Final version published as: Viljoen, J. L., Vargen, L. M., Cochrane, D. M., Jonnson, M. R., Goossens, I., & Monjazeb, S. (2021). Do structured risk assessments predict violent, any, and sexual offending better than unstructured judgment? An umbrella review. Psychology, Public Policy, and Law, 27(1), 79–97. https://doi.org/10.1037/law0000299

Do Structured Risk Assessments Predict Violent, Any, and Sexual Offending Better than Unstructured Judgment? An Umbrella Review

Jodi L. Viljoen, Lee M. Vargen, Dana M. Cochrane, Melissa R. Jonnson, Ilvy Goossens, and Sanam Monjazeb

Simon Fraser University

Psychology, Public Policy, and Law

Author Note

Jodi L. Viljoen, Lee M. Vargen, Dana M. Cochrane, Melissa R. Jonnson, Ilvy Goossens, & Sanam Monjazeb. Department of Psychology, Simon Fraser University. Funding for this review was provided by Simon Fraser University/Social Sciences and Humanities Research Council of Canada's Institutional Grants Program (23142). The funders had no role in the study design, data collection, data analysis, or manuscript composition. We thank Shanna Li and Julia Schillaci-Ventura for data entry and editing and Dr. Matthew Sigal for statistical assistance with the ggplot2 package in R.

Jodi L. Viljoen designed the study, wrote the protocol, conducted analyses, and wrote the manuscript. Lee M. Vargen, Dana M. Cochrane, Melissa R. Jonnson, Ilvy Goossens, and Sanam Monjazeb extracted data from the studies, conducted analyses, and contributed to the final manuscript. Also, Lee M. Vargen and Dana M. Cochrane conducted the systematic searches. Jodi L. Viljoen is an author of an adolescent risk assessment tool. This tool was not included in any of the studies in this umbrella review, and the author does not receive any proceeds from the sale of this tool.

Correspondence concerning this article should be addressed to Jodi L. Viljoen, Department of Psychology, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, V5A 1S6 Canada. Email: jviljoen@sfu.ca

Abstract

Although it is widely believed that risk assessment tools lead to more accurate estimates of risk of violence and offending than unstructured clinical judgments, the nature and quality of evidence that supports this view is unclear. As such, we conducted an umbrella review of systematic reviews. Through a search of 15 databases, we identified nine systematic reviews, including six meta-analyses and three narrative systematic reviews, that compared unstructured and structured risk judgments for any, violent, and sexual offending. Each review was independently coded by two raters. Raters also coded the 46 primary studies on unstructured judgment included in these reviews. Although the reviews concluded that structured risk judgments are superior to unstructured judgments, the data supporting these conclusions have limitations. None of the systematic reviews directly compared risk assessment tools to unstructured judgments. In addition, two-thirds of the primary studies included in the systematic reviews were from the 1980s or earlier, and 89% had serious methodological limitations that created a high risk of bias. In many cases, the primary studies did not examine unstructured judgments per se but instead used proxies such as legal and administrative decisions. As such, there is a pressing need for an updated systematic review that focuses on direct comparison studies and carefully addresses study limitations. To address this gap, we have initiated a preregistered systematic review (PROSPERO ID: CRD42020187585).

Keywords: risk assessment, violence, offending, clinical judgment, risk assessment tools

Do Structured Risk Assessments Predict Violent, Any, and Sexual Offending Better than Unstructured Judgment? An Umbrella Review

Mental health and criminal justice professionals are often asked to judge the likelihood that people will engage in future violence or offending. These risk assessments play a central role in many legal decisions, such as decisions about pretrial detention, incarceration, involuntary psychiatric hospitalization, and transfer to adult court (Skeem & Monahan, 2011; United States 115th Congress, 2018). The consequences of these assessments are serious; a failure to identify people who pose a high risk could jeopardize public safety. Conversely, erroneously deeming people to be high risk when they are not could lead to unjustified stigma and unnecessary restrictions to liberty. However, accurately assessing risk of future violence or offending is not an easy task, especially when clinicians do not have structured assessment devices to guide them.

In the mid-1950s, Meehl (1954) demonstrated the fallibility of unstructured clinical judgments. Meehl argued that clinical data, such as interviews or psychometric tests, could be combined using only one of two methods: (1) a formal, mechanical approach that involves explicit rules, such as an equation or actuarial table or (2) an informal, nonmechanical approach (herein referred to as unstructured judgment). In his review of 20 studies, Meehl found that mechanical approaches "were either approximately equal or superior to those made by a clinician" in all but one study (p. 119).

Meehl's work sparked a considerable amount of research and debate in many areas of psychology, including violence risk assessment. Consistent with his assertions, early reviews highlighted that unstructured risk judgments were susceptible to error. For instance, in 1974, Ennis and Litwack concluded that predictions of dangerousness are "usually wrong" (p. 719) and equated these assessments to "flipping coins in the courtroom" (p. 694). Similarly, based on his

review of the five studies that had been conducted at that time, Monahan (1981) estimated that clinicians were accurate only one-third of the time at best. In other words, they were "wrong at least twice as often as they were right" (Monahan, 1984, p. 10).

The American Psychiatric Association (1983) presented these early findings to the United States Supreme Court in the case of *Barefoot v. Estelle* (1983). In its amicus brief, it argued that, given the inherent difficulty of accurately predicting violence, psychiatrists should not be permitted to provide expert testimony on long-term dangerousness. However, the Court dismissed this assertion, concluding that it had "no merit" (p. 882), reasoning that, "it makes little sense, if any, to submit that psychiatrists, out of the entire universe of persons who might have an opinion on the issue, would know so little about the subject that they should not be permitted to testify" (p. 897). The Court further asserted that a decision to bar psychiatric testimony about dangerousness would be akin to asking them to "disinvent the wheel" (p. 896).

In the years to follow, several additional court cases reinforced the legitimacy of unstructured risk predictions in guiding legal decisions. In *Schall v. Martin* (1984), the United States Supreme Court affirmed that judges' unstructured predictions of reoffense risk can be used as a basis for preventative detention for adolescents, noting that "there is nothing inherently unattainable about a prediction of future criminal conduct" and that "such a judgment forms an important element in many decisions" (p. 45). They further noted that risk predictions are based on numerous factors and thus "cannot be readily codified" (p. 47). In *Kansas v. Hendricks* (1997), the United States Supreme Court upheld the constitutionality of sexually violent predator statutes that included dangerousness as a criterion. In effect, rather than negating efforts to predict violence and reoffending, these cases served to endorse and extend the use of such predictions in court. Accordingly, researchers shifted away from arguing why risk assessments

should not be conducted to focusing on how to make these assessments as accurate as possible (Monahan, 1984).

To achieve these improvements, researchers directed their efforts towards the development of risk assessment tools. These instruments structure the risk assessment process by providing direction on which risk factors to consider and how to rate or score these factors (see Skeem & Monahan, 2011). Studies conducted in the 1990s reported promising results, demonstrating that instruments were able to predict violence and reoffending at statistically significant levels (Borum, 1996). Spurred on by these findings, the development of risk assessment tools quickly gained momentum. Two main types of tools emerged: actuarial tools, which prompt assessors to calculate a total score using a predetermined formula, and structured professional judgment tools, which allow assessors to combine information using their own discretion.

As research support for risk assessment tools grew, some researchers concluded that the debate between tools and unstructured judgment had been indisputably resolved in favor of tools. Most notably, Quinsey et al. (1998) asserted that actuarial methods should completely replace existing practices, arguing that "actuarial methods are too good and clinical judgment too poor to risk contaminating the former with the latter" (p. 171). However, questions about the relative merit of unstructured judgment lingered. At that time, some authors pointed out that many of the early studies that had been cited in the rush to reject clinical assessments of dangerousness suffered from serious methodological limitations (Litwack, 2001; Melton et al., 1997). For instance, many studies that were used to argue that unstructured risk judgments were inaccurate did not, in fact, examine risk judgments. Instead, they examined administrative decisions, such as release decisions, inferring that such decisions were tantamount to a risk prediction (Litwack,

2001). Mossman (1994) reanalyzed earlier research using receiver operating characteristic analyses and concluded that “clinicians are able to distinguish violent from nonviolent patients with a modest, better-than-chance level of accuracy" (p. 790). In addition, a rigorous study by Lidz et al. (1993) indicated that the accuracy of unstructured risk judgments was not as dismal as previously thought.

Despite the finding that unstructured judgments were not as poor as believed, actuarial approaches continued to outperform unstructured judgements in those studies (Gardner et al., 1996; Mossman, 1994). Reassured by these results, researchers created new tools for a variety of populations (e.g., adolescents, adults), settings (e.g., pretrial, probation), and forms of offending (e.g., violent, sexual, general). By 2010, researchers had developed over 400 risk assessment tools (Singh et al., 2014). Furthermore, many justice and mental health agencies throughout the world had adopted these tools, using them to guide a multitude of decisions, such as decisions about pretrial detention, sentencing, security level, treatment, and release. For instance, by 2009, 88% of American pretrial detention agencies (Pretrial Justice Institute, 2009), and 87% of residential sex offending treatment programs (McGrath et al., 2010) had implemented risk assessment tools.

In contrast to the rapid uptake and extensive research on risk assessment tools, research on unstructured judgment slowed down substantially, grinding nearly to a complete halt. In 2011, the American Psychological Association (APA) submitted an amicus brief to the courts asserting that unstructured judgment was no longer a useful or recommended approach. In this court case, a psychiatrist, Dr. Coons, concluded that the defendant posed a high risk of reoffending based solely on his unstructured judgment (*Coble v. Texas*, 2011). In addition, Dr. Coons, at the time of his assessment, had not met with the defendant in 18 years. In its amicus brief, the APA argued

that unstructured judgments should not be relied on and instead encouraged the use of structured risk assessment tools, arguing that these approaches "can be scientifically reliable and provide a modest advantage over unstructured approaches" (p. 19). Nevertheless, the defendant's petition for the case to be heard by the U.S. Supreme Court was denied and his death penalty sentence was upheld.

Although debates about the superiority of risk assessment tools over unstructured risk judgments seemed to briefly disappear from the limelight at around that time, renewed criticisms of structured approaches have recently begun to reappear. Many of these criticisms have emanated from legal scholars and policymakers. Specifically, as risk assessment tools have gained widespread use in high stakes sentencing and preventative detention decisions, some authors have questioned the underlying evidence for these tools (see Monahan & Skeem, 2016). For instance, legal scholar Starr (2014) argued that although the superiority of risk assessment tools is interpreted as "gospel," there is "no persuasive evidence" that tools lead to more accurate predictions of recidivism than judges' unstructured judgments (p. 807), asserting that:

> [W]hile scores of studies have found that actuarial prediction methods outperform clinical judgment, this finding is not universal, the average accuracy edge is not drastic, and the vast majority of studies are from wholly different contexts (such as medical diagnosis or business failure prediction). (p. 807)

As another example, the Pretrial Justice Institute (2020), once a supporter of risk assessment tools, recently concluded that pretrial risk tools "can no longer be part of our solution for building equitable pretrial justice systems" (p. 1) due to concerns that tools may deepen racial and ethnic inequities.

