publications which did not report empirical data were excluded (e.g., editorials). Studies written in a language other than English were excluded due to difficulties in obtaining reliable translation.

If articles met the eligibility criteria they were put forward for potential inclusion in the review and subjected to a quality assessment. Where an article was considered relevant a hardcopy was obtained for further consideration.

2.5. Screening

Figure 1 provides a visual appraisal of the data selection process.

Insert Figure 1 about here.

The database search yielded 256 articles. After the removal of duplicates this left 194 articles. A further four articles were identified through Google Search Engine and through searching reference lists in relevant publications. No articles were identified through contact with experts. The titles and abstracts of these 198 articles were screened according to the exclusion/inclusion criteria and 136 articles were deemed irrelevant. The full-text versions of the remaining 62 articles were obtained and a second level of screening was conducted whereby each full-text article was subjected to the inclusion/exclusion criteria. Forty articles were deemed irrelevant following this second screening, leaving a potential 22 articles for inclusion in the review.

2.6 Quality Assessment

The remaining 22 articles were subjected to a quality assessment. There is no universal framework for assessment of quality in observational studies. As such, to assess risk of bias within primary studies, an adapted tool was created using the Critical Appraisal Skills Programme (CASP, 2013), Effective Public Health Practice (EPHPP, 1998) and CRD (2009) guidelines. The assessment of quality was completed in two steps. First, threshold criteria were applied to each study. Threshold criteria included having a clear description of the female offenders, the tools used and the outcome measure, as well as sufficiently detailed statistical analyses regarding the prediction of recidivism and violence. Seven studies failed to meet the threshold criteria and were therefore excluded. Second, the methodological quality of the remaining 15 primary studies was assessed using the adapted quality assessment form. This form contained 18 questions relating to a range of methodological considerations including; selection bias, measurement bias, attrition bias, and reporting bias.

Salient questions related to the generalizability of the study, how the outcome was measured, statistical reporting standards, and consideration of confounding variables. Additionally, the quality assessment form also considered pragmatism by assessing the practical utility of the studies assessed. A scoring system was applied to each of the questions. Where conditions were not met, a score of 0 was allocated. If conditions were partially met, a score of 1 was applied. Where conditions were met and there was no ambiguity in the study regarding the condition, a score of 2 was applied. If it was unclear whether a condition had been met or not, a question was scored as 'Unclear'. The total number of unclear scores were calculated for each study.

The primary author reviewed all 15 studies. In order to ensure the reliability of quality assessment, eight studies (53%) were dual-assessed by an independent rater, qualified to postgraduate level. Inter-rater agreement was assessed using an intraclass correlation coefficient (ICC). An ICC of .86 was found indicating strong agreement between the raters. Any differences in opinion between raters were resolved by consensus.

2.7. Data Extraction

A predefined form was used to extract data, provide an overview of the quality of the study and clarity of reporting, and record limitations for each study. Information extracted included: the population studied and characteristics, sample size, offender type, the risk assessment tool(s) used and any inter-rater reliability estimates, the outcome measure including how the outcome was defined, length of follow-up, statistical analysis used to predict the outcome, and strengths and limitations of the study.

3.Results

3.1.Description of included studies

Of the 15 studies included in the review, three followed a prospective design; ten were retrospective in nature, while two of the studies followed a mixed prospective-retrospective design. Thirteen used a correctional sample while two used a psychiatric sample of female offenders. A total of 12 risk instruments were included in the review, although it is noted that there were four variations of the LSI risk assessment used across studies. The risk assessment tools included: Correctional Offender Management Profiling for Alternative Sanction (COMPAS; Brennan & Oliver, 2000), Child and Adult Taxon Scale-Self-Report (CAT-SR; Quinsey, Harris, Rice, & Cormier, 2006), Historical Clinical and Risk Management Scale (HCR-20; Webster, Douglas, Eaves, & Hart, 1997); Level of Service Inventory (LSI; Andrews, 1982), Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995), Level of Service Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormirth, 2004), Level of Service Inventory Ontario Revision (LSI-OR; Andrews, Bonta, & Wormirth, 2004), Offender Group Reconviction Scale (OGRS; Copas & Marshall, 1998), Psychopathy Checklist-Revised (PCL-R; Hare, 1991), Risk Assessment Scales (RISc; Van Montfoort & Reclassering Nederland, 2004), Risk Matrix 2000 Violence Scale (RM2000V; Thornton et al., 2007), and Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 2006).

The most commonly used measure to assess risk was the LSI with eight studies evaluating its predictive validity. Four studies examined the predictive validity of both the HCR-20 and PCL-R while four studies assessed the remaining six instruments. The size of the total relevant sample was 7,893 participants (M = 526, Range = 42-2,831). Based on the 11 studies that provided the age of their samples, the overall mean age of participant included in the review was 34 years (Range = 28-42). Of the studies included in the review, seven were conducted in the US, four were undertaken in Canada, three were undertaken in the Netherlands, and one study was conducted in the UK. For six of the seven studies that

provided percentage breakdowns for ethnicity, the majority of participants were Caucasian. None of the included studies examined offender subtypes. Twelve studies used recidivism only as their outcome measure, one used violence only as their outcome measure, while two measured both violence and recidivism as an outcome. In 10 of the studies, the follow-up periods were stated. From these studies, the mean range of follow-up for studies assessing recidivism (n = 10) was 2.78 years. No follow-up period was reported for the study which assessed violence only. In total, five studies provided insufficient details of follow-up period. Table 1 provides a summary table of the main characteristics of the studies included in the review. It also includes AUC and *r* statistics for the studies as well as the quality score for each study.

3.2. Data synthesis

As noted above, AUC values <.70 are generally considered as indicative of 'moderate' predictive accuracy and values < .75 are 'good' (Douglas et al., 2008); therefore, studies were examined in relation to the recommended benchmark of .70. It should be noted that not all studies reported AUC estimates; some studies provided bivariate analyses only (Reisig et al., 2006; Rettinger & Andrews, 2010; Salisbury et al., 2009; Vose et al., 2009). As such, these effect sizes were examined in relation to Cohen's (1988) criteria where values greater than .30 are considered moderate and those greater than .50 are large.

Few of the instruments achieved either an AUC or *r* indicative of a moderate or large effect size. Of the nine risk assessment tools used to predict recidivism or violence in female offenders, only the HCR-20, LSI, and PCL-R yielded either an AUC or *r* above the recommended threshold. The tool with the worst performance was the VSC.

3.2.1. Recidivism.

There were conflicting results both within and across tools in terms of their ability to predict recidivism. Nevertheless, some tools achieved the .70 recommended level (or higher) for subtypes of recidivism. The HCR-20 achieved moderate predictive accuracy for violent recidivism (AUC = .70) in one correctional sample (Coid et al., 2009), but not in another (Warren et al. 2005). However, the 95% confidence interval for the AUC in Coid et al.'s (2009) study was wide suggesting that obtained AUC values cannot be interpreted with confidence. Other studies found the HCR-20 to perform no better than chance at distinguishing recidivists from non-recidivists (Schaap et al., 2009). The primary author, compared the effect sizes for the HCR-20 for correctional and psychiatric samples. This was done using an independent samples *t*-test which revealed no differences between the two samples ($p > .05$) .

The LSI and its variants (LSI-R/LSI-OR/LS/CMI) obtained the highest AUC estimates for predicting recidivism. Rettinger (1998) found large AUCs for both general recidivism (AUC = .93) and violent recidivism (AUC = .85). However, other studies did not obtain such strong predictive accuracy. Van Voorhis et al. (2010) found that the LSI-R accurately predicted recidivism in two correctional samples (AUC = .71 and .72, significant at the .01 level). Using correlation coefficients as an estimate, Rettinger and Andrews (2010) found the LSI-R to be accurate in predicting general and violent recidivism ($r = .63$ and .44 respectively). Conversely, other studies (Salisbury et al., 2009; Vose et al., 2005), using bivariate analyses, did not find the LSI-R to be predictive of recidivism with no correlation coefficient reaching .30. The LS-CMI was found to be an accurate predictor of any recidivism (AUC = .87) and violent recidivism (AUC = .86; Rettinger & Andrews, 2010). Similarly, the LSI-OR was found to be a valid predictor of recidivism among a correctional sample (AUC = .79, Brews, 2009). This held true for the different types of sentences the

correctional sample received (custodial, conditional, conditional and probational sentences). None of the LSI instruments were evaluated with psychiatric samples.