Two recent studies have added further fuel to debates about the value of risk assessment

tools. In a study by Dressel and Farid (2018), laypeople recruited via Amazon's Mechanical Turk were provided with descriptions of real-world defendants and asked to assess these defendants' risk. Despite their lack of specialized training, laypeople's unstructured predictions showed similar accuracy to that of a commercial risk assessment tool, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS; Northpointe Institute for Public Management, 1996). Although these findings are provocative, Lin et al. (2020) pointed out that it is unclear how well that study approximates real world practice. For example, contrary to what occurs in practice, participants were provided with immediate feedback on the accuracy of each of the 50 predictions they made before making the next prediction. Thus, participants could learn from this feedback, which practicing clinicians are seldom, if ever, able to receive. Due to concerns about Dressel's and Farid's methodology, Lin et al. (2020) conducted a new study and found that when participants were not provided with feedback, risk assessment tools significantly outperformed unstructured judgments.

In sum, rather than fading into history, debates about the superiority of risk assessments over unstructured judgments have resurfaced and are growing in intensity. In fields where views are polarized and findings vary, systematic reviews can help ground debates, as they can provide a careful, rigorous, and transparent examination of evidence. To date, the predictive validity of risk assessment tools has been the focus of many systematic reviews, including both quantitative systematic reviews (i.e., meta-analyses) and narrative systematic reviews. These systematic reviews demonstrate that risk assessment tools can significantly predict violence and other forms of reoffending with moderate effect sizes (Olver et al., 2014; Singh et al., 2011; Yang et al., 2010). However, what is less clear is the extent to which reviews demonstrate that tools are superior to unstructured judgments. This question is key, as the belief that tools are superior is

the core assumption upon which the current risk assessment practices rest. As such, in this study we examined the following research questions:

1. How many and what type of prior systematic reviews have compared the predictive validity of structured and unstructured risk judgments for violent, any, and sexual offending?
2. What are the findings of these systematic reviews?
3. What are the limitations of these reviews and the studies included in these reviews?

To answer these questions, we conducted an umbrella review. Umbrella reviews, or overviews of systematic reviews, are a form of research synthesis in which researchers synthesize the results of systematic reviews (Higgins et al., 2019). Umbrella reviews provide researchers with a systematic method to generate a "big picture" overview of a body of research and identify gaps and limitations in knowledge (Hunt et al., 2018; Ng & Benedetto, 2016). As such, the popularity of umbrella reviews has grown tremendously during the past 5 years (Hossain, 2020).

By conducting an umbrella review, we aimed to identify research gaps and limitations, and determine whether a new systematic review and new primary studies comparing unstructured judgment to risk assessment tools may be needed. Thus, we used standardized study appraisal measures to evaluate existing reviews. In addition, we evaluated the quality of the primary studies included in these reviews. Although examining the limitations of primary studies goes beyond the scope of a standard umbrella review, it enabled us to gain a better understanding of specific limitations of this body of research in order to establish an agenda for future research.

Method

Prior to initiating our umbrella review, we preregistered our review with the International Prospective Register of Systematic Reviews (PROSPERO ID = CRD42019132461; Booth et al.,

2012). In our preregistered protocol, we specified our review questions, search strategy, inclusion/exclusion criteria, risk of bias assessment, and analytic plan. In conducting this review, we adhered to recommended practices, such as those in the Cochrane Collaboration guidelines (Becker & Oxman, 2011) and the Assessing the Methodological Quality of Systematic Reviews 2 tool (Shea et al., 2017). In addition, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (Moher et al., 2009). Throughout this article, we use the term "review" to encompass meta-analyses and narrative systematic reviews.

Inclusion and Exclusion Criteria

To be included, a review needed to (1) be a meta-analysis or narrative systematic review that used an explicit, reproducible methodology (e.g., provide a list of search terms and databases), and (2) compare the predictive validity of unstructured and structured risk assessment for predictions of violent, any, or sexual offending. Given that risk assessments are conducted within a variety of contexts (e.g., civil commitment, release from prison), we did not restrict our umbrella review to specific populations or settings.

We defined unstructured risk assessment as assessments in which the assessor "selects, measures, and combines risk factors and produces an estimate of violence risk solely according to his or her clinical experience and judgment" (Skeem & Monahan, 2011, p. 39). In contrast, with structured risk assessments, assessors are provided with a list of risk factors and rating criteria. Our definition of structured risk assessment encompassed (a) risk assessment tools (i.e., instruments that were designed for use in real-world practice) and (b) statistical models (i.e., prediction models with two or more predictors [Wolff et al., 2019] that were used solely for research purposes rather than clinical practice). No restrictions were imposed on publication date, publication status, or language; we included published and unpublished works, and English

and non-English studies. When two reports were based on the same review, we selected the most comprehensive report.

Search Strategy

We searched four repositories that contain systematic reviews (e.g., Cochrane Database of Systematic Reviews), and 11 broader databases (e.g., PsycINFO) using the following search terms: (clinical or unstructure* or unaid*) and (predict* or judgmen* or judge*) and "risk assessment" and (violen* or reoffen* or recidiv* or offen* or crime*) and (meta-analy* or metaanaly* or systematic review) (see Figure 1). Searches were conducted on April 22, 2019 and updated on March 23, 2020. Consistent with recommended practices, our Google Scholar search (conducted March 23, 2020) examined the first 250 search records (Haddaway et al., 2015). We also hand-searched the reference lists of included reviews.

Screening and Full Text Review

After removing duplicates, a total of 679 reports were identified. Two authors (LV, DC) reviewed the abstracts and titles to determine if they met eligibility for full-text review. Twenty percent of abstracts (n = 106) from the April 22, 2019 search were independently screened by the two authors. The interrater agreement was 92.5%, indicating a high level of consistency in screening decisions. Discrepancies were discussed and resolved through consensus. Once a report was screened in (n = 61), we conducted a full-text review to determine if it met inclusion criteria. The most common reason that reports were excluded is that they did not examine unstructured judgments (see Figure 1). A study-by-study list of reasons for exclusion is provided in the Supplemental Materials.

Data Extraction from Systematic Reviews and Meta-Analyses

Each review was independently coded by two raters. Prior to beginning data extraction,

raters completed a 6-hour training which included a didactic session and two practice cases. All five raters (DC, IG, LV, MJ, SM) were graduate students in psychology who had completed coursework and practicums related to risk assessment, and four of the five raters had prior experience in conducting systematic reviews. Interrater reliability for the complete set of reviews (n = 9) was calculated using Cohen's (1960) kappa (κ) coefficients for categorical variables and intraclass correlation coefficients (ICCs) for continuous variables (one-way random effects model, absolute agreement, average measures; Hallgren, 2012). We interpreted κ coefficients as follows: slight (0.00 – 0.20), fair (0.21 – 0.40), moderate (0.41 – 0.60), substantial (0.61 – 0.80), and almost perfect (0.81 – 1.00; Landis & Koch, 1977). In addition, ICCs were interpreted with the following guidelines: poor (< 0.40), fair (0.40 – 0.59), good (0.60 – 0.74), excellent (0.75 – 1.00; Cicchetti, 1994). Discrepancies in ratings were discussed and resolved through consensus.

Data extraction form. We developed a 39-item data extraction form through a review of previous studies and guidelines (e.g., Moons et al., 2014; Singh & Fazel, 2010), as well as pilot testing with three reviews. The form included items pertaining to the characteristics of the report (e.g., systematic review or meta-analysis), study eligibility, search procedures, data collection (e.g., number of databases searched), and results (e.g., aggregated effect size). Raters showed perfect agreement about type of review (i.e., systematic review or meta-analysis; κ = 1.00), number of databases searched (ICC = 1.00), attempts to search for unpublished data (κ = 1.00), search date (ICC = 1.00), and aggregated effect sizes (ICC = 1.00).

AMSTAR 2. The Assessing the Methodological Quality of Systematic Reviews 2 (AMSTAR 2; Shea et al., 2017) is an appraisal tool for meta-analyses and systematic reviews. It includes 16 items (e.g., Did the review authors use a comprehensive literature search strategy?), scored on a two-point (yes/no) or three-point (yes, partial, no) scale, that help to assess bias

arising from confounds, selection of studies, methodological flaws, analytic weaknesses, and reporting issues. Previous research has found that the AMSTAR 2 shows moderate reliability and adequate validity (Lorenz et al., 2019). In the present study, ICCs for the total score fell in the excellent range (ICC = 0.84, *n* = 9; Cicchetti, 1994).

ROBIS. Although the AMSTAR 2 assesses numerous aspects of study quality, it does not generate an overall risk of bias rating. As such, we also used the Risk of Bias in Systematic Reviews tool (ROBIS; Whiting et al., 2016). Bias refers to the "presence of systematic error in a study that leads to distorted or flawed results and hampers the study's internal validity" (Moons et al., 2019, p. W4). The ROBIS examines four domains of potential biases: study eligibility, identification and selection of studies, data collection and study appraisal, and syntheses of findings. After rating bias on a three-point scale (low, high, unclear) for each domain, raters are prompted to make an overall risk of bias rating. Although reviews are typically rated as high overall risk of bias if there is a bias in one or more domains, raters can lower their risk of bias rating if they believe the authors of the review have addressed all these concerns in their interpretation of the findings. In the current study, only one of these adjustments occurred. The interrater reliability for the overall rating was in the good range prior to this adjustment (ICC = 0.67). However, the ICC for the post-adjustment ratings dropped to 0.34 because, for the one case in which an adjustment was made, one rater believed the review authors had adequately addressed the domain concerns whereas the other rater did not. Discrepancies in bias ratings per domain and total risk of bias were resolved through consensus.

Data Extraction from Primary Studies

After identifying and coding reviews, we created a list of all primary studies that the reviews used to make conclusions about the predictive validity of unstructured versus structured

risk assessments; this included studies that compared unstructured judgments to risk assessment tools or statistical models (i.e., direct comparisons), and studies that reported unstructured judgments alone (i.e., indirect comparisons). We made an a priori decision to extract data from the primary studies because we anticipated that few if any reviews would systematically appraise the quality of primary studies. Each primary study was coded by two raters out of the pool of five raters (DC, IG, LV, MJ, SM), and discrepancies were resolved through consensus. Raters completed an 8-hour training which included a didactic session and four practice cases. Interrater reliability for the complete set of primary studies (n = 46) was examined using κ coefficients and ICCs (one-way random effects model, absolute agreement, average measures; Hallgren, 2012).

Data extraction form. We developed a 65-item data extraction form through a review of protocols used in previous studies (e.g., Guy, 2008) and pilot testing with four cases. It included variables pertaining to characteristics of the report (such as whether it was peer-reviewed), sample, method, outcomes, and results. The ICC for sample size fell in the excellent range (ICC = 0.94). Similarly, κ coefficients for the dichotomous variables fell in the perfect or almost perfect range for age of the sample (i.e., adult vs. adolescent, κ = 1.00), sex of sample (i.e., male vs. mixed, κ = 1.00), population type (i.e., justice vs. mental health, κ = 0.95), source of outcome data (i.e., justice records vs. other sources of information, κ = 0.93), and study design (i.e., direct comparison to risk assessment tools, direct comparison to statistical models, and unstructured clinical judgment only, κ = 0.89).

PROBAST. The Prediction model Risk of Bias ASsessment Tool (PROBAST; Wolff et al., 2019) is a risk of bias tool for prediction modelling studies. Raters code risk of bias in four domains (participants, predictors, outcome, and analysis), and subsequently rate the study's overall risk of bias as low, unclear, or high. The PROBAST also prompts raters to examine the

applicability of the study in three domains (participants, predictors, and outcome), and then rate overall concerns about applicability. Applicability pertains to the study's relevance to the research question at hand (Wolff et al., 2019, p. 54). For instance, if a study did not examine clinical judgments of risk but instead examined general prognosis, it may have limited applicability to the questions of whether structured risk judgments outperform unstructured risk judgments. In the current study, the interrater reliability for applicability was fair to excellent (ICC = 0.79, 0.54, 0.63, 0.66, and 0.53 for participants, unstructured and structured predictors, outcome, and overall). However, interrater reliability for risk of bias ratings fell within or near the poor range for a few items (ICC = 0.43, 0.30, 0.23, and 0.41 for participants, unstructured and structured predictors, and analyses). This may be due, in part, to the restricted range of ratings (e.g., 89.1% of studies were rated as high overall risk of bias). As such, our team further clarified PROBAST's rating instructions and recoded items with poor reliability. Interrater reliability for risk of bias ratings improved, with ICCs falling in the good to excellent range (ICC = 0.90, 0.90, 0.95, 0.70, 0.93, and 0.96 for participants, unstructured and structured predictors, outcome, analysis, and overall, respectively). However, key findings remained consistent (i.e., 89.1% were rated as high overall risk of bias both prior to and after recoding).

Data Analyses

The list of included reports is provided in the reference list. To examine the characteristics and quality of reviews, we calculated descriptive statistics using IBM SPSS, Version 26.0. We examined the findings of the reviews using two approaches.

First, we extracted aggregated effect sizes from the meta-analyses. To enable comparisons across meta-analyses, we converted all results to area under the curve (AUC) of the receiver operating characteristic curve (Hanley & McNeil, 1982). AUC values are the most

widely used statistic in risk assessment research and represent the probability that a randomly selected recidivist will score higher on the measure than a randomly selected non-recidivist (Helmus & Babchishin, 2017). Z*r* scores were converted to *r* using formulas in Card (2012) and then DeCoster's (2012) Excel spreadsheet was used to transform these *r* values to AUCs. Cohen's *d* scores were converted to AUCs using the tables provided by Salgado (2018) and cross-referenced with DeCoster's (2012) Excel spreadsheet for conversions. To ensure accuracy and reproducibility, these conversions were made independently by two raters (IG, MJ) and checked by each other. Of note, umbrella reviews do not empirically aggregate prior meta-analyses' aggregated results into a single effect size (Becker & Oxman, 2011); doing so would be inappropriate because reviews typically include overlapping studies (Singh & Fazel, 2010). Instead, to visually present results, we created a figure in RStudio using the ggplot2 package to show the range of AUC values reported by meta-analyses (Wickham, 2016). AUC values were interpreted as follows: 0.56 (small), 0.64 (medium), 0.71 (large; Rice & Harris, 2005).

Second, we conducted directed content analysis to examine the conclusions made by the authors of the reviews (Hsieh & Shannon, 2005). Prior to coding, we developed three initial codes based on the literature: (1) structured judgment outperforms unstructured approaches, (2) the difference between structured and unstructured approaches is not as large as proclaimed, and (3) the relative difference between approaches is moderated by other factors, such as the type of tool. Raters (IG, MH, SM) read through the abstract and discussion of each review with attention to the conclusions offered by the authors. They then annotated the conclusions with a priori codes using NVivo 12 software (QSR International, 2018). Each review was independently coded by two raters. The κ coefficients were .77, .73, and .77, respectively, indicating substantial agreement ($k = 9$; Landis & Koch, 1977). The raters also used an inductive approach to identify

new codes. Based on this procedure, one new code was added (i.e., the relative superiority of structured approaches is robust to moderators, $\kappa = .61$).

Results

Systematic Reviews and Meta-Analyses

Characteristics. We identified six meta-analyses and three narrative systematic reviews that compared the predictive validity of structured and unstructured clinical judgments of violent, any, and sexual offense risk (see Table 1). All but one of these reviews were conducted over a decade ago and were published in peer-reviewed journals. On average the reviews included seven studies on unstructured judgments of risk for offending ($M = 7.33$, SD = 5.48, $Mdn = 6.00$), but this ranged from 0 to 17.

None of the reviews were designed primarily to compare unstructured judgments to risk assessment tools. Instead, they had broader aims: two meta-analyses examined the predictive validity for a variety of clinical judgments, such as vocational assessments and "diagnoses of homosexuality" (Ægisdottir et al., 2006; Grove et al., 2000, p. 23), one meta-analysis examined risk factors for offending among offenders with mental disorders (Bonta et al., 1998), and another was a methods-oriented paper conducted to demonstrate AUC analyses (Mossman, 1994). Given that many of these meta-analyses were conducted prior to the development of risk assessment tools, most of the meta-analyses ($k = 4$) examined statistical models of risk that were calculated for research purposes rather than tools that were developed for clinical use. Only two meta-analyses examined risk assessment tools per se (Guy, 2008; Hanson & Morton-Bourgon, 2009), and in those meta-analyses the comparison between tools and unstructured judgment was indirect. In indirect comparison studies, researchers include all eligible studies that examine unstructured judgment *or* a risk assessment tool (Takwoingi et al., 2015). In contrast, in direct

comparison studies, researchers include only those studies that compare unstructured judgment *and* a tool in the same sample.

Study Appraisal and Risk of Bias. Although some of the reviews were influential and high in quality for their time, they did not meet criteria that are now considered important. Based on AMSTAR 2 ratings, none of the reviews included an explicit statement that review methods were established a priori (Item 2), provided a list of potentially relevant studies that were excluded and reasons for exclusion (Item 7), assessed risk of bias (Item 9), examined impact of risk of bias on meta-analytic results (Item 12), or reported possible conflicts of interest (Item 16; see Table 2). In addition, none of the reviews met full criteria for using a comprehensive search strategy because they did not search reference lists of included studies, consult with content experts, search grey literature, and/or provide justifications for restrictions (e.g., English language restrictions; Item 4).

On the ROBIS, three reviews were rated as showing a high overall risk of bias (33.3%), four were unclear (44.4%), and two were low risk (22.2%; see Table 3). The most common bias was in the Data Collection and Appraisal domain. In this domain, 88.9% of reviews were rated as high or unclear risk of bias ($k = 8$). This is largely because none of the reviews formally appraised the quality of the primary studies. In addition, 77.8% ($k = 7$) of the reviews were rated as high or unclear risk of bias in the Selection of Studies because they did not search at least two databases, use search methods other than databases, or have a process whereby at least two reviewers selected studies.

Results of Meta-Analyses. Five of the six meta-analyses (83.3%) found that structured judgments had significantly higher predictive validity than unstructured judgments for at least one outcome (i.e., any, violent, or sexual offending; see Table 4 and Figure 2). In the remaining

meta-analysis (i.e., Grove et al., 2000), structured judgments also showed higher predictive validity than unstructured judgments of offense risk, but it was unclear if this reached significance as the exact statistical value was not reported. The aggregated AUCs for unstructured judgment fell in the small range for all three meta-analyses that examined any offending (100%), the one meta-analysis on sexual offending (100%), and two of the three meta-analyses on violent offending (66.7%). In the remaining meta-analysis on violent offending (i.e., Mossman, 1994), the AUC for unstructured judgment fell in the medium range. Notably, three meta-analyses (75.0%) found significant heterogeneity in the predictive validity of unstructured judgment (i.e., Bonta et al., 1998; Hanson & Morton-Bourgon, 2009; Mossman, 1994; cf. Guy, 2008), indicating that the effect sizes varied between studies.

Content Analysis of Conclusions. Next, we conducted a content analyses of the review authors' conclusions. These analyses expanded on our summary of meta-analytic findings in two ways: (1) they captured the narrative systematic reviews in addition to the meta-analyses, and (2) they added depth to our analyses by examining the authors' interpretations of their findings.

In three of the nine reviews, the difference between structured and unstructured judgment was not a focal point and thus was not discussed in the authors' conclusions (i.e., Blank, 2001; Guy, 2008; Nicholls et al., 2013). For instance, Nicholls et al. (2013) did not identify any studies that examined the predictive validity of unstructured judgments of intimate partner violence, and as such, this issue was not discussed. In each of the remaining six reviews, the authors concluded that structured judgments showed higher predictive validity than unstructured judgments (i.e., Ægisdóttir et al., 2006; Bonta et al., 1998; Grove et al., 2000; Hanson & Morton-Bourgon, 2009; Mossman, 1994; Turgut et al., 2006).

That said, in two meta-analyses, the authors reported that the difference between

structured and unstructured judgment was not as large or as consistent as commonly believed. Specifically, in their meta-analysis of broad types of clinical decisions, Ægisdóttir et al. (2006) noted that the evidence for the relative superiority of statistical measures is "not overwhelming" (p. 367) and encouraged "more temperance on both sides" of the debate (p. 367). Grove et al. (2000) offered similar conclusions.

In three reviews, the authors concluded that the superiority of structured versus unstructured judgment was moderated by other variables. For instance, various unstructured judgments showed poorer predictive validity relative to structured judgments when clinicians were provided with clinical interview data only (Grove et al., 2000), but the difference between unstructured and structured judgments was smaller when clinicians also had information about base rates (Ægisdóttir et al., 2006). In discussing violence risk assessments more specifically, Hanson and Morton-Bourgon (2009) concluded that structured judgments were still more accurate than unstructured judgments regardless of methodological factors or study design. They also reported that empirical actuarial tools showed greater improvement over unstructured judgment than did structured professional judgment tools but noted a lack of studies that directly compared these approaches.

Primary Studies on Unstructured Judgment

After accounting for nine studies that were included in multiple reviews, the reviews included 50 separate studies on unstructured judgment. Four studies (a) consisted only of raw data and did not have any accompanying reports (i.e., Reddon et al., 1996; Schiller, 2000), (b) had been erroneously identified in a prior meta-analysis as examining unstructured clinical judgment when in fact they did not (i.e., Klassen & O'Connor, 1989), or (c) used the same sample as a more comprehensive study that was already included (i.e., Lidz et al., 1993). As

such, based on our a priori criteria, we excluded these four studies. We examined the characteristics and quality of the remaining 46 studies which included 11,174 participants.

Characteristics. Nearly two-thirds of the primary studies included in the reviews were published in the 1980s or earlier (65.2%, *k* = 30; see Table 5), and most studies were with adult samples (85.4%, *k* = 35). In approximately one-third of studies, the sample comprised people in justice settings, such as prisons or forensic hospitals (39.1%, *k* = 18). In the remaining studies, the sample consisted of people in general mental health settings such as psychiatric hospitals (33.3%, *k* = 15), or males in treatment for sexual offending (28.9%, *k* = 13). Typically, researchers measured violence and offending outcomes using official justice records (75.6%, *n* = 31), although in some cases hospital records were used (22.0%, *k* = 9). Few studies examined time to reoffense, offense severity, or level of harm. In nearly half of studies (48.6%, *k* = 17) the follow-up period was 1 to 5 years, but in 17.1% of studies the follow-up period was 7 days or less (*k* = 6), and in 25.7% of studies it was more than 5 years (*k* = 9).

Only a small proportion of the studies included in the reviews directly compared unstructured judgments with risk assessment tools in the same sample (26.1%, *k* = 12). Instead, most studies examined only unstructured judgment without making comparisons to a structured approach (54.3%, *k* = 25). In addition, in nine studies (19.6%), researchers compared unstructured judgments with statistical models that were developed for research purposes. The contexts in which unstructured and structured judgments were conducted differed. In almost all cases (84.4%, *k* = 38), unstructured risk assessments were conducted by professionals who were making judgments as part of real-world practice. In contrast, structured risk ratings were always made by researchers who were making judgments for research purposes only (100%, *k* = 22).