Insert Table 1 about here

The PCL-R was able to predict recidivism with moderate accuracy in a correctional sample (AUC = .73; Coid et al., 2009), but the same was not found for a psychiatric sample (AUC = .57; Schaap et al., 2009). When the primary author compared effect sizes between the correctional and psychiatric sample using independent *t*-tests, no significant differences were found ($p > .05$). The widths of the 95% confidence intervals across the majority of studies assessing the PCL-R were large which would suggest that there is variability with regards to where the true effect size falls for each study (Warner, 2008). The PCL-SV was found to be predictive of any recidivism (AUC = .90) and violent recidivism (AUC = .87) in a community psychiatric sample (Nicholls et al., 2004).

The remaining instruments (CAT-SR, COMPAS, OGRS-II, RISc, RM2000V, VRAG) failed to demonstrate acceptable effect sizes using AUC or *r* estimates. Some tools (OGRS) performed no better than chance at differentiating recidivists from non-recidivists. However, it should be noted that the RISc approached the recommended .70 threshold (AUC = .68, Van der Knaap et al., 2012) and the confidence interval for this tool was narrower than for other instruments included in the review, suggesting that the RISc may moderately predict recidivism.

In terms of recognising how individual differences might affect risk assessment, two studies considered the effect of ethnicity (Brews, 2009; Rettinger & Andrews, 2010). These evaluated the LS-CMI and the LSI-OR. The LS-CMI was found to be most accurate in predicting recidivism in Black female offenders (AUC = .95). Estimates for White and Aboriginal female offenders were also above the recommended .70 level (AUC = .86 and .84

respectively; Rettinger & Andrews, 2010). Similarly, the LSI-OR was found to be more accurate in predicting recidivism among Black female offenders (AUC= .81) which was higher than the AUC for the sample overall (AUC = .74; Brews, 2009).

Some studies also provided effect size estimates for individual subscales of the risk instruments. When the subscales of the instruments that did demonstrate some predictive validity (HCR-20, PCL-R, LSI-R) were examined, some discrepancies regarding the predictive power of subscales were uncovered. The Historical subscale of the HCR-20 predicted recidivism more accurately than the Clinical scale in a correctional sample (AUC = .73; Coid et al., 2009). Rettinger and Andrews (2010) found that the LSI/CMI was better at predicting general recidivism than other types of recidivism. 'Minor risk factors' (that includes items such as accommodation, financial, personal/emotional, general risk/need) and 'moderate risk factors' (that includes family/marital, education employment, alcohol/drug, leisure/recreation) were more accurate in the prediction of general recidivism (r = .65 and .64, respectively) than violent recidivism (r = .47 for both) or new convictions (r = .59). This is in comparison with the 'major factors' (which includes criminal history, antisocial pattern, pro-criminal attitude, companions) of the instrument where the major factors were better at predicting general recidivism (r = .61) and new convictions (r = .59) than violent (r = .45) recidivism as well. The PCL-R Factor 2 was found to be more predictive of violent recidivism in a correctional sample (AUC = .71, Coid et al., 2009), but was less accurate when predicting 'acquisitive' or 'any' recidivism. Conversely, Factor 2 was less accurate in predicting violent recidivism in a psychiatric sample (AUC = .62; Schaap et al., 2009). The primary author, using an independent samples t-test compared the differences in effect size between the correctional and psychiatric samples. There were no significant differences (p > .05).

3.2.2. Violence.

The ability of two risk tools, HCR-20, PCL-R, to accurately predict violence was evaluated by studies included in the review. The results for the tools were very variable and only the HCR-20 reached acceptable levels of predictive accuracy. The predictive accuracy of the HCR-20 was examined with one psychiatric and one correctional sample (De Vogel & de Ruiter, 2005; Warren, 2005). The HCR-20 was found to be no better than chance in predicting violence among either the psychiatric or correctional samples (AUC = .59; de Vogel & de Ruiter, 2005; AUC = .55; Warren et al., 2005). When the differences between effect sizes between the psychiatric and correctional samples were compared by the primary author, no significant differences were found ($p > .05$).

However, there was variability with respect to the accuracy of individual scales for the HCR-20 for psychiatric patients. De Vogel and de Ruiter (2005) found that the Final Risk Judgement of the HCR-20 had strong predictive accuracy for future violence (AUC = .86). No effect sizes were reported for their correctional sample.

3.3. Quality Assessment

Quality assessment forms were completed on all 16 studies included in the review. A copy of the quality assessment is included in the appendix. Table 2 provides a summary table of the elements included in the quality review and the scores each study obtained.

Insert Table 2 about here

There was little variability in the total quality scores for studies included in the review. The mean quality score for the included studies was 22.80 (*SD* = 3.76; Range = 16-29) out of a possible 36. The number of unclear items ranged from 2 to 4. All studies scored similarly with respect to selection bias, attrition bias and clinical judgement but there was more

variability with respect to measurement and reporting bias. These differences made it difficult to draw comparisons within and across studies and to extrapolate from the findings.

The quality scores for studies assessing recidivism and assessing violence did not differ significantly from one another as assessed by independent samples *t*-tests ($p > .05$). The quality scores for the correctional sample and the psychiatric sample also did not differ significantly from each other as determined by independent samples *t*-tests ($p > .05$). The study with the highest quality score in the review also obtained the highest predictive validity estimates (Rettinger, 1998). When the quality scores between studies who achieved higher predictive validity estimates ($M = 23$, $SD = 5.22$) were compared with those who obtained lower estimates ($M = 23$, $SD = 5.22$), there were no significant differences found ($p > .05$).

4. Discussion

4.1. General Findings

The findings suggest that there is great variability with respect to the accuracy of risk assessment tools in predicting either violence or recidivism with female offenders. Risk instruments were found to be more accurate at predicting recidivism than violence.

Additionally, the widths of the confidence intervals do not give confidence with respect to either AUC or *r* estimates obtained in studies that were above the recommended thresholds (Warner, 2008). Of the studies included within the review, it seems that the HCR-20, PCL-R predict recidivism more accurately for female offenders and the LSI and its variants (LSI-R/LSI-OR/LS/CMI) is the most accurate tool for predicting recidivism. The poor predictive accuracy of the VRAG for recidivism with female offenders is in contrast with research studies with male offenders (Glover, Nicholson, Hemmati, Bernfield, & Quinsey, 2002). Although the majority of tools did not reach recommended statistical

thresholds, a comparison of estimates between correctional and psychiatric samples in the review suggests that risk assessment tools may be as valid in either setting.

When subscales were examined, the Historical scale of the HCR-20 was a more accurate predictor of recidivism suggesting that for correctional samples the best predictor of future behavior is past behavior. This supports the literature demonstrating the relevance of static factors in the prediction of recidivism (Hare et al., 2000). Conversely the LSI-R subscales demonstrated the opposite relationship, whereby dynamic factors were found to be more accurate predictors of recidivism. This supports the research advocating the adoption of 'gender-responsive' approaches to the assessment of female offenders (Blanchette, 2002; Chesney-Lind & Pasko, 2013; Chesney-Lind & Sheldon, 2004; Farr, 2000).

The variability in predictive validity across risk assessment tools may be due to risk assessments not capturing, in full, the relevant risk factors associated with the onset and maintenance of female violence. Theories on pathways to female offending highlight the role of victimization in predisposing women to violence (Daly, 1994; Simpson, Yahner, & Dugan, 2008). Furthermore, salient risk factors for female violence include relationships with others and mental-health difficulties (Blanchette, 2002; Blanchette & Brown, 2006; Reisig, Holtfreter, & Morash, 2006). However, these risk factors are also relevant for male offenders as evidenced by their inclusion as risk factors within risk assessments such as the HCR-20. However, the manifestation and function of these risk factors for future violence in females may be unique (Caulfield, 2010; Nicholls, Greaves, & Moretti, 2008). Additionally, unique risk factors for violence for women include prostitution, pregnancy at a young age and self-harm (Blanchette & Brown, 2006). Current risk assessments may not be capturing these risk factors. The implications of this then may mean that current tools may not be as valid for

female offenders than their male counterparts which may explain why violence was not well predicted in the current review.

Overall, the review demonstrates that SPJ tools perform better than the actuarial tools evaluated in the studies in this review (e.g., PCL-R, CAT-SR, RM2000, VRAG). However, there was very little difference between all the instruments incorporated within the review; as such, the predictive potency of SPJ tools here is less than perfect. These findings give further weight to those authors in the field (Blanchette, 2002; Caulfield, 2010; Chesney-Lind, 2013; Davidson, & Chesney-Lind, 2009) raising concerns about the uncritical application of risk assessment tools developed on male samples to females.

4.2. Strengths and Limitations of the Current Review

The review contributes uniquely to previous research assessing the validity of risk assessment tools through evaluating available research on the predictive validity of risk assessment tools for female offenders. The current review also used a comprehensive search strategy (going beyond electronic databases) and utilized search terms that were based on previous reviews in the area which were cross referenced with key publications as a measure of quality control. Additionally, reference lists were hand-searched and a free-search strategy was adopted to identify relevant publications and grey literature. This ensured an inclusive review (Egger, Dickerson, & Smith, 2007).

Another strength of the review is that it incorporated a quality assessment of the studies included. This is unlike previous reviews in the area (Singh et al., 2011; Yang et al., 2009). When the reference lists of the articles included in these two reviews were hand-searched and compared with the studies included in the current review as a measure of quality control, some studies were identified that were not included in the current review as they failed to meet inclusion or threshold criteria (i.e., Coulson et al., 1996; Raynor, 2007).

Reasons for their exclusion from the current review included lack of a clear description of female offenders (Raynor, 2007) and the authors changing some of the questions of the risk tool, thereby undermining the tool's standardisation (Coulson et al., 1996). The question remains as to whether these studies would have been included in past reviews if these reviews had incorporated a quality assessment stage. Additionally, the adoption of a quality threshold enabled an objective means of selecting studies for inclusion.

Nevertheless, the review may be limited by publication bias as only two non-peer reviewed papers were included (Brews, 2009; Rettinger, 1998). Aside from these papers, no other grey literature was identified for inclusion. Although relevant experts were contacted, this yielded no results. Given that research in this area is still in its infancy it may be that not all experts within this area were identified. This may be another reason for the lack of identifiable grey literature.

The inclusion/exclusion criteria also led to the omission of articles that were not in the English language. Although such exclusion has been found to have minimal effect in reviews (Juni, Holenstein, Sterne, Bartlett, & Egger, 2002), it may nevertheless introduce systematic bias (Song et al., 2010). Another point for consideration is that the current review focused exclusively on predictive validity. Although attempts were made to consider pragmatism and the practical utility of risk assessment instruments, focusing solely on predictive validity within research has been criticized by researchers assessing risk in female offenders as ignoring content validity (Davidson & Chesney-Lind, 2009). While this position is grounded in consideration of implications for practice, a necessary step in evaluating the worth of any measure is determining its validity and reliability (Breakwell et al., 2008; Warner, 2008). This review is, therefore, a necessary step in contributing to the literature on demonstrating the value of risk assessment tools for the female offending population.

4.3. Strengths and weaknesses of the studies

The review appears to be the first to systematically evaluate the predictive accuracy of risk assessment tools for adult female offenders *and* appraise the quality of research in the area. Findings were variable and this heterogeneity between studies assessing predictive accuracy may be due to 'legitimate' and 'illegitimate' variability (Andrews et al., 2010). Legitimate variability includes; the reliability of the measure used, the climate and culture of the agency, the accuracy of how the outcome measure is measured, and the heterogeneity of the population to whom the risk tool is applied. On the other hand, illegitimate variability artificially distorts validity estimates and can include; experimenter bias, reporting bias and manipulation of scoring or data. On the basis of the quality score, studies varied both legitimately (measurement of outcome, population being measured) and illegitimately (reporting bias) as evaluated in the quality assessment. In terms of the studies' quality scores, the variability across studies and predictive validity estimates obtained would suggest that it is difficult to draw firm conclusions about the predictive validity estimates obtained. It is noted that the studies that achieved a higher level of quality also obtained the highest predictive validity estimate (Rettinger, 1998; Rettinger & Andrews, 2010). However, this was not true for all studies (Warren et al., 2005).

In terms of sampling quality, there was homogeneity of scores across studies with most scores being high. The main reason for studies losing scores here was due to the sample not being randomly selected and therefore being unlikely to be representative of female offenders. In relation to external validity, 11 of the 15 studies were conducted in the US and Canada (Brennan et al., 2009; Brews, 2009; Folsom & Atkinson, 2007; Hastings et al., 2011; Reisig et al., 2006; Rettinger, 1998; Rettinger & Andrews, 2010; Salisbury et al., 2009; Van Voorhis et al., 2010; Vose et al., 2009; Warren et al., 2005). Additionally, most of the risk

assessment tools have been validated on samples from these geographical areas. Of the studies that were conducted elsewhere, one was conducted in the UK (Coid et al., 2009) while the remaining three were conducted in the Netherlands (De Vogel & De Ruiter, 2005; Schaap et al., 2008; Van der Knaap et al., 2012). There was variability with respect to the predictive accuracy of risk assessment tools among non-US/Canadian samples. Some studies demonstrated predictive accuracy (Coid et al., 2009) while others did not (de Vogel & de Ruiter, 2005; Van der Knaap et al., 2012). This suggests that: (a) risk assessment tools may not be as valid in other cultures, and/or (b) the methodological quality of the study may have impacted results.

With respect to offender subtypes (violent/sex offenders) there were no analyses conducted in any of the studies for the predictive validity of risk tools for different types of offenders. Given that differences have been found in risk levels for future offending or future violence in different types of offenders, such as sex offenders (Rettenberger, Matthes, Boer, & Eher, 2010) versus violent offenders (Valliant, Gristey, Pottier, & Kosmyna, 1999), this is a particular weakness of the literature reviewed. Only two studies attempted to account for individual differences in terms of ethnicity (Rettinger, 1998; Rettinger & Andrews, 2010) and these did show varied performance of the LSI dependent on the ethnicity of the client assessed.

In terms of measurement bias, there was more heterogeneity between studies. There was variability in follow-up periods between and within studies. For example, the range of follow-up periods varied from 23 days to nearly five years in the sample overall. This made it difficult to make meaningful comparisons and may have contributed to the differences in validity estimates between studies. Additionally, some studies reported a follow-up period using the mean as an estimate with large standard deviations (Coid et al., 2009; Salisbury et

al., 2009) which may suggest that for some offenders the follow-up period was too short to enable a recorded re-offence or violent incident. This is particularly pertinent when assessing violence by women in the community given the proposed underreporting of female crime (Monahan et al., 2001). It should also be noted that some studies did not report follow-up periods at all (Brennan et al., 2009; Folsom & Atkinson, 2007; Schaap et al., 2008; Warren et al., 2005). These measurement issues may have impinged upon validity estimates reported in the review.

There were also differences in terms of how outcome was measured across studies. For recidivism, some studies used re-arrest (Brennan et al., 2009), some used reconviction (Coid et al., 2009), some differentiated reconviction into subtypes (e.g., violent and non-violent) (Folsom & Atkinson, 2007), while others also included 'violation of supervision' as a form of recidivism (Reisig et al., 2006). These were also obtained using official records which have been criticized for being an underestimate of true crime rates (Howitt, 2009). Similarly, violence was variably defined in the studies included in the review with some studies defining community violence as recidivism. Additionally, violence was measured in a sample of offenders where only 10% of their previous offenses were categorized as 'violent' (de Vogel & de Ruiter, 2005).

Other notable issues in the quality of studies reviewed were the lack of inter-rater reliability assessments for scoring risk tools and the reporting standards of the studies. Some studies performed poorly (Reisig et al., 2006) while others performed well (Rettinger & Andrews, 2010). In reporting predictive validity estimates, it is strongly suggested that confidence intervals, standard errors, and significance parameters are reported alongside the effect sizes (Field, 2009; Singh, 2013; Warner, 2008). However, not all studies reported these statistics.

4.4. Implications for clinical practice

In spite of the limitations, an important consideration is an awareness that statistical significance does not always equate to practical significance (Baert, 2005). As such, pragmatism is offered as a framework where the quality of evidence that is put forward and the manner in which it is conducted and contextualized is more worthwhile than statistical significance (Baert, 2005; Finn, Bothe, & Bramlett, 2005). In light of this, the estimates used to calculate predictive validity have been found wanting. In a review of validity performance indicators, Singh (2013) highlights a few concerns. The predictive validity estimates can be categorized in two types; calibration and discrimination. Calibration refers to how well an instrument in predicting risk coincides with observed risk, while discrimination refers to how well an instrument separates those who engage in a particular activity (such as violence or reoffending behaviors) from those who do not (Cook, 2007 as cited in Singh, 2013). AUC and *r* estimates are examples of discrimination indices. Singh (2013) argues that measuring only one of these (i.e., discrimination or calibration) may not provide an accurate account of a tool's predictive capacity. None of the studies in the review used measures of calibration to assess predictive accuracy of risk assessment tools and therefore may not have provided an accurate account of a tool's predictive capacity.