In addition, although some studies did examine assessors' explicit unstructured

judgments of risk (32.6%, *k* = 15), most studies relied upon proxy variables to make inferences about unstructured judgments. Specifically, in 41.3% of studies (*k* = 19), researchers made inferences about assessors' unstructured judgments based on legal decisions. For instance, if an individual was recommended for release it was assumed that this meant that the assessors considered that person to be low risk. In 26.1% of studies (*k* = 12), researchers used other proxies for unstructured risk judgments, such as by examining general prognosis or whether the patient had completed treatment.

Risk of Bias and Applicability. On the PROBAST, 89.1% (*k* = 41) of the primary studies were rated as having a high overall risk of bias (see Figure 3). The individual rating for each study is provided in the Supplemental Materials. One of the most common domains in which biases occurred was the Analyses domain, in which 76.1% (*k* = 35) of studies were rated as having a high risk of bias. Many studies (*k* = 18) presented only basic frequencies or proportions, such as the proportion of released people who reoffended, without any acceptable statistical indicators of performance (e.g., AUCs, survival analyses, regression; Singh, 2013). In general, the studies that compared structured versus unstructured judgments (*k* = 12) did not test whether the predictive validity of these approaches differed significantly using statistical analyses such as z-tests. Instead, they based their conclusions on a simple visual inspection of the results. Many studies were also rated as having a high risk of bias in the Structured Predictors domain (45.5%, *k* = 10). This is because, in nine of the 10 studies that compared unstructured judgment to statistical models, these models typically had not been cross validated in another sample, leading to the possibility of overfitting.

In addition to concerns about risk of bias, raters had high concerns about the overall applicability of 30.4% (*k* = 14) of the primary studies (see Figure 3). The most common domain

in which concerns arose was in how well the researchers' operationalization of unstructured judgments truly captured unstructured judgments of risk; 28.3% (k = 13) of studies were rated as presenting high or unclear concerns in this domain. For instance, in four studies, evaluators in the unstructured judgment condition were given results from standardized tests (e.g., Minnesota Multiphasic Personality Inventory [MMPI], Rorschach). Evaluators then used this information to make an "unstructured" clinical judgment. Thus, in those studies, the unstructured judgment was not truly unstructured according to common definitions (see Hanson & Morton-Bourgon, 2009). In five studies, researchers made inferences about professionals' unstructured judgments by listening to meetings or reading clinical reports and looking for statements such as whether a person is predisposed to violence (Dix, 1976). Although these sources may include references to risk, it is not discussed in a consistent manner which makes it challenging to convert these qualitative descriptions into a common metric. Similarly, in four studies, researchers examined clinicians' general prognoses rather than risk ratings per se (e.g., Glaser, 1955). As such, these studies may have limited applicability or relevance to the question of whether structured risk judgments are superior to unstructured risk judgments.

Discussion

In this umbrella review, we identified nine systematic reviews that compared the accuracy of risk assessment tools to unstructured judgment. Although these systematic reviews provide some indication that risk assessment tools outperform unstructured judgments, the single most important conclusion from our review is that the quality of this evidence is limited. One of the most obvious problems is that prior reviews were conducted at least 10 to 25 years ago. As a result, they do not capture newer studies (e.g., Lin et al., 2020), nor do they adhere to practices that methodologists now recommend (Shea et al., 2017; Whiting et al., 2016). For instance, none

of the reviews used a risk of bias tool to assess methodological flaws in the primary studies that they included, and few offered justifications for search restrictions (such as decisions to restrict their search to certain dates or languages) or used search methods other than databases (such as by contacting subject experts).

Another major issue is that several reviews did not focus exclusively on risk assessment tools. More specifically, although the Ægisdóttir et al. (2006) and Grove et al. (2000) meta-analyses have been cited in support of structured judgment in hundreds of risk assessment studies, they included only 10 or fewer studies on offending, and those studies focused on statistical models that were used for research purposes rather than risk assessment tools per se. The two meta-analyses that examined risk assessment tools (Guy, 2008; Hanson & Morton-Bourgon, 2009) did not provide direct, head-to-head comparisons between tools and unstructured judgments. Instead, different sets of studies were used to estimate effect sizes for unstructured judgment and tools. This means that observed differences between approaches could be due, in part, to methodological differences between studies such as differences in follow-up length or sample type (see Takwoingi et al., 2015).

When we examined the primary studies included in the systematic reviews, the limitations of this body of evidence became even more apparent. In fact, only one-third of primary studies included in the reviews measured an "unstructured judgment" in which the assessor "selects, measures, and combines risk factors and produces an estimate of violence risk solely according to his or her clinical experience and judgment" (Skeem & Monahan, 2011, p. 39). Instead, 41.3% of studies examined legal and administrative decisions (e.g., release decisions). This is problematic because in making case processing decisions, judges consider not only risk to the public, but also other factors such as defendants' blameworthiness and practical

constraints such as jail capacity (Maddan & Hartley, 2018). In an additional 26.1% of studies, unstructured judgment was measured using other approaches which did not fully align with common definitions of unstructured judgment (i.e., Hanson & Morton-Bourgon, 2009; Skeem & Monahan, 2011). For instance, in four studies, clinicians were provided with MMPI-2 or Rorschach results prior to making an "unstructured judgment." Thus, contrary to Skeem's and Monahan's (2011) definition, they did not select and measure risk factors solely based on their judgment.

Besides these problems with the applicability or relevance of primary studies, 89.1% of primary studies showed a high risk of bias or, in other words, barriers to internal validity (Moons et al., 2019). One problem was that, in 90.0% of the studies that compared risk assessment tools to statistical models, the models had not been cross validated, meaning that the estimates of their predictive validity may have been inflated (Wolff et al., 2019). Statistical models have been shown to have much higher AUC values when they are fitted to the sample in which they were developed than when they are cross-validated (AUCs = .89 and .71 for non-cross-validated and cross-validated samples, respectively; Mossman, 1994). Another problem is that whereas structured judgments were always made for research purposes, in most studies (84.4%), unstructured judgments were made as part of real-world practice. As such, the comparison between approaches was susceptible to bias because for unstructured judgment, a treatment effect could have occurred wherein assessors may have made efforts to mitigate risk, thereby proving their predictions wrong and attenuating their predictive validity.

Although this body of research is clearly imperfect, it is tempting to ask whether the evidence that risk assessment tools outperform unstructured judgments is nevertheless "good enough." By current standards, the answer to this question is no. According to the GRADE

framework (Grading of Recommendations Assessment, Development, and Evaluating Work Group; Guyatt et al., 2011), which is widely used in medicine and other fields, the level of evidence for a prognostic test is downgraded if problems occur in any of the following domains: the studies in the field show risk of bias, are indirect, find inconsistent results, are imprecise, or demonstrate publication bias (Iorio et al., 2015). In this umbrella review, we found that, although research is generally consistent in reporting that risk assessment tools are superior to unstructured judgment, studies used to support this statement showed serious problems in terms of risk of bias and a lack of direct comparison.

Limitations

In interpreting these findings, some caveats are important to note. Although we searched 15 different databases, it is possible that we missed some less well-known reviews. In addition, although we were able to locate almost all the primary studies included in the reviews (96.0%), one review included two unpublished datasets that did not have accompanying written materials, and our attempts to contact their authors were unsuccessful.

Another limitation is that although two raters coded all reviews and primary studies and raters generally demonstrated good interrater reliability, a couple of PROBAST items had poor reliability. Given that the PROBAST is a new instrument which is designed specifically for prediction modelling studies, no other research has examined its interrater reliability. However, even in exemplary umbrella reviews (e.g., Cochrane reviews), researchers have found that that it can be challenging to achieve high interrater reliability on risk of bias tools due to the complex nature of these ratings (Hartling et al., 2013; Pollock et al., 2017). As such, we clarified and recoded several PROBAST items, which resulted in acceptable interrater reliability.

Future Directions

Although risk assessment tools continue to have greater research support than unstructured judgments, this review suggests that researchers and practitioners should be careful to avoid overly zealous and exaggerated claims about the relative merit of risk assessment tools versus unstructured risk judgments. Similarly, they should avoid resting their conclusions about the superiority of tools on older reviews that are dated and have significant methodological limitations. Instead, new research is needed. Most pressing, there is a need for an updated systematic review that captures new studies (e.g., Lin et al., 2020) and adheres to current best practices (e.g., Shea et al., 2017). To address this gap, we have initiated a pre-registered meta-analysis (PROSPERO ID: CRD42020187585). Rather than relying on indirect evidence, this meta-analysis will focus on studies that directly compare unstructured judgments to risk assessment tools. It will also test moderators, such as whether the relative superiority of tools is smaller or larger for studies that are high in methodological quality, or varies depending on the context (e.g., sentencing decisions, civil commitment).

In addition to an updated meta-analysis, new primary studies are needed that overcome the limitations of prior research. These new studies should measure unstructured judgment in a direct manner (e.g., asking clinicians to provide unstructured judgments on a 10-pt. scale) rather than relying on convenient but crude proxies (e.g., dichotomous release decisions). Furthermore, to provide a fair comparison between unstructured judgments and risk assessment tools, studies must hold constant other variables, such as who is conducting the assessment (e.g., clinician, research assistant) and the nature of the assessment (e.g., whether or not the assessment is being used to guide real-world treatment and risk management decisions). However, achieving this methodological control may be difficult. As such, we recommend that researchers use a combination of experimental designs (e.g., Lin et al., 2020), which provide greater rigor and

control, and field studies, which typically offer greater generalizability to real-world contexts. To minimize the potential for biases, researchers should preregister their studies, as is becoming the norm in psychological research (Nosek et al., 2018).

Despite the widespread belief that risk assessment tools are superior to unstructured judgment, research has not yet provided clear explanations for why this might be. As such, researchers need to delve deeper into potential mechanisms. One possibility is that tools may help ensure that assessors consider the right information, namely factors that research has shown to predict recidivism. A second possibility is that risk assessment tools might help assessors better combine information, such as by placing the correct weight on each factor (see Meehl, 1956). A third possibility is that tools might help mitigate certain social cognitive biases (see Neal & Grisso, 2014). For instance, assessors' unstructured judgments may be heavily swayed by their first impressions of an evaluee (i.e., anchoring bias), and they may subsequently seek out information that confirms their initial hypotheses (i.e., confirmation bias). Risk assessment tools may reduce these biases by reorienting assessors and ensuring they consider a broader set of risk factors. In addition, although research shows that that clinicians tend to overpredict violence (Melton et al., 2018; Monahan & Cummings, 1974; i.e., base rate neglect), some tools might help reduce assessors' tendency to overestimate reoffense rates by providing reoffense rate norms. To date, these possibilities have not been adequately tested.