It is highly important that both researchers and practitioners take these considerations into account. Although a tool may not demonstrate predictive accuracy as measured by one particular estimate, this does not imply that the tool should be disregarded. Rather, it suggests that other statistical procedures should be taken into consideration such as those that account for calibration as well as discrimination (Singh, 2013). Equally, it does not imply that the tool has little value clinically. A risk assessment may provide invaluable information on an offender's protective as well as risk factors and can guide the clinician as to targets for

intervention and treatment. However, its ability as an assessment tool remains uncertain without reaching acceptable standards of predictive accuracy.

An important consideration for practitioners and researchers working in forensic risk assessment is acknowledging the potential effects of base rates in accurately predicting an outcome. A risk assessment should predict who will perpetrate an undesirable outcome and who will not. However, the rate at which criminal behavior occurs in the population of interest is critical to determining the level of predictive validity an instrument holds (Szmukler, 2001). When an event is infrequent, it has a low base rate, which makes it extremely difficult to predict. Both violence and recidivism are infrequent events and, as such, have low base rates. The implications for assessing risk and for the predictive utility of risk assessments is best illustrated using an example from Barbaree (1997): In a population of 100,000, assume that 15% are violent and that a risk assessment has 80% accuracy in predicting violence. Here, 12,000 people would be correctly classified as being violent in the future (hits), but 3,000 of the violent people would be predicted as being non-violent (misses). Of those who were not violent, 68,000 would be correctly classified (correct rejections). However, 17,000 of the non-violent people would be misclassified as being violent (false alarms). If sentencing or release decisions depended solely on such tools, 3,000 people predicted to be non-violent would engage in harmful acts whilst 17,000 people would be wrongly detained. Furthermore, with the low base rates of violence and reoffending in offending populations, which are more pronounced in female offenders, even where an offender is categorized as being at 'high risk' of violence or reoffending according to given risk tool they are still less likely to engage in the form of criminal behavior being assessed.

This also means that risk assessments tools based on outcomes with low base rates are less likely to demonstrate predictive accuracy. For instance, base rates affect the correlation coefficient as the range of possible values in determining association between two variables

may be narrower than the conventional ±1.00 if the prevalence of the outcome (such as violence or recidivism) is low. While there are statistical calculations which are independent of base rates, such as the AUC (Rice & Harris, 1995), the significance values obtained may be dependent on sample size and they, therefore, should be interpreted as approximations (Singh, 2013).

Risk may never be predicted with complete accuracy (Hart, Laws, & Kroop, 2003). Risk prediction tools are not sacrosanct, but provide estimates. They are relatively easy to use and establish a common vocabulary that can provide rich knowledge for case management. As such, the ecological validity of risk assessment tools cannot be understated. Statistical findings within the review should thus be evaluated against the quality of the studies. For this review, the variability in the quality across studies, particularly with respect to selection, measurement and reporting bias, implies that attempts to extrapolate and draw conclusions from the findings are difficult.

Accurate risk assessment should: (a) enhance public safety, (b) be financially viable and cost-effective (c) enable the identification of future risk, and (d) enable the identification of treatment targets (Harris & Hanson, 2010). The current review highlights that current tools to assess risk in female offenders have some way to go before they can achieve such targets.

4.5.Implications for future research.

For future research, strategies for improving the quality of studies in this area may include empirical research that: examines the predictive accuracy across subtypes of female offenders, ensures sufficient follow-up for recidivism/violence, assesses both correctional and psychiatric samples either independently or through comparative research, is undertaken in areas outside the US and Canada, and ensures consistency in how the outcome is measured. Future systematic reviews may make more efforts to access non-English language research

and subsequently appraise the differences in predictive accuracy between studies of low and high quality.

In terms of current literature, the review neither supports nor contests the adoption of gender-responsive approaches. Rather, it advocates the need for further research in the area on both gender-specific and gender-neutral risk factors and risk assessment tools. It does, however, give some, albeit limited, confidence to the potential ability of gender-neutral tools to predict risk. As such, rather than questioning whether the origins of female offending is qualitatively different, a more pertinent question may be whether gender-specific research is practically meaningful in the measurement of risk in female offending. The implications of this may be examining whether the items within the tool are adequately capturing risk for female offenders. This would also support the research that highlights that while the risk factors may be similar for males and females, the manner in which the risk factors are expressed may be different for females. The challenge may be to empirically and pragmatically conduct validation studies on gender-specific risk assessment tools as well as gender-neutral risk tools. Research in this area has already begun with the development of the Security Reclassification Scale for Women in Canada (SRSW; Blanchette & Taylor, 2005), the Women's Risk Need Assessment in the US (WRNA; Van Voorhis, Salisbury, Wright, & Bauman, 2008) and the publication of the Female Additional Guidelines (FAM) for the HCR-20 (de Vogel et al., 2012). The task will now be for researchers to demonstrate the predictive validity of such tools.

References

Andrews, D. A. (1982). *The Level of Supervision Inventory (LSI): The first follow-up.* Toronto: Ontario Ministry of Correctional Services.

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

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

Andrews, D. A., Guzzo, L., Raynor, P., Rowe, R. C., Rettinger, J., Brews, A., & Wormith, S. (2012). Are the major risk/need factors predictive of both female and male reoffending?: A test with the eight domains of the level of service/case management inventory. *International Journal of Offender Therapy and Comparative Criminology, 56*, 113-133.

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

Baert, P. (2005). *Philosophy of the Social Sciences*. Cambridge: Polity Press.

Barbaree, H. (1997). Evaluating treatment efficacy with sexual offenders: The insensitivity of recidivism studies to treatment effects. *Sexual Abuse: A Journal of Research and Treatment, 9*, 111-128.

Bonta, J. (1996). Risk-needs assessment and treatment. In A.T. Hartland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp.18-32). Thousand Oaks, CA: Sage Publications.

Blanchette, K. (2002) Classifying female offenders for effective intervention: application of the case-based principles of risk and need. *Forum on Correctional Research, 14*, 31–35.

Blanchette, K., & Brown, S. L. (2006*). The assessment and treatment of female offenders: an integrative perspective*. West Sussex: John Wiley & Sons.

Blanchette, K. & Taylor, K. (2005). *Development and field test of a gender-informed security reclassification scale for women offenders.* Ottawa, ON: Corresctional Service Canada.

Borum, R., Bartel, P., & Forth, A. (2006). *Structured Assessment of Violence Risk in Youth (SAVRY).* Lutz, FL: Psychological Assessment Resources.

Breakwell, G. M., Hammond, S., Fife-Schaw, C., & Smith, J. (2008). Research methods in psychology (3rd ed). London: Sage Publications.

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

Brennan, T., & Oliver, W. L. (2000). *Evaluation of reliability and validity of COMPAS scales: National aggregate sample.* Traverse City, MI: Northpointe Institute for Public Management.

Brews, A. L. (2009). *The Level of Service Inventory and female offenders: addressing issues of reliability and predictive validity* (Unpublished doctoral dissertation). University of Saskatchewan: Saskatoon, Canada.

Critical Appraisal Skills Programme (CASP). (May 2013). Critical Appraisal Checklists. Retrieved from http://www.casp-uk.net/.

Caulfield, L. (2010). Rethinking the assessment of female offenders. *The Howard Journal, 49*, 315-327.

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*, 567-590.

Centre for Review and Disseminations. *Systematic Reviews: CRDs guidance for undertaking reviews in health care*. (2009, January). Retrieved from http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf.

Chesney-Lind, M., & Shelden, R. G. (2004). *Girls, delinquency, and juvenile justice*, (3rd ed). Belmont, CA: Thompson Wadsworth.

Chesney-Lind, M., & Pasko, L. (2013). *The female offender: girls, women and crime* (3rd ed). Thousand Oaks, CA: Sage Publications.

Cohen, J. (1988). *Statistical power analyses for the behavior al sciences* (2nd ed). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., Farrington, D. P., & Rogers, R.D. (2009). Gender differences in structured risk assessment: comparing the accuracy of five instruments. *Journal of Consulting and Clinical Psychology, 77*, 337-348.

Copas, J. B., & Marshall, P. (1998). The Offender Group Reconviction Scale: the statistical reconviction score for use by probation officers. *Journal of Royal Statistical Society, 47*, 159-171.

Corston, J. (2007). *The Corston Report: A review of women with particular vulnerabilities in the Criminal Justice System*. London: Home Office.

Coulson, G., Ilacqua, G., Nutbrown, V., Giulekas, D., & Cudjoe, F. (1996). Predictive utility of the LSI for incarcerated female offenders. *Criminal Justice and behavior* , 23 (3); 427-439.

Craig, L. A., Browne, K. D., & Beech, A. R. (2008). *Assessing risk in sexual offenders: a practitioner's guide.* West Sussex: John Wiley & Sons.

Daly, K. (1994). *Gender, crime, and punishment*. New Haven, CT: Yale University Press.

Davidson, J. T., & Chesney-Lind, M. (2009). Discounting women: context matters in risk and need assessment. *Critical Criminology, 17*, 221-245.

De Vogel, V. (2005). *Structured risk assessment of (sexual violence) on forensic clinical practice. The HCR-20 and SVR-20 in Dutch forensic psychiatric patients.* Amsterdam, The Netherlands: Dutch University Press.

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 Vries Robbé, M. (2013). Working with women: towards a more gender-sensitive violence risk assessment. In Johnstone, L., & Logan, C (Eds.), *Managing Clinical Risk: A Guide to Effective Practice* (pp.224-241). London: Routledge

De Vogel, V., de Vries Robbé, M., van Kalmthout, W., & Place, C. (2012). *Female Additional Manual (FAM): additional guidelines to the HCR-20 for assessing risk for violence in women.* Amsterdam: Van der HoevenKliniek

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., Ogloff, J. R. P., Nicholls, T. L., & Grant, I. (1999). Assessing risk for violence among psychiatric patients: The HCR–20 violence risk assessment scheme and the Psychopathy Checklist— Screening Version. *Journal of Consulting and Clinical Psychology, 67*, 917–930.

Douglas, K. S., Guy, L. S., & Reeves, K. A. (2008). HCR-20 violence risk assessment scheme: Overview and annotated bibliography. Retrieved from http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1362&context=psych_cmhsr.

Egger, M., Dickerson, K., & Smith, D. G. (2007). Problems and limitations in conducting systematic reviews. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care. Meta-analysis in context* (2nd ed., pp. 43-66). London, UK: BMJ Publishing Group.

Effective Public Heath Practice Project (EPHPP). (2008). Quality Assessment for quantitative studies. Retrieved from http://www.ephpp.ca/tools.html.

Farr, K. A. (2000). Classification for female inmates: moving forward. *Crime and Delinquency, 46*, 3-17.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed). London: Sage Publications.

Finn, P., Bothe, A. K., & Bramlett, R. E. (2005). Science and pseudoscience in communication disorders: criteria and applications. *American Journal of Speech-Language Pathology, 14*, 172-186.

Folsom, J., & Atkison, J. L. (2007). The generalisability of the LSI-R and the CAT to the prediction of recidivism in female offenders. *Criminal Justice and Behavior, 34*, 1044-1056.

Funk, S. J. (1999). Risk assessment for juveniles on probation: a focus on gender. *Criminal Justice and Behavior, 26,* 44–68.

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.

Glover, A. J. J., Nicholson, D. E., Hemmati, T., Bernfeld, G. A., & Quinsey, V. L. (2002). A comparison of predictors of general and violent recidivism among high risk federal offenders. *Criminal Justice and Behavior, 29*, 235-249.

Grann, M., Långstrom, N., Tengström, A., & Gunnar, K. (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., Hill, C., McGleish, A., Timmons, D., MacCulloch, M. J. & Snowden, R. J. (2003). Prediction of violence and self harm in mentally disordered offenders: A prospective study of HCR-20, PCL-R, and psychiatric symptomatology. *Journal of Consulting and Clinical Psychology, 71*, 443-451.

Gray, N. S., Taylor, J., & Snowden, R. J. (2008). Predicting violent reconvictions using the HCR-20. *British Journal of Psychiatry, 192,* 384-387.

Greenfeld, L.A., & Snell, T.L. (1999). Women Offenders- Bureau of Justice statistics special report. Retrieved from http://www.bjs.gov/content/pub/pdf/wo.pdf.

Hannah-Moffat, K., & Shaw, M. (2001). *Taking risks: Incorporating gender and culture into the classification and assessment of federally sentenced women in Canada.* Ontario: Status of Women Canada.

Hanson, R. K., & Bussiére, M. T. (1998). Predicting relapse: a meta-analysis of sexual recidivism studies. *Journal of Consulting and Clinical Psychology, 66,* 348-362.

Hare, R D. (1991). *The Hare Psychopathy Checklist—Revised*. Multi-Health Systems: Toronto, ON.

Hare, R. D., Clark, D., Grann, M., & Thornton, D. (2000). Psychopathy and the predictive validity of the PCL-R: an international perspective. *Behavioral Sciences and the Law, 18*, 623-45.

Harris, A. J. R., & Hanson, R. K. (2010). Clinical, actuarial and dynamic risk assessment of sexual offenders: why do things keep changing? *Journal of Sexual Aggression, 16*, 296-310.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315-335.

Hart, S.D. (1998a). Psychopathy and risk for violence. In Cooke, D., Forth, A.E., & Hare, R.D (Eds), *Psychopathy, theory, research and implications for society* (pp. 355-379). Dodrecht, The Netherlands: Kluwer Academic Publications.

Hart, S. D. (1998b). The role of psychopathy in assessing risk for violence: conceptual and methodological issues. *Legal and Criminological Psychology, 3*, 121– 137.

Hart, S. D., Laws, D. R., & Kropp, R. P. (2003). The promise and peril of sex offender risk assessment. In T. Ward, D. R. Laws, & S. M. Hudson (Eds.), *Sexual deviance: Issues and controversies* (pp. 207-205). Thousand Oaks, CA: Sage Publications.

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

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

Home Office. (2007). *The Corston Report: the need for a distinct, radically different, visibly-led, strategic, proportionate, holistic, woman-centred, integrated approach.* London: The stationary office.

Howitt, D.H. (2009). *Introduction to forensic and criminal psychology* (3rd Ed). Essex: Pearson Education Limited.

Juni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal of Epidemiology, 31*(1), 115-23.

Kemshall, H. (2002). *Risk assessment and management of serious violence and sexual offenders: A review of current issues.* Edinburgh: Scottish Executive.

McKeown, A. (2010). Female offenders: assessment of risk in forensic settings. Aggression and Violent Behavior, 15, 422-429.

McNiel, D.E., & Binder R.L. (1994). Screening for risk of inpatient violence: Validation of an actuarial tool. *Law and Human Behavior, 18*, 579–586.

Ministry of Justice. (2014). *Prison Population Figures: 2014*. Population Bulletin-Weekly 25 April 2014. Retrieved from https://www.gov.uk/government/publications/prison-population-figures-2014.

Ministry of Justice. (2012a). Statistics on women in the criminal justice system 2011: A Ministry of Justice publication under Section 95 of the Criminal Justice Act 1991. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220081/statistics-women-cjs-2011-v2.pdf.

Ministry of Justice. (2012b). 2012 Compendium of re-offending statistics and analysis- Ministry of Justice statistics bulletin. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220081/statistics-women-cjs-2011-v2.pdf.

Ministry of Justice. (2011). *Statistics on women in the criminal justice system 2011: a ministry of justice report under Section 95 of the Criminal Justice Act*. London: Ministry of Justice.

Monahan, J., Steadman, H.J., Silver, S., Applebaum, P.S., Robbins, P.C., Mulvey, E.P., . . . Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental illness and violence*. New York: Oxford University Press.

Mossman, D. (1994). Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 64* (2), 783-792.

Nedopil, N. (2009). Gender specific and structured approach to violence risk assessment- the Munich prognosis project (MMP). *Proceedings from 17th EPA Congress, Lisbon, Portugal: European Psychiatry, 24* (1), S148.

Nicholls, T.L., Greaves, C., & Moretti, M. (2008). Female aggression. In Jamieson, A., & Moenssens, A (Eds.). *Wiley Encyclopedia of Forensic Science*, Sussex, UK. Wiley.

Nicholls, T.L., Ogloff, J.R.P., & Douglas, K.S. (2004). Assessing risk for violence among male and female civil psychiatric patients: the HCR-20, PCL-SV, and VSC. *behavior al Sciences and the Law, 22,* 127-158.