Research on the impact of risk assessments tools on racial and ethnic biases is also urgently needed (Shepherd & Anthony, 2018; Vincent & Viljoen, 2020). Black, Indigenous, and People of Color (BIPOC) are incarcerated at much higher rates than White people (Mauer, 2011; Roberts & Reid, 2017), and some policymakers and scholars have raised important questions about whether tools might increase these disparities (Holder, 2014; Starr, 2014). However,

testing racial and ethnic biases is difficult because the outcome measures themselves can be biased (Vincent & Viljoen, 2020). For instance, Black and Indigenous people are more heavily policed than White people, leading to inflated arrest rates (Clark, 2019; Pierson et al., 2020). As such, researchers could test whether risk assessment tools outperform unstructured judgments when reoffending is measured by self-report rather than arrests. Another challenge in examining racial and ethnic biases is that such biases can be insidious and embedded within the items included in tools. For instance, some tools include multiple items related to arrest history, which can be impacted by police biases, whereas other tools do not focus on such items. Therefore, researchers should compare results by tool rather than lumping all tools together (Vincent & Viljoen, 2020). Finally, even if predictive validity looks comparable on the surface (e.g., AUCs are equivalent across groups), the nature of errors might differ between groups (see Muir et al., 2019 for a discussion). For instance, White people may be more likely to be incorrectly judged as low risk (i.e., false negatives), whereas Black people may be more likely to be incorrectly judged as high risk (i.e., false positives). Given that mean differences in scores across groups could lead to disparate and unfair consequences (see Skeem & Lowenkamp, 2018), researchers need to examine calibration and types of errors rather than focusing solely on tools' ability to discriminate between people who do and do not reoffend.

Thus far, many of the discussions about the impact of tools on racial and ethnic disparities have failed to compare tools to the alternative approach, namely unstructured judgments. This comparison is important because, like tools, unstructured judgments are not immune to racial and ethnic biases. Indeed, some studies suggest that when professionals rely on unstructured judgments, they assume that Black youth are more dangerous than White youth (Graham & Lowery, 2004), and they are more likely to attribute crimes committed by Black

youth to internal deficits rather than situational factors (Bridges & Steen, 1998).

Not only does research need to compare how the use of risk assessment tools affects risk *predictions*, so too does research need to directly test the impact of tools on risk *management* outcomes, such as whether the use of tools improves treatment-planning, reduces unnecessary use of incarceration, and decreases rates of violence (Viljoen & Vincent, 2020). A recent meta-analysis found that when agencies adopt risk assessment tools, incarceration rates showed modest declines, but further research is needed to determine whether the size of declines vary across racial and ethnic groups (Viljoen et al., 2019). In addition, although some studies have reported that the use of tools can improve treatment-planning and reduce rates of violence, the results are scarce and variable (Viljoen et al., 2018).

Summary

In sum, although researchers and policymakers often assume that the superiority of risk assessment tools over unstructured judgments is a conclusion that is backed by a mass of rigorous studies, our umbrella review indicates that the evidence for this belief is not as strong as assumed. To be clear, we are not saying that risk assessment tools do not significantly predict violence and reoffending; many studies demonstrate that they do (e.g., Singh et al., 2011). Nor are we saying that risk assessment tools are *not* more accurate than unstructured judgment; we hypothesize that they are. Instead, our conclusion is that, to ensure that any such claims are sound and verifiable, we need more rigorous research. Given the recently renewed debates about risk assessment, alongside the serious impacts that these assessments can have on public safety and people's liberty, it is time to revisit and strengthen the body of evidence that underlies current practices in risk assessment.

References

Note. * References marked with an asterisk indicate reviews that were included in the umbrella review. § References marked with a § indicate primary studies that were included in the reviews.

*Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of Clinical Judgment Project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34*(3), 341-382. https://doi.org/10.1177/0011000005285875

Augimeri, L. K., Koegl, C. J., Webster, C. D., & Levene, K. S. (2001). *Early Assessment Risk List for Boys (EARL-20B): Version 2*. Earlscourt Child and Family Centre.

American Psychological Association. (2011). *Brief for Amici Curiae, Coble v. Texas, 564 U.S. 1020* (2011) (No. 10-1271). https://www.apa.org/about/offices/ogc/amicus/coble

American Psychiatric Association (1983). *Brief for Amicus Curiae, Barefoot v. Estelle, 463 U.S. 880* (1983) (No. 82-6080). https://www.psychiatry.org/File%20Library/Psychiatrists/Directories/Library-and-Archive/amicus-briefs/amicus-1982-barefoot.pdf

Barefoot v. Estelle, 463 U.S. 880, 883 (1983).

Becker, L., & Oxman, A. (2011). Chapter 22: Overviews of reviews. In J. P. T. Higgins and S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Cochrane Collaboration. https://handbook-5-1.cochrane.org/

§Bengtson, S., & Långström, N. (2007). Unguided clinical and actuarial assessment of re-offending risk: A direct comparison with sex offenders in Denmark. *Sexual Abuse: A Journal of Research and Treatment, 19(2),* 135–153. https://doi.org/10.1007/s11194-007-9044-5

*Blank, A. (2001). Patient violence in community mental health: A review of the literature. *British Journal of Occupational Therapy, 64(12),* 584–589. https://doi.org/10.1177/030802260106401202

§Bloom, J. D., Williams, M. H., Rogers, J. L., & Barbur, P. (1986). Evaluation and treatment of insanity acquittees in the community. *Bulletin of the American Academy of Psychiatry & the Law, 14*(3), 231–244.

*Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123*(2), 123–142. https://doi.org/10.1037/0033-2909.123.2.123

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: An international prospective register of systematic reviews. *Systematic Reviews, 1*(1), 2. https://doi.org/10.1186/2046-4053-1-2

Borum, R. (1996). Improving the clinical practice of violence risk assessment: Technology, guidelines, and training. *American Psychologist, 51*(9), 945–956. https://doi.org/10.1037/0003-066X.51.9.945

Borum, R., Bartel, P., & Forth, A. (2006). *Manual for the Structured Assessment for Violence Risk in Youth (SAVRY).* Psychological Assessment Resources.

Bridges, G. S., & Steen, S. (1998). Racial disparities in official assessments of juvenile offenders: Attributional stereotypes as mediating mechanisms. *American Sociological Review, 63*(4), 554–570. https://doi.org/10.2307/2657267

Card, N. A. (2012). *Methodology in the social sciences: Applied meta-analysis for social science research.* Guilford Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and

standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Coble v. Texas, 2011 U.S. LEXIS 4622, 564 U.S. 1020, 131 S. Ct. 3030, 180 L. Ed. 2d 846, 79 U.S.L.W. 3710 (U.S. June 20, 2011).

§Cocozza, J. J., & Steadman, H. (1976). The failure of psychiatric predictions of dangerousness: Clear and convincing evidence. *Rutgers Law Review, 29(5),* 1084-1101.

Clark, S. (2019). Overrepresentation of Indigenous people in the Canadian criminal justice system: Causes and responses. Research and Statistics Division, Department of Justice Canada. https://www.justice.gc.ca/eng/rp-pr/jr/oip-cjs/oip-cjs-en.pdf

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

§de Vogel, V., de Ruiter, C., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health, 3*(2), 149-165. https://doi.org/10.1080/14999013.2004.10471204

DeCoster, J. (2012). *Converting effect sizes* [Excel spreadsheet]. http://www.stat-help.com/spreadsheets/Converting%20effect%20sizes%202012-06-19.xls

§Dix, G. E. (1976). Differential processing of abnormal sex offenders: Utilization of California's Mentally Disordered Sex Offender Program. *Journal of Criminal Law and Criminology, 67(2),* 233-243.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), 1–5. https://doi.org/10.1126/sciadv.aao5580

§Enebrink, P., Långström, N., Hultén, A., & Gumpert, C. H. (2006). Swedish validation of the

early assessment risk list for boys (EARL-20B), a decision aid for use with children presenting with conduct-disordered behaviour. *Nordic Journal of Psychiatry, 60*(6), 438–446. https://doi.org/10.1080/08039480601021795

Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review, 62*(3), 693-752.

Fazel, S., P. Singh, J., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *British Medical Journal, 345*(7868), 17. https://doi.org/10.1136/bmj.e4692

§Florida Department of Health and Rehabilitative Services. (1985). *Status of the sex offender treatment programs, fiscal year 1983–1984.*

§Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Consulting and Clinical Psychology, 64*(3), 602–609. https://doi.org/10.1037/0022-006X.64.3.602

§Glaser, D. (1955). The efficacy of alternative approaches to parole prediction. *American Sociological Review, 20(3),* 283–287. https://doi.org/10.2307/2087386

§Glaser, D., & Hangren, R. F. (1958). Predicting the adjustment of federal probationers. *National Probation and Parole Association Journal, 4*(3), 258-267.

Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior, 28*(5), 483–504. https://doi.org/10.1023/B:LAHU.0000046430.65485.1f

*Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19–

30. https://doi.org/10.1037/1040-3590.12.1.19

*Guy, L. S. (2008). Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey. [Unpublished doctoral dissertation, Simon Fraser University]. http://summit.sfu.ca/item/9247

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., … Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE*, *10*(9), 1–17. https://doi.org/10.1037/1040-3590.12.1.19

§Hall, G. C. N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology, 56*(5), 773–775. https://doi.org/10.1037/0022-006X.56.5.773

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143(1),* 29–36. http://doi.org/10.1148/radiology.143.1.7063747

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism* (User report 1997-04). Department of the Solicitor General of Canada.

*Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments

for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1–21. https://doi.org/10.1037/a0014421

Hanson, R. K., & Thornton, D. (1999). *Static–99: Improving actuarial risk assessments for sex offenders* (User Rep. No. 1999–02). Department of the Solicitor General of Canada.

Hanson, R. K., & Thornton, D. (2003). Notes on the development of Static-2002 (Rep. No. 2003–01). Public Safety and Emergency Preparedness Canada.

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(4),* 315–335. https://doi.org/10.1177/0093854893020004001

Harris, G. T., Rice, M. E., Quinsey, V. L., & Cormier, C. A. (2015). *Violent offenders: Appraising and managing risk* (3rd edition). American Psychological Association. https://doi.org/10.1037/14572-007

Hartling, L., Hamm, M. P., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., Tsertsvadze, A., Hempel, S., Shekelle, P., & Dryden, D. M. (2013). Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology*, *66*(9), 973–981. https://doi.org/10.1016/j.jclinepi.2012.07.005

Hathaway, S. R., & McKinley J. C. (1942). *Manual for the Minnesota Multiphasic Personality Inventory.* University of Minnesota Press.

Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior*, *44*(1), 8–25. https://doi.org/10.1177/0093854816678898

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A.

(Eds.) (2019). *Cochrane handbook for systematic reviews of interventions version 6.0* (updated July 2019). Cochrane. https://training.cochrane.org/handbook

Holder, E. (2014). Speech presented at the National Association of Criminal Defense Lawyers 57th Annual meeting and 13th State Criminal Justice Network conference, Philadelphia, PA. *Federal Sentencing Reporter, 27*(4), 252–255. http://doi.org/10.1525/fsr.2015.27.4.252

§Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology, 68*(2), 203–211. https://doi.org/10.1037/0021-9010.68.2.203

§Hood, R., Shute, S., Feilzer, M., & Wilcox, A. (2002). Sex offenders emerging from long-term imprisonment: A study of their long-term reconviction rates and of parole board members' judgments of their risk. *British Journal of Criminology, 42*(2), 371–394. https://doi.org/10.1093/bjc/42.2.371

Hossain, M. M. (2020). Umbrella review as an emerging approach of evidence synthesis in health sciences: A bibliometric analysis. *SSRN Electronic Journal.* http://doi.org/10.2139/ssrn.3551055

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hunt, H., Pollock, A., Campbell, P., Estcourt, L., & Brunton, G. (2018). An introduction to overviews of reviews: planning a relevant research question and objective for an overview. *Systematic Reviews, 7*(1), 39. https://doi.org/10.1186/s13643-018-0695-8

Iorio, A., Spencer, F. A., Falavigna, M., Alba, C., Lang, E., Burnand, B., McGinn, T., Hayden,

J., Williams, K., Shea, B., Wolff, R., Kujpers, T., Perel, P., Vandvik, P. O., Glasziou, P., Schunemann, H., & Guyatt, G. (2015). Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *British Medical Journal, 350,* h870. https://doi.org/10.1136/bmj.h870

§Janofsky, J. S., Spears, S., & Neubauer, D. N. (1988). Psychiatrists' accuracy in predicting violent behavior on an inpatient unit. *Hospital & Community Psychiatry, 39*(10), 1090–1094. https://doi.org/10.1176/ps.39.10.1090

§Johansen, S. H. (2006). *Accuracy of predictions of sexual offense recidivism: A comparison of actuarial and clinical methods.* [Doctoral dissertation, Fielding Graduate University]. ProQuest Information and Learning Company.