O'Shea, L.E., Mitchell, A.E., Picchioni, M.M., Dickens, G.L. (2012). Moderators of the predictive efficacy of the Historical, Clinical and Risk Management-20 for aggression in psychiatric facilities: systematic review and meta-analysis. *Aggression and Violent Behavior, 18* (2), 255-270.

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

Raynor, P. (2007). Risk and need assessment in the British probation: the contributions of the LSI-R. *Psychology, Crime and Law, 13*, 125-138.

Reisig, M.D., Holtfreter, K., & Morash, M. (2006). Assessing recidivism risk in female pathways to crime. *Justice Quarterly, 23* (3), 384-405.

Rettinger, J.L. (1998). *A recidivism follow-up study investigating risk and need within a sample of provincially sentenced women* (Unpublished doctoral dissertation). Carleton University: Ontario, Canada.

Rettinger, J.L., & Andrews, D.A. (2010). Gender risk and need, gender specificity, and the recidivism of female offenders. *Criminal Justice and behavior , 37* (1), 29-46.

Rettenberger, M., Matthes, A., Boer, D.P., Eher, R. (2010). Prospective actuarial risk assessment: a comparison of five risk assessment instruments in difference sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology, 54* (2), 169-186.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737-748.

Salisbury, E.J., Van Voorhis, P., & Spiropoulos, G.V. (2009). The predictive validity of a gender-responsive needs assessment: an exploration study. *Crime & Delinquency, 55* (4), 550-585.

Schaap, G., Lammers, S., & de Vogel, V. (2009). Risk assessment in female forensic psychiatric patients: a quasi-prospective study into the validity of the HCR-20. *Journal of Forensic Psychiatry and Psychology, 20* (3), 354-365.

Scmukler, G. (2001). Violence risk prediction in practice. *British Journal of Psychiatry, 178*, 84-85.

Simpson, S.S., Yahner, J.L., & Dugan, L. (2008). Understanding women's pathways to jail: Analysing the lives of incarcerated women. *The Australian and New Zealand Journal of Criminology, 41*, 84-108. Retrieved from http://crim.umd.edu/sites/ccjs.umd.edu/files/pubs/Pathways.pdf.

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

Singh, J.P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and meta analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31* (3), 499-513.

Skeem, J., Schubert, C., Stowman, S., Beeson, S., Mulvey, E., Gardner, W., &Lidz, C. (2005). Gender and risk assessment accuracy: underestimating women's violence potential. *Law and Human Behavior, 29* (2), 173-186.

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,373 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183-208.

Song, F., Parekh, S., Hooper, L., Loke, Y.K., Ryder, J., Sutton, A.J., Hing, C, Kwok, C.S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment, 14*, S(8).

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risk for sexual and violent recidivism. *Annals of New York Academy of Sciences, 989*, 225-235.

Valliant, P.M., Gristey, C., Pottier, D., & Kosmyna, R. (1999). Risk factors in violence and nonviolent offenders. *Psychological Reports, 85*, 675-680.

Van der Knaap, L. M., Alberda, D. L., Oosterveld, P., & Born, M. P. (2012). The predictive validity of criminogenic needs for male and female offenders: comparing the relative impact of needs in predicting recidivism. *Law and Human Behavior, 36*, 413-422.

Van Montfoort, A., & Relcasserig Nederland. (2004). *RISc version 1.0 Risk Assessment Scales*. Harderwijk, the Netherlands: Flevodruk.

Van Voorhis, P., Salisbury, E., Wright, E., & Bauman, A. (2008). Achieving accurate pictures of risk and identifying gender responsive needs: two new assessments for women offenders. Retrieved from http://www.uc.edu/content/dam/uc/womenoffenders/docs/NIC%20Summary%20Report.pdf.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: the current status of a gender-responsive supplement. *Criminal Justice and Behavior, 37*, 261-288.

Vose, B., Lowenkamp, C.T., Smith, P., & Cullen, F.T. (2009). Gender and predictive validity of the LSI-R: a study of parolees and probationers. *Journal of Contemporary Criminal Justice, 25* (4), 459-471.

Warner, R.M. (2008). *Applied statistics: from bivariate through multivariate analyses*. Thousand Oaks, California: Sage Publications.

Warren, J.I., South, S.C., Burnette, M.L., Rogers, A., Friend, R., Bale, R., & Van Patten, I. (2005). Understanding the risk factors for violence and criminality in women: the concurrent validity of the PCL-R and HCR-20. *International Journal of Law and Psychiatry, 28,* 269-289.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence (version 2).* Burnaby, British Columbia: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

Yang, M., Wong, S.C.P., & Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136* (5), 740-767.

Appendix
Quality Assessment Tool:

Adapted from CASP Critical Appraisal Tools (2010), EHPP (1998) and CRD (2009) guidelines

Step 1:

Threshold Criteria:

- Clear description of female offenders
- Clear description of measurements used
- Clear description of outcome measure
- Sufficient statistical analysis regarding the prediction of recidivism or violence

Step 2:

Assessment of Quality **(only score for relevant items)**	**Guidance for Scoring** **each sub-section**	**Overall rating of quality,**
A. Selection Bias		**Total: /6** **Unclear: /3**
Q1Were the study objectives clear? a)Yes b) Partially c) No d) Unsure		Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q2 Were the participants recruited in an acceptable way? a)Yes b) Partially c) No d) Unsure	***Yes***= participants were appropriately selected, recruitment process described, and ethical principles adhered to (i.e., female offenders, was vulnerability of the population considered) ***Partially*** = meet some of the expectations of the sample, unclear recruitment process ***No*** = no recruitment process not described ***Unsure*** = lack of description to make comprehensive judgement	Yes = 2 Partially = 1 No = 0 Unsure = unclear

Q3 Are the individuals selected to participate in the study likely to be representative of the target population? a)Very likely b) Somewhat likely c) Not likely d) Unsure	***Very likely***= randomly selected from female offending population ***Somewhat likely*** = they are referred from a source list in a systematic manner e.g., clinic, prison, mental health facility ***Not likely***- if they are self-referred ***Can't tell***- if participants characteristics not appropriately described	Very likely = 2 Somewhat Likely = 1 Not likely = 0 Unsure = unclear
B. Measurement Bias		**Total: /18** **Unclear /9**
Q1 Was the operational definition of outcome clearly stated? a)Yes b) Partially c) No d) Unsure	***Yes***- clear definition of types of recidivism (reconviction/rearrest) and/ or violence (eg., verbal, physical) underpinned by strong rationale/theory ***Partially***- recidivism/violence used as outcome but not clearly defined ***No***- no clear definition or rationale for recidivism/violence Unsure – not described	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q2 Were the methods for obtaining the outcome clearly described? a)Yes b) Partially c) No d) Unsure	***Yes-***reliable system for sourcing data described e.g: Recidivism: reconviction data, police records, Violence: police records, hospital records ***Partially-*** sources mentioned but methods on how they were obtained not adequately described or methods but no sources identified ***No-*** no system to measure outcome established ***Unsure-*** authors do not report establishing any system but the method/results suggest they may have	Yes = 2 Partially = 1 No = 0 Unsure = unclear

Q3 Was the outcome measured in the same way across all participants? a)Yes b) Partially c) No d) Unsure	***Yes*-**recidivism/Violence measured in the same way for all participants ***No*-**outcome not measured in same way for all participants ***Unsure*-** measurement of outcome for participants not adequately described	Yes = 2 No = 0 Unsure = unclear
Q4 Was the risk assessment tool administered by trained professionals? a)Yes b) Partially c) No d) Unsure	***Yes*-** trained professionals (psychologists or others trained to administer the tool and/or trainees/researchers under supervision) ***Partially*-**research assistants/trainees with no experience or supervision ***No*-** no professional was trained to administer the tool ***Unsure*-** not adequately described	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q5 Did the authors use multiple sources of information to score risk assessments? a)Yes b) Partially c) No d) Unsure	***Yes*-** multiple sources of information used (file info, interviews, psychometrics, ***Partially*-** more than one source used but not all potential sources (*e.g.*, file info + interview but not psychometric or hospital records) ***No*-** only one source of information used ***Unsure*-** not adequately described	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q6 Were inter-reliability sought for scoring the risk assessments? Was this above .8? a)Yes b) Partially	**Yes-** inter-rater reliability reported for all assessments and was above .8 **Partially-** inter-rater reliability sought for all/some assessments and/or estimate was below .8 or	Yes = 2

c) No d) Unsure	statistic not reported **No-** authors did not seek inter-rater reliability **Unsure-** not sufficiently described to make judgement	Partially = 1 No = 0 Unsure = unclear
Q7 Was the follow-up period sufficiently described & reported? a)Yes b) Partially c) No	**Yes-** follow-up period described and reported **Partially-** follow up period described or follow-up period reported **No-**no follow up period described or reported	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q8 Was the follow-up period long enough to determine outcome defined in the study? a)Yes b) Partially c) No d) Unsure	A follow up period of 2 years is typically deemed sufficient for recidivism studies For violence minimum follow up period = 3 months	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q9 Was missing data dealt with appropriately? a)Yes b) Partially c) No d) Unsure e) N/A	***Yes-*** missing data (if any) was reported and taken into account for risk assessment tool (i.e., not included in analyses or adjustments made) ***Partially-*** missing data was reported but not taken into consideration in measuring risk ***No-*** missing data was not dealt with at all ***Unsure-*** not sufficiently described, study did not report whether there was any missing data ***Not Applicable***-the study did not have any missing data and	Yes = 2 Partially = 1 No = 0 Unsure = unclear N/A = N/A

	reported this	
C. Attrition Bias		**Total: /2** **Unclear /1**
Q1 Were drop-out rates recorded on the studies? a)Yes b) Partially c) No d) Unsure e) N/A	***Yes*-** Drop-out rates recorded & stage of drop-out recorded or not relevant to study ***Partially*-** Drop-out rate reported but stage of drop-out not ***No*-** drop-out rate not recorded ***Unsure*-** not sufficiently described	Yes or N/A = 2 Partially = 1 No = 0 Unsure = unclear
D. Reporting Bias		**Total: /8** **Unclear /4**
Q1 Were appropriate statistical tests used for the research design and question? a)Yes b) Partially c) No d) Unsure	Was the quantitative analysis appropriate for the research? (ROC/AUC statistics, correlations, Multivariate statistics e.g., regressions)	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q2Was the predictive validity of the tests reported (e.g., ROC analyses, incidents of violence) a)Yes b) Partially c) No d) Unsure	***Yes*-** ROC/AUC analyses or DOC, NPV or PPV or correlations reported for recidivism, and/or incidents of violence reported by category (verbal, physical) using above analyses **and** range reported i.e., **(**CIs, SE) ***Partially*-** other statistics used to report recidivism, type of violent incidences not reported or correlations only reported or AUC etc estimates but no range reported (i.e., CI or SE)	Yes = 2 Partially = 1 No = 0 Unsure = unclear

	No- neither recidivism nor rates of violent acts recorded ***Unsure-*** not adequately described	
Q3 Were potential confounders taken into account? a)Yes b) Partially c) No d) Unsure	**Yes-** any or most of potential confounders were taken into consideration **Partially-** Some efforts made to control for confounders **No-** no effort made to control for potential confounders **Unsure-** not enough information given	Yes = 2 Partially = 1 No = 0 Unsure = unclear
Q4 Can the results be generalized to other female offending populations? a)Yes b) Partially c) No d) Unsure	Can recidivism/violence be predicted in other female populations? Consider age, ethnicity, offender type, correctional (prison/community) vs psychiatric sample	Yes = 2 Partially = 1 No = 0 Unsure = unclear
E. Clinical Judgement/ Pragmatism		**Total: /2** **Unclear: /1**
Q1 Is the study worth continuing? a)Yes b) Maybe c) No d) Unsure	Based on the overall study does the study have credibility? Do you believe the results? Is the design of the study sufficiently flawed to render the results unreliable? Also consider ***Pragmatism***: are there any benefits to research *and* practitioners to continuing studies of this nature?	Yes = 2 Maybe = 1 No = 0 Unsure = unclear

Quality Score	**/36**	**Unclear**	**/18**

Tables and Figures

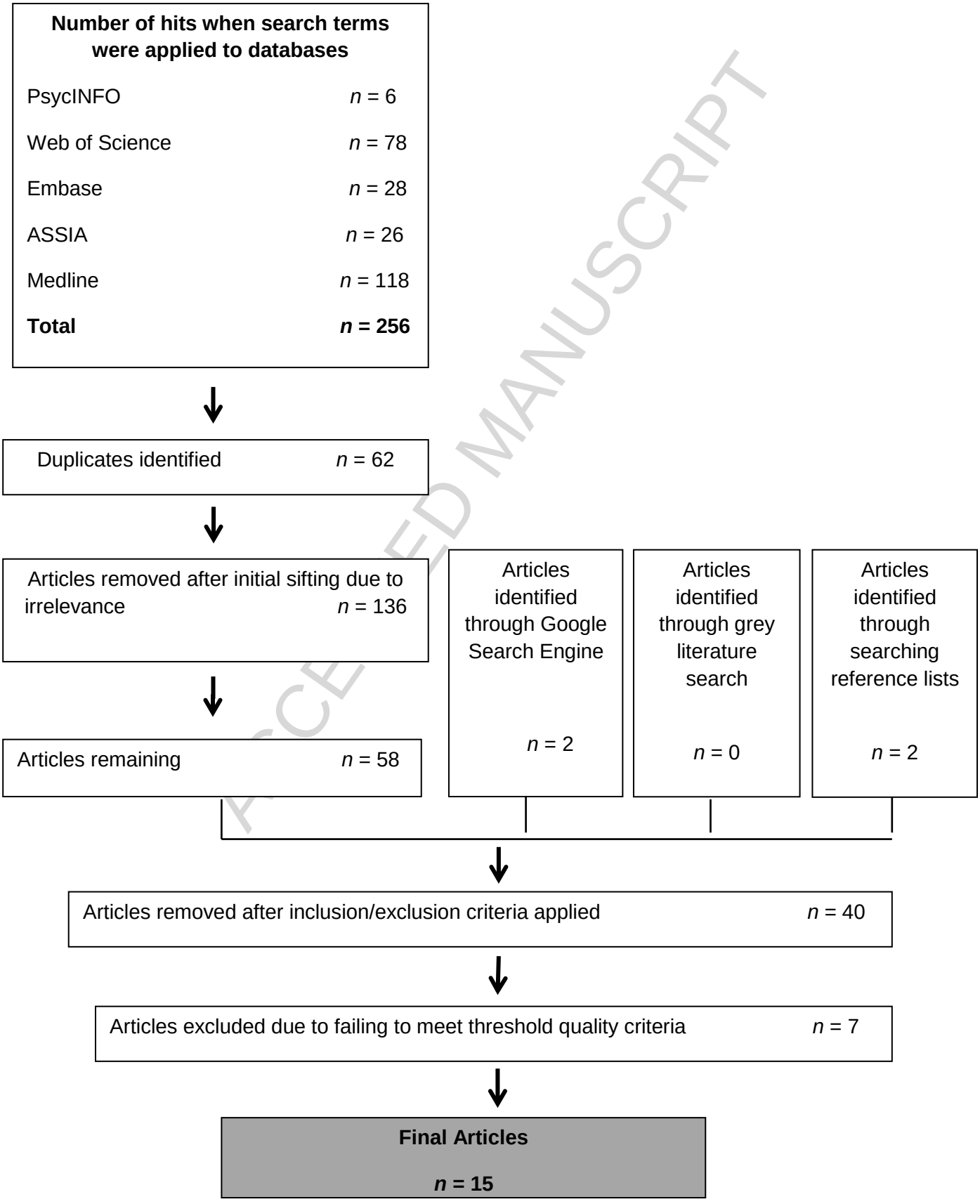

*Figure 1***.** Flow chart of search process

Table 1.