Kansas v. Hendricks, 521 U.S. 346 (1997).

§Kirk, A. (1989). The prediction of violent behavior during short-term civil commitment. *Bulletin of the American Academy of Psychiatry & the Law, 17*(4), 345–353.

Klassen, D., & O'Connor, W. A. (1989). Assessing the risk of violence in released mental patients: A cross-validation study. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*(2), 75–81. https://doi.org/10.1037/1040-3590.1.2.75

§Kolko, D. J. (2005, November). *Court-mandated juvenile sexual offenders in the community: Characteristics, collaborative treatment, and recidivism*. Talk conducted at the annual conference of the Association for the Treatment of Sexual Abusers, Salt Lake City, UT.

§Kozol, H. L., Boucher, R. J., & Garofalo, R. F. (1972). The diagnosis and treatment of dangerousness. *Crime and Delinquency, 18(4),* 371–392. https://doi.org/10.1177/001112877201800407

§Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior, 24*(1), 101–118. https://doi.org/10.1023/A:1005430904495

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363-374. https://doi.org/10.2307/2529786

§Langton, C. M. (2003). *Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy* (UMI No. NQ78052) [Doctoral dissertation, University of Toronto]. Dissertation Abstracts International.

§Levinson, R. M., & Ramsay, G. (1979). Dangerousness, stress, and mental health evaluations. *Journal of Health and Social Behavior, 20*(2), 178–187. https://doi.org/10.2307/2136438

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of American Medicine, 269*(8), 1007-1011. https://doi.org/10.1001/jama.1993.03500080055032

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances, 6*(7), eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law, 7*(2), 409–443. https://doi.org/10.1037/1076-8971.7.2.409

§Lodewijks, H. P., Doreleijers, T. A., De Ruiter, C., & Borum, R. (2008). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential

treatment. *International Journal of Law and Psychiatry*, *31*(3), 263-271. https://doi.org/10.1016/j.ijlp.2008.04.009

Lorenz, R. C., Matthias, K., Pieper, D., Wegewitz, U., Morche, J., Nocon, M., Rissling, O., Schirm, J., & Jacobs, A. (2019). A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. *Journal of Clinical Epidemiology*, *114*, 133–140. https://doi.org/10.1016/j.jclinepi.2019.05.028

Maddan, S., & Hartley, R. D. (2018). Towards the development of a standardized focal concerns theory of sentencing. In J. T. Ulmer & M. S. Bradley (Eds.), *Handbook on punishment decisions: Locations of disparity* (pp. 311–335). Routledge, Taylor & Francis.

Mauer, M. (2011). Addressing racial disparities in incarceration. *The Prison Journal*, *91*(3, Suppl), 87S–101S. h https://doi.org/10.1177/0032885511415227

McGrath, R. J., Cumming, G. F., Burchard, B. L., Zeoli, S., & Ellerby, L. (2010). *Current practices and emerging trends in sexual abuser management: The Safer Society 2009 North American Survey*. Brandon, VT: Safer Society Press. http://robertmcgrath.us/files/6414/3204/5288/2009_Safer_Society_North_American_Survey.pdf

§McNiel, D. E., & Binder, R. L. (1987). Predictive validity of judgments of dangerousness in emergency civil commitment. *American Journal of Psychiatry, 144*(2), 197–200. https://doi.org/10.1176/ajp.144.2.197

§McNiel, D. E., & Binder, R. L. (1991). Clinical assessment of the risk of violence among psychiatric inpatients. *American Journal of Psychiatry, 148*(10), 1317–1321. https://doi.org/10.1176/ajp.148.10.1317

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of*

the evidence. University of Minnesota Press. https://doi.org/10.1037/11281-000

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (1997). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (2nd ed.). Guilford Press.

Melton, G. B., Petrila, J., Poythress, N. G., Slobogin, C., Otto, R. K., Mossman, D., & Condie, L. O. (2018). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (4th ed.). Guilford Press.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLoS Medicine, 6(7),* 1–6. https://doi.org/10.1371/journal.pmed.1000097

Monahan, J. (1981). *The clinical prediction of violent behavior.* U.S. Department of Health and Human Services.

Monahan, J. (1984). The prediction of violent behavior: Toward a second generation of theory and policy. *American Journal of Psychiatry, 141*(1), 10–15. https://doi.org/10.1176/ajp.141.1.10

Monahan, J., & Cummings, L. (1974). Prediction of dangerousness as a function of its perceived consequences. *Journal of Criminal Justice, 2*(3), 239–242. https://doi.org/10.1016/0047-2352(74)90035-X

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12,* 489–513. http://dx.doi.org/10.1146/annurev-clinpsy-021815-092945

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for

systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moons, K. G., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Johannes B. R., Jos K., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), W1-W33. https://doi.org/10.7326/M18-1377

*Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*(4), 783–792. https://doi.org/10.1037/0022-006X.62.4.783

§Mullen, J. M., & Reinehr, R. C. (1982). Predicting dangerousness of maximum security forensic mental patients. *Journal of Psychiatry and Law*, *10*(2), 223–231. https://doi.org/10.1177/009318538201000207

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, *24*(10), 1889–1897. https://doi.org/10.1177/0956797613481812

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and Caucasian female and male adolescents on probation. *Psychological Assessment*, *32*(6), 594–607. https://doi.org/10.1037/pas0000816

Neal, T. M. S., & Grisso, T. (2014). The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy, and Law*, *20*(2), 200–211. https://doi.org/10.1037/a0035824

Ng, C., & Benedetto, U. (2016). Evidence hierarchy. In G. Biondi-Zoccai (Ed.), *Umbrella*

reviews: Evidence synthesis with overviews of reviews and meta-epidemiologic studies (pp. 11–19). Springer. https://doi.org/10.1007/978-3-319-25655-9_2

*Nicholls, T. L., Pritchard, M. M., Reeves, K. A., & Hilterman, E. (2013). Risk assessment in intimate partner violence: A systematic review of contemporary approaches. *Partner Abuse, 4*(1), 76–168. https://doi.org/10.1891/1946-6560.4.1.76

Northpointe Institute for Public Management. (1996). *COMPAS* [Computer software]. Author.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service Scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*(1), 156–176. https://doi.org/10.1037/a0035080

Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological Reports, 10(3),* 799–812. https://doi.org/10.2466/pr0.1962.10.3.799

§Perez, F. I. (1976). Behavioral analysis of clinical judgment. *Perceptual and Motor Skills, 43*(3), 711-718. https://doi.org/10.2466/pms.1976.43.3.711

§Philipse, M. (2002, March). Postdictive validity of the HCR-20 in a Dutch forensic psychiatric sample. Talk conducted at the annual conference of the International Association of Forensic Mental Health Services, Munich, Germany.

§Phillips, P., & Nasr, S. J. (1983). Seclusion and restraint and prediction of violence. *American Journal of Psychiatry, 140*(2), 229-232. https://doi.org/10.1176/aip.140.2.229

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-

scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour, 4*(7), 736–745. https://doi.org/10.1038/s41562-020-0858-1

Pollock, A., Campbell, P., Brunton, G., Hunt, H., & Estcourt, L. (2017). Selecting and implementing overview methods: implications from five exemplar overviews. *Systematic Reviews, 6*(1), 145. https://doi.org/10.1186/s13643-017-0534-3

§Polvi, N. H. (1999). *The prediction of violence in pre-trial forensic patients: The relative efficacy of statistical versus clinical predictions of dangerousness* [Doctoral dissertation, Simon Fraser University]. Dissertation Abstracts International.

Pretrial Justice Institute. (2009). *Survey of pretrial services programs*. https://university.pretrial.org/viewdocument/survey-of-pretrial-s

Pretrial Justice Institute (2020, Feb). *Updated position on pretrial risk assessment tools.* https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk.* American Psychological Association. https://doi.org/10.1037/10304-000

QSR International Pty Ltd. (2018). *NVivo qualitative data analysis software* (Version 12).

Reddon, J. R., Studer, L., & Estrada, L. (1996). Recidivism data from the Phoenix Program for sex offender treatment [Data set].

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Roberts, J. V., & Reid, A. A. (2017). Aboriginal incarceration in Canada since 1978: Every picture tells the same story. *Canadian Journal of Criminology and Criminal*

Justice, *59*(3), 313-345. doi:10.3138/cjccj.2016.E24

§Rofman, E. S., Askinazi, C., & Fant, E. (1980). The prediction of dangerous behavior in emergency civil commitment. *American Journal of Psychiatry, 137*(9), 1061–1064. https://doi.org/10.1176/ajp.137.9.1061

§Sacks, H. R. (1977). Promises, performance, and principles: An empirical study of parole decision making in Connecticut. *Connecticut Law Review, 9(3),* 349-422.

Salgado, J. F. (2018). Transforming the area under the normal curve (AUC) into Cohen's d, Pearson's rpb, odds-ratio, and natural log odds-ratio: Two conversion tables. *European Journal of Psychology Applied to Legal Context, 10*(1), 35-47. http://doi.org/10.5093/ejpalc2018a5

Schall v. Martin, 467 U.S. 253 (1984).

Schiller, G. (2000). Hospital and aftercare clinicians' estimates of recidivism probability compared with actual reoffense rates by type of outcome and type of prediction [Data set].

§Schram, D. D., Milloy, C. D., & Rowe, W. E. (1991). *Juvenile sex offenders: A follow-up study of reoffense behavior.* Washington State Institute for Public Policy.

§Sepejak, D., Menzies, R. J., Webster, C. D., & Jensen, F. A. (1983). Clinical predictions of dangerousness: Two-year follow-up of 408 pre-trial forensic cases. *Bulletin of the American Academy of Psychiatry and the Law, 11*(2), 171–181.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal*, *358*, j4008.

https://doi.org/10.1136/bmj.j4008

§Shergill, S. S., Murray, R. M., & McGuire, P. K. (1998). Auditory hallucinations: A review of psychological treatments. *Schizophrenia Research, 32*(3), 137–150. https://doi.org/10.1016/S0920-9964(98)00052-8

Shepherd, S. M., & Anthony, T. (2018). Popping the cultural bubble of violence risk assessment tools. *Journal of Forensic Psychiatry & Psychology, 29*(2), 211–220. https://doi.org/10.1080/14789949.2017.1354055

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8–22. https://doi.org/10.1002/bsl.2052

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., Doyle, M., Folino, J. O., Godoy-Cervera, V., Grann, M., Ho, R. M. Y., Large, M. M., Nielsen, L. H., Pham, T. H., Rebocho, M. F., Reeves, K. A., Rettenberger, M., de Ruiter, C., Seewald, K., & Otto, R. K. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*(3), 193–206. https://doi.org/10.1080/14999013.2014.922141

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice and Behavior, 37(9)*, 965–988. https://doi.org /10.1177/0093854810374274

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499–513. https://doi.org/10.1016/j.cpr.2010.11.009

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680–712. https://doi.org/10.1111/1745-9125.12123

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20(1),* 38-42. https://doi.org/10.1177/0963721410397271

Slomen, D. J., Webster, C. D., Butler, B. T., Jensen, F. A. S., Turrall, G. M., Pepper, J., Penfold, M., Sepejak, D. S. Loftus, L., Byers, D., Chapeskie, T., Mahabir, R. J., Schlager, M., Beckett, K., Ronald, M., Shinkoda, A., McDonald, A., Glasberg, R., Jackson, M., Allgood, R., Harman, R., Keeling, K., Taylor, C., Murray, M., Farquharson, D., Lawson, I., Hermanstyne, L., & Bendall, L. (1979). *The assessment of dangerous behaviour: Two new scales. METFORS Working Paper #14.* Metropolitan Toronto Forensic Service.

§Smith, J., & Lanyon, R. I. (1968). Prediction of juvenile probation violators. *Journal of Consulting and Clinical Psychology, 32(1),* 54–58. https://doi.org /10.1037/h0025451

Starr, S. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*(4), 803–872. http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2014/04/66_Stan_L_Rev_803-Starr.pdf

§Steadman, H. J. (1977). A new look at recidivism among Patuxent inmates. *Journal of the American Academy of Psychiatry and the Law, 5(*2), 200–209. http://jaapl.org/content/5/2/200

§Storment, C. T., & Finney, B. C. (1953). Projection and behavior: A Rorschach study of assaultive mental hospital patients. *Journal of Projective Techniques, 17*(3), 349-360. https://doi.org/10.1080/08853126.1953.10380498

§Sturgeon, V., & Taylor, J. (1980). Report of a five-year follow-up study of mentally disordered

sex offenders released from Atascadero State Hospital in 1973. *Criminal Justice Journal, 4*(1), 31–63.

Takwoingi, Y., Riley, R. D., & Deeks, J. J. (2015). Meta-analysis of diagnostic accuracy studies in mental health. *Evidence-Based Mental Health, 18*(4), 103-109. https://doi.org/10.1136/eb-2015-102228

§Thompson, R. E. (1952). A validation of the Glueck social prediction scale for proneness to delinquency. *Journal of Criminal Law & Criminology, 43(4),* 451–470. https://doi.org/10.2307/1139334

*Turgut, T., Lagace, D., Izmir, M., & Dursun, S. (2006). Assessment of violence and aggression in psychiatric settings: Descriptive approaches = Psikiyqtri kliniklerinde şiddet ve agresyonun değerlendirilmesi: Tanisal yaklaşimlar. *Klinik Psikofarmakoloji Bülteni / Bulletin of Clinical Psychopharmacology, 16*(3), 179–194.

United States 115th Congress. (2018). *U.S.757, First Step Act.* https://www.congress.gov/bill/115thcongress/senate-bill/756/text#toc-idf6d6f8724d0d4799b9f45809dca1a3fa

§Van Emmerick, J. (1987). Detention at the government's pleasure: A follow-up study of patients released from the Dr. Henri Van Der Hoeven Clinic. In M. J. M. Brand-Koolen (Ed.), *Studies on the Dutch prison system* (pp. 117-149). Kugler.

Viljoen, J. L., Cochrane, D. M., & Jonnson, M. R. (2018). Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review. *Law and Human Behavior, 42*(3), 181–214. https://doi.org/10.1037/lhb0000280

Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., Vargen, L. M., & Vincent, G. M. (2019). Impact of risk assessment instruments on rates of pretrial detention, postconviction placements,

and release: A systematic review and meta-analysis. *Law and Human Behavior*, *43*(5), 397–420. https://doi.org/10.1037/lhb0000344

Viljoen, J. L., & Vincent, G. M. (2020). Risk assessments for violence and reoffending: Implementation and impact on risk management. *Clinical Psychology: Science and Practice.* Advance online publication. https://doi.org/10.1111/cpsp.12378

Vincent, G. M., & Viljoen, J. L. (2020). Racist algorithms or systemic problems? Risk assessments and racial disparities. *Criminal Justice and Behavior.* Advance online publication. https://doi.org/10.1177/0093854820954501

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

§Werner, P. D., Rose, T. L., Yesavage, J. A., & Seeman, K. (1984). Psychiatrists' judgments of dangerousness in patients on an acute care unit. *American Journal of Psychiatry*, *141*(2), 263–266. https://doi.org/10.1176/ajp.141.2.263

Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies P., Kleijnen, J., Churchill, R., & ROBIS group (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225-234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd edition). Springer.

§Wieand, L. A. (1983). *Judgments of dangerousness: An analysis of mental hospital decisions to release mentally disordered sex offenders* (UMI No. 8323449) [Doctoral dissertation, University of California, Riverside]. Dissertation Abstracts International.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and

applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51-58. https://doi.org/10.7326/M18-1376

§Wormith, J. S., & Goldstone, C. S. (1984). The clinical and statistical prediction of recidivism. *Criminal Justice & Behavior, 11*(1), 3–34. https://doi.org /10.1177/0093854884011001001

§Wormith, J., & Ruhl, M. (1986). Preventive detention in Canada. *Journal of Interpersonal Violence, 1*(4), 399–430. https://doi.org/10.1177/088626086001004002

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. https://doi-org/10.1037/a0020473

§Yesavage, J. A., Werner, P. D., Becker, J. M., & Mills, M. J. (1982). Short-term civil commitment and the violent patient. *American Journal of Psychiatry, 139*(9), 1145–1149. https://doi.org/10.1176/ajp.139.9.1145

§Zeiss, R. A., Tanke, E. D., Fenn, H. H., & Yesavage, J. A. (1996). Dangerousness commitments: Indices of future violence potential? *Bulletin of the American Academy of Psychiatry and Law, 24*(2), 247-253.

Table 1

Characteristics of Meta-Analyses and Systematic Reviews

First Author	Year	Type of Systematic Review	Sample Type	Dates Included	Total Citations	Citations/ Year	# Studies on UCJ	# Studies on Statistical Models	# Studies on Risk Tools	Direct or Indirect Comparison
Ægisdottir	2006	Meta-analysis	Various	1940-1996	695	49.64	6	6	-	Direct
Blank	2001	Narrative	Psychiatric	1990-2000	3	0.16	3	1	-	NA
Bonta	1998	Meta-analysis	MDOs	1959-1995	1632	74.18	7	10	1	Indirect
Grove	2000	Meta-analysis	Various	1945-1994	1872	93.60	10	10	-	Direct
Guy	2008	Meta-analysis	Various	Up to 2008	158	13.17	6	-	104	Indirect
Hanson	2009	Meta-analysis	SO	Up to 2008	1071	97.36	14	-	110	Indirect
Mossman	1994	Meta-analysis	Various	1972-1993	1116	69.75	17	22	-	Indirect
Nicholls	2013	Narrative	IPV	Up to 2011	101	14.43	0	-	36	NA
Turgut	2006	Narrative	Unclear	1996-2006	20	1.43	3	-	-	NA

Note. Direct = analyses compared unstructured and structured judgment in the same samples; Indirect = analyses compared unstructured and structured judgment in different samples; IPV = individuals who perpetrated intimate partner violence; MDOs = offenders with mental disorders; NA = not applicable because no analyses were conducted; SO = individuals who committed sexual offenses; Statistical = model designed for estimating risk for violent, any, or sexual offending that is used for research purposes rather than being a risk assessment tool used in practice; UCJ = unstructured clinical judgment of risk for violent, any, or sexual offending. Total citations were derived from Google Scholar on October 20, 2020. Citations per year were calculated on the same date.

Table 2

AMSTAR 2 Ratings

First Author	Year	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16
Ægisdottir	2006	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No	No	Yes	Yes	No
Blank	2001	No	No	No	No	No	No	No	Partial	No	No	NA	NA	No	No	NA	No
Bonta	1998	Yes	No	Yes	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No
Grove	2000	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No	Yes	Yes	No	No
Guy	2008	Yes	No	No	Partial	No	Yes	No	Partial	No	No	No	No	No	No	Yes	No
Hanson	2009	Yes	No	No	Partial	No	Yes	No	Partial	No	No	Yes	No	No	Yes	Yes	No
Mossman	1994	Yes	No	No	No	No	No	No	No	No	No	Yes	No	No	Yes	No	No
Nicholls	2013	Yes	No	No	No	Yes	No	No	Yes	No	No	NA	NA	No	NA	NA	No
Turgut	2006	No	No	No	No	No	No	No	No	No	No	NA	NA	No	No	NA	No

Note. NA = not applicable (was not a meta-analysis and therefore the item was not relevant). Item 1: Did the research questions and inclusion criteria for the review include the components of PICO? Item 2: Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol? Item 3: Did the review authors explain their selection of the study designs for inclusion in the review? Item 4: Did the review authors use a comprehensive literature search strategy (e.g., provide a justification for including only English studies)? Item 5: Did the review authors perform study selection in duplicate? Item 6: Did the review authors perform data extraction in duplicate? Item 7: Did the review authors provide a list of excluded studies and justify the exclusions? Item 8: Did the review authors describe the included studies in adequate detail? Item 9: Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review? Item 10: Did the review authors report on the sources of funding for the studies included in the review? Item 11: If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results? Item 12: If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis? Item 13: Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review? Item 14: Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review? Item 15: If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review? Item 16: Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

Table 3

ROBIS Ratings

First Author	Year	Risk of Bias in ROBIS Domains				
		Study Eligibility	Selection of Studies	Data Collection & Appraisal	Syntheses of Findings	Overall Risk of Bias
Ægisdottir	2006	Low	Unclear	Unclear	Low	Unclear
Blank	2001	High	High	High	High	High
Bonta	1998	Low	High	High	Unclear	Unclear
Grove	2000	Low	High	Low	Unclear	Unclear
Guy	2008	Low	Low	Unclear	Low	Low
Hanson	2009	Low	Low	Unclear	Low	Low
Mossman	1994	High	High	High	High	High
Nicholls	2013	High	High	Unclear	NA	Unclear
Turgut	2006	High	High	Unclear	Unclear	High

Note. Low = low risk of bias; Unclear = unclear risk of bias; High = high risk of bias; NA = not applicable (no studies identified).