Summary table of predictive validity of instruments and study characteristics

Risk Tool	Study	Study Type	Population		Outcome	Follow-up period	Rate of Recidivism/Violence	Quality Score	AUC (Type of behavior)
			Sample Size	Sample Type					
CAT-SR	Folsom & Atkinson (2007)	Prospective	*N* =100	Correctional	Recidivism	Not reported	38%	22	*Any criminal behavior* AUC = .68 (C I= ±.12, *r* .30, *p* < .01) *Non-Violent behavior* AUC = .61 (CI = ±.13, *r* .14, n.s) *Violent behavior* AUC = .68 (CI =

									±.14, *r* .23, *p* < .05)
COMPAS	Brennan et al. (2009)	Prospective	*N* = 449	Correctional	Recidivism	Not reported	Not reported	21	*Any Recidivism* AUC = .68 *Person* AUC = .78 *Felony* AUC = .68
HCR-20	Coid et al. (2009)	Prospective	*N* = 304	Correctional	Recidivism	2 years	88%	27	*Violent* AUC = .70 (95% CI = .60 - .80) *Acquisitive* AUC =.62 (95% CI =

								.53 - .69) *Any* AUC =.67 (95% CI = .60 -.73)
De Vogel & de Ruiter (2005)	Prospective + Retrospective	*N* = 42	Psychiatric	Violence	Not reported	Recidivism 13% Inpatient Violence 30%	21	AUC =.59 (SE = .11, *r* = .22, n.s)
Warren et al. (2005)	Retrospective	*N* = 132	Correctional	Violence/Recidivism	Not reported	Not reported	27	*Violent Crime* AUC = .55 (SE = .06, 95% CI = .43

									- .67) *Nonviolent crimes* AUC = .68 (SE =.06, 95% CI = .56 -.79)
	Schaap et al. (2009)	Retrospective	*N* = 45	Psychiatric	Recidivism	Not reported	36%	19	*Violent Recidivism* AUC = .54 (SE = .12) *General Recidivism* AUC = .55 (SE = .09)
LSI	Rettinger (1998)[1]	Retrospective	*N* = 441	Correctional	Recidivism	4.75 years	46.5%	29	*General Recidivism*

									AUC = *.93* *Violent Recidivism* *AUC* = *.85*
LSI-R	Folsom & Atkinson (2007)	Prospective	*N* =100	Correctional	Recidivism	Not reported	38%	22	*Any* AUC = .67 (CI = ± .12) *Nonviolent* AUC .62 (CI = ± .12) *Violent* AUC = .67 (CI = ± .18)

Reisig et al. (2006)[2]	Retrospective	*N* = 411	Correctional	Recidivism	1.5 years	35-50%	21	*r* =07, n.s
Rettinger & Andrews (2010)	Retrospective	*N* = 411	Correctional	Recidivism	4.75 years	General Recidivism 45% Violence Recidivism 13%	27	*General Recidivism* *r* = .63, n.s *Violent Recidivism* *r* = .44, n.s
Salisbury et al. (2009)	Retrospective	*N* = 134	Correctional	Recidivism	3.67 years	54.5%	21	*Prison Misconduct* *r* = .12, *p*<.10 *Technical Violations* *r* = .18, *p*< .05 *Any failure*

								r = .21, p<.001**
								Rearrest r not given
Vose et al. (2009)[2]	Retrospective	N = 401	Correctional	Recidivism	3.79 years	Not reported	18	*Time 1*
								r = .11, p<.05
								Time 2
								r = .20, p<.01
Van Voorhis et al. (2010)	Retrospective	N = 356	Correctional	1) Violence(Prison Misconducts) 2) Recidivism	1.42 years	Not reported	16	*Institutional Misconduct*
								1) AUC = .58, p<.05
								2) AUC = .68, p<.01

									Recidivism 1) AUC = .72, *p*<.01 2) AUC = .71, *p*<.01
LS/CMI	Rettinger & Andrews (2010)	Retrospective	*N* = 411	Correctional	Recidivism	4.75 years	General Recidivism 45% Violent Recidivism 13%	27	*General Recidivism* AUC = .87 (95% CI = .83-.90) *Violent Recidivism* AUC = .86 (95% CI = .82-.91)
LSI-OR	Brews	Retrospective	*N* =	Correctional	Recidivism	2 years	28.3%	25	*AUC* .78 (CI =

	Risk assessment in female offenders (2009)		2831						±.18)
	Rettinger (1998)[1]	Retrospective	*N* = 441	Correctional	Recidivism	4.75 years	46.5%	29	*General Recidivism* *AUC = .93* *Violent Recidivism* *AUC = .85*
PCL-R	Coid et al. (2009)	Prospective	*N* = 304	Correctional	Recidivism	2 years	88%	27	*Violent* AUC = .73 (95% CI = .63-.83) *Acquisitive* AUC = .63 (95% CI = .53-.72) *Any*

								AUC = .67 (95% CI = .64-.74)
de Vogel & de Ruiter (2005)	Prospective & Retrospective	*N* = 42	Psychiatric	Violence	Unclear	Recidivism 13% Inpatient Violence 30%	21	AUC = .34, *r* = -21, n.s
Warren et al. (2005)	Retrospective	*N* = 132	Correctional	Violence/Recidivism	Not reported	Not reported	27	*Violent Crimes* AUC = .55 (SE = .06, 95% CI = .43-.67) *Non Violent Crimes* AUC = .67

									(SE = .06, 95% CI = .56-.79)
	Schaap et al. (2009)	Retrospective	*N* = 45	Psychiatric	Recidivism	Not reported	36%	19	*Violent Recidivism* AUC = .57 (SE = .11) *General Recidivism* AUC = .60 (SE=.09)
OGRS-II	Coid et al. (2009)	Prospective	*N* = 304	Correctional	Recidivism	2 years	88%	27	*Violent* AUC = .54 (95% CI = .43-.66) *Acquisitive* AUC = .69 (95%

									CI = .61-.77)
									Any
									AUC = .68 (95%
									CI= .61-.74)
RISc	Van der Knaap et al. (2012)	Retrospective	*N* = 1691	Correctional	Recidivism	2 years	22.4%	23	AUC = .68 (95%
									CI = .64-.71)
RM2000 V	Coid et al. (2009)	Prospective	*N* = 304	Correctional	Recidivism	2 years	88%	27	*Violent*
									AUC = .66 (95%
									CI = .55-.77)
									Acquisitive
									AUC = .61 (95%
									CI = .55-.70)
									Any

									AUC = .62 (95% CI = .55-.69)
VRAG	Coid et al. (2009)	Prospective	*N* = 304	Correctional	Reconviction	2 years	88%	27	*Violent* AUC = .65 (95% CI = .55-.75) *Acquisitive* AUC = .66 (95% CI = .59-.74) *Any* AUC = .66 (95% CI = .59-.72)

Hastings et al. (2011)	Retrospective + Prospective	*N* = 145	Correctional	Recidivism	1 year	49.4%	25	(a)Arrests AUC = .62 (95% CI = .49-.77) (b)Undetected offenses AUC = .61 (95% CI = .49-.73) (c)(a)OR (b) AUC = .66 (95% CI = .54-.78) (d)Violent Arrests/Undetected AUC = .66 (95% CI = .47-.85)

[1] Used Regression and RIOC analysis. First author calculated AUC from R^2, [2]Violence coded by offence but reoffending not measured.

Table 2.

Quality scores for each study

Study	Selection Bias *(out of 6)*		Measurement Bias *(out of 18)*		Attrition Bias *(out of 2)*		Reporting Bias *(out of 8)*		Clinical Judgement *(out of 2)*		Overall Quality Score *(out of 36)*	
	Total	Unclear	Total	Unclear	Total	Unclear	Total	Unclear	Total	Unclear	Total	Unclear
Brennan et al. (2009)	4	0	10	2	1	0	5	1	1	0	21	2
Brews (2009)	3	1	12	2	0	0	8	0	2	0	25	3
Coid et al. (2009)	5	0	15	0	1	0	5	0	1	0	27	0
De Vogel & de Ruiter (2005)	4	0	7		1	0	7	0	2	0	21	4
Folsom & Atkinson (2007)	4	0	11	0	2	0	4	0	1	0	22	0
Reisig et al. (2006)	4	0	12	1	1	0	3	0	1	0	21	1

Rettinger (1998)	5	0	16	0	0	0	6	0	2	0	29	0
Rettinger & Andrews (2010)	4	0	14	0	1	0	7	0	1	0	27	0
Salisbury et al. (2009)	4	0	10	0	1	0	5	0	1	0	21	1
Schaap et al. (2008)	3	0	9	3	1	0	5	0	1	0	19	3
Vose et al. (2009)	4	0	5	4	1	0	6	0	2	0	18	4
Van der Knaap et al. (2012)	4	0	13	0	1	1	4	0	1	0	23	0
Warren et al. (2005)	5	0	14	3	1	0	6	0	1	0	27	2
Hastings et al. (2011)	5	0	14	1	1	0	4	1	1	0	25	2
Van Voorhis et al. (2010)	4	0	5	3	1	1	5	0	1	0	16	4

Highlights

(max of 85 characters, core results to be presented)

1. We present a systematic review of risk assessment tools for female offenders
2. Fifteen studies were included which assessed nine risk assessment tools
3. The quality of studies were systematically appraised
4. The LSI demonstrated most accuracy for assessing violence and recidivism
5. The implications of results for assessing risk in female offenders is considered