Table 4

Results of Prior Meta-Analyses

First Author	Year	Offense Type	Unstructured Judgment		Structured Judgment (i.e., Statistical Model or Risk Assessment Tool)			Was the Model or Tool Significantly Better Than Unstructured Judgment?	
			k	AUC	Type	*k*	AUC	Significant?	Statistics Reported
Statistical Model (*k* = 4)									
Ægottisdotir	2006	Any	4	NR	Statistical model	4	NR	Yes	$d+ = -.17$
Bonta	1998	Violence	3	.55	Statistical model	8	.65 [a]	Yes	$\chi^2 = 21.98$
		Any	5	.56 [b]	Statistical model	6	.71	Yes	$\chi^2 = 77.25$
Grove	2000	Any	10	NR	Statistical model	10	NR	Unclear	$M_{diff} = 0.89$
Mossman	1994	Violence	17	.67	Cross-validated model	14	.71	Yes	$z = 2.88$
					Non-cross-validated	14	.89	Yes	CIs did not overlap
Risk Assessment Tools (*k* = 2)									
Guy	2008	Any	6	.58	SPJ	104	.68	Yes	CIs did not overlap
					Actuarial	45	.67	Unclear [c]	CIs – overlap unclear
Hanson [d]	2009	Violence	7	.56	Actuarial for violence	15	.71	Yes	CIs did not overlap
					Mechanical for violence	3	.59	No	CIs overlapped
		Any	9	.53	Actuarial for any	10	.75	Yes	CIs did not overlap
		Sexual	11	.62	Actuarial for sexual	81	.68	Yes	CIs did not overlap
					Mechanical for sexual	29	.68	Yes	CIs did not overlap
					SPJ for sexual	6	.63	No	CIs overlapped

Note. AUC = aggregated area under the receiver operating characteristics curve; CI = confidence intervals; *d*+ = mean difference in deviation units corrected for sample size; *k* = number of studies; M_{diff} = mean difference in the transformed effect size; NR = not effect size was reported; SPJ = structured professional judgment tool. To aggregate effect sizes, Guy (2008) used a random effects model, Hanson and Morton-Bourgon (2009) used a fixed effects model, and Mossman (1994) averaged AUC scores. [a] After outliers were excluded, the AUC was .67. [b] After outliers were excluded, the AUC was .52. [c] CIs were .65-.70 for actuarial tools and .50-.65 for unstructured judgment so it was unclear if they overlapped as numbers may have been rounded up or down. [d] Hanson and Morton-Bourgon (2009) define mechanical tools as tools which use explicit a priori methods for combining items into total scores; in contrast, actuarial tools also include tables to estimate expected recidivism rates.

Table 5

Characteristics of Primary Studies on Unstructured Judgment Included in Prior Meta-Analyses and Systematic Reviews

First Author	Year	Sample Size	Population	Age	Sex	Explicitly Rated Risk	Unstructured Clinical Judgment	Research or Practice	Structured Judgment	Research or Practice	Offense Type	Source	Follow-Up	Included In
Direct Comparison with Risk Assessment Tools (*k* = 12)														
Bengtson	2007	121	SO	Adult	M	No	Coded pretrial reports	Practice	Static-99 & Static-2002	Research	V, S	J	~16 yr	H
De Vogel	2004	120	Justice	Adult	M/F	No	Discharge decision	Practice	HCR-20	Research	V, A	J	~6 yr	Gu
Enebrink	2006	76	MH	Child	M	Yes	Risk rating (10-pt. scale)	Research	EARL-20B	Research	V	C	~2.5 yr	Gu
Holland	1983	339	Justice	Adult	M	Legal	Jail recommendation	Practice	Salient Factor Scale	Research	V, A	J	2 yr	Æ, G
Hood	2002	162	SO	Adult	M	No	Observed parole board meetings	Practice	Static-99	Research	V, A, S	J	2-6 yr	H
Johansen	2006	280	SO	Adult	NR	Yes	Risk rating (3 & 5-pt. scale)	Practice	RRASOR, Static-99, VRAG, etc.	Research	V, A, S	J	7 yr	H
Kropp	2000	102	IPV	Adult	M	Yes	Risk rating (10-pt. scale)	NR	SARA	Research	V	J	NR	Gu
Langton	2003	476	SO	Adult	M	Yes	Risk rating (5-pt. scale)	Practice	RRASOR, Static-99, VRAG, etc.	Research	V, A, S	J	~6 yr	H
Lodewijks	2008	117	Justice	Adol.	M/F	No	Coded forensic reports	Practice	SAVRY	Research	V	J	3 yr	Gu
Philipse	2002	69	Justice	Adult	M/F	Yes	Risk rating (6-pt. scale)	Practice	HCR-20	Research	A	J	~4 yr	Gu
Polvi	1999	215	Justice	Adult	M/F	Yes	Risk rating (7-pt. scale)	Practice	HCR-20, VRAG, DBRS	Research	V	J	~6 yr	Gu
Wormith	1984	222	Justice	Adult	M	Legal	Release decision	Practice	Recidivism Prediction Score	Research	A	J	NR	G
Direct Comparison with Statistical Models (*k* = 9)														
Gardner	1996	784	MH	Adult /adol.	M/F	Yes	Risk rating (5-pt. scale)	Practice	6-item model (e.g., past violence)	Research	V	J, C	6 mo.	Æ, Bl, T
Glaser	1955	2545	Justice	NR	NR	No	General prognosis	Practice	7-item model (e.g., criminal record)	Research	Vn	NR	NR	G
Glaser	1958	190	Justice	NR	M	No	Coded presentence reports	Practice	6-item model (e.g., prior conviction)	Research	Vn	NR	NR	G
Hall	1988	342	SO	Adult	M	Legal	Safety determination	Practice	19-item model (e.g., age, MMPI)	Research	V, S	J	5 yr	Æ, G
Perez	1976	40	Justice	Adult	M	No	Sorted based on MMPI & Rorschach	Research	Multivariate model (Rorschach, MMPI)	Research	V	J	NR	Æ
Sacks	1977	226	Justice	Adult	M	Legal	Release decision	Practice	4-item model (e.g., criminal records)	Research	A	J	1 yr	G
Smith	1968	287	Justice	Adol.	M	No	Sorted based on MMPI	Research	5-item model (e.g., court referrals, age)	Research	A	J	1 yr.	G
Thompson	1952	100	MH	Child	M	Yes	Risk rating (11-pt. scale)	Research	Glueck Social Prediction Scale [a]	Research	A	J	NR	Æ, G

Werner	1984	40	MH	NR	M	No	Classified based on BPRS	Research	19-item model (BPRS, assault)	Research	V	H	7 days	Æ, G, M
Studies with No Structured Judgment (*k* = 25)														
Bloom	1986	123	Justice	Adult	M/F	Legal	Discharge decision	Practice	NA	NA	A	J	2-4 yr	B
Cocozza	1976	257	Justice	Adult	M	Legal	Testimony on dangerousness	Practice	NA	NA	V, A	J	3 yr	B, M
Dix	1976	130	SO	NR	NR	Legal	Coded reports	Practice	NA	NA	A, S	J	7 yr	H
Florida DHRS	1985	129	SO	Adult	M	No	Prognosis	Practice	NA	NA	A, S	J	>0.75 yrs	H
Janofsky	1988	47	MH	Adult	M/F	Yes	Prediction of assault	Practice	NA	NA	V	H	7 days	M
Kirk	1989	68	MH	Adult	M/F	Legal	Commitment for danger to others	Practice	NA	NA	V	H	~3 days	M
Kolko	2005	171	SO	Adol.	M	Yes	Identified as high risk	Practice	NA	NA	A	J	2 yrs	H
Kozol	1972	435	SO	Adult	M	Legal	Release decision	Practice	NA	NA	V	NR	NR	H, M
Levinson	1979	53	MH	Adult	M/F	No	Home visit notes	Practice	NA	NA	V	S	NR	M
McNiel	1987	101	MH	Adult	M/F	Legal	Commitment for danger to others	Practice	NA	NA	V	H	3 days	M
McNiel	1991	149	MH	Adult	M/F	Yes	Risk rating (11-pt. scale)	Practice	NA	NA	V	H	7 days	M
Mullen	1982	165	Justice	Adult	M	Yes	Staff consensus if dangerous	Research	NA	NA	A	J	4 yrs	M
Phillips	1983	69	MH	Adult	M/F	Legal	Commitment for danger to others	Practice	NA	NA	V	H	NR	M
Rofman	1980	118	MH	Adult	M/F	Legal	Commitment for danger to others	Practice	NA	NA	V	H	45 days	M
Schram	1991	197	SO	Adol.	M	Yes	Risk rating (3-pt. scale)	Practice	NA	NA	A, S	J	~6 yr	H
Sepejak	1983	408	Justice	Adult/ adol.	M/F	Yes	Risk rating (4-pt. scale)	Practice	NA	NA	A	J	2 yrs.	B
Shergill	1998	318	MH	NR	M/F	Yes	Risk rating (3-pt. scale)	Practice	NA	NA	V	NR	6 mo.	Bl
Steadman	1977	85	Justice	Adult	NR	Legal	Defective delinquent determination	Practice	NA	NA	V, A	J	3 yrs.	B, M
Storment	1953	46	MH	Adult	M	No	Sorted based on Rorschach	Research	NA	NA	V	H	NR	G
Sturgeon	1980	260	SO	Adult	M/F	Legal	Release decision	Practice	NA	NA	V, A, S	J	5 yrs.	B, H
Van Emmerick	1987	589	Justice	Adult/ adol.	M	Legal	Discharge decision	Practice	NA	NA	V, A	J	>5 yrs.	B
Wieand	1983	182	SO	Adult	M	Legal	Release decision	Practice	NA	NA	A	J	~2 yr.	H
Wormith	1986	75	SO	Adult	M	No	General prognosis	Practice	NA	NA	A	J	>8 yrs.	H
Yesavage	1982	84	MH	Adult	NR	Legal	Commitment for danger to others	Practice	NA	NA	V	H	7 days	M
Zeiss	1996	62	MH	Adult	M/F	Legal	Commitment for danger to others	Practice	NA	NA	V	J, H	1-5 yrs.	M

Note. NA = not applicable (did not examine); NR = not reported or unclear. [a] Statistical model had been previously cross-validated. **Population:** IPV = people who committed intimate partner violence; Justice = people in justice settings (e.g., prison, probation, forensic psychiatric hospital); MH = people in a mental health setting (e.g., psychiatric hospital); SO = people who sexually offended. **Age:** Child = mean age was less 13 years old; Adol. = mean age was 13 to 18 years old; Adult = mean age was greater than 18 years old. **Sex:** M = males; M/F = males and females. **Explicitly Rated Risk:** risk = professionals made an explicit rating of risk (e.g., rated risk on a 5-pt. scale); legal judgment = risk inferred based on a legal judgment (e.g., decision to detain); no = did not explicitly rate risk or examine legal judgment but, instead, used another approach to infer risk. **Structured Judgment:** BPRS = Brief Psychiatric Rating Scale (Overall & Gorham, 1962); DBRS = Dangerous Behavior Rating Scale (Slomen et al., 1979); EARL-20B = Early Assessment Risk List for Boys (Augimeri et al., 2001); HCR-20 = Historical, Clinical, Risk Management-20 (Webster et al., 1997); MMPI = Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1942); RRASOR = Rapid Risk Assessment of Sex Offender Recidivism (Hanson, 1997); SAVRY = Structured Assessment of Violence Risk (Borum et al., 2006); Static-99 (Hanson & Thornton, 1999); Static-2002 (Hanson & Thornton, 2003); VRAG = Violence Risk Appraisal Guide (Harris et al., 1993). **Offense Type:** A = any or general offense (e.g., nonviolent offending, parole failure); S = sexual offense; V = violence. **Source:** C = collaterals (e.g., friends); H = hospital records or staff observations; J = justice records; S = self-report. **Included In:** Æ = Ægottisdotir et al. (2006); Bl = Blank (2001); B = Bonta et al. (1998); G = Grove et al. (2000); Gu = Guy (2008); H = Hanson & Morton-Bourgon (2009); M = Mossman (1994); T = Turgut.

Figure 1

Search Strategy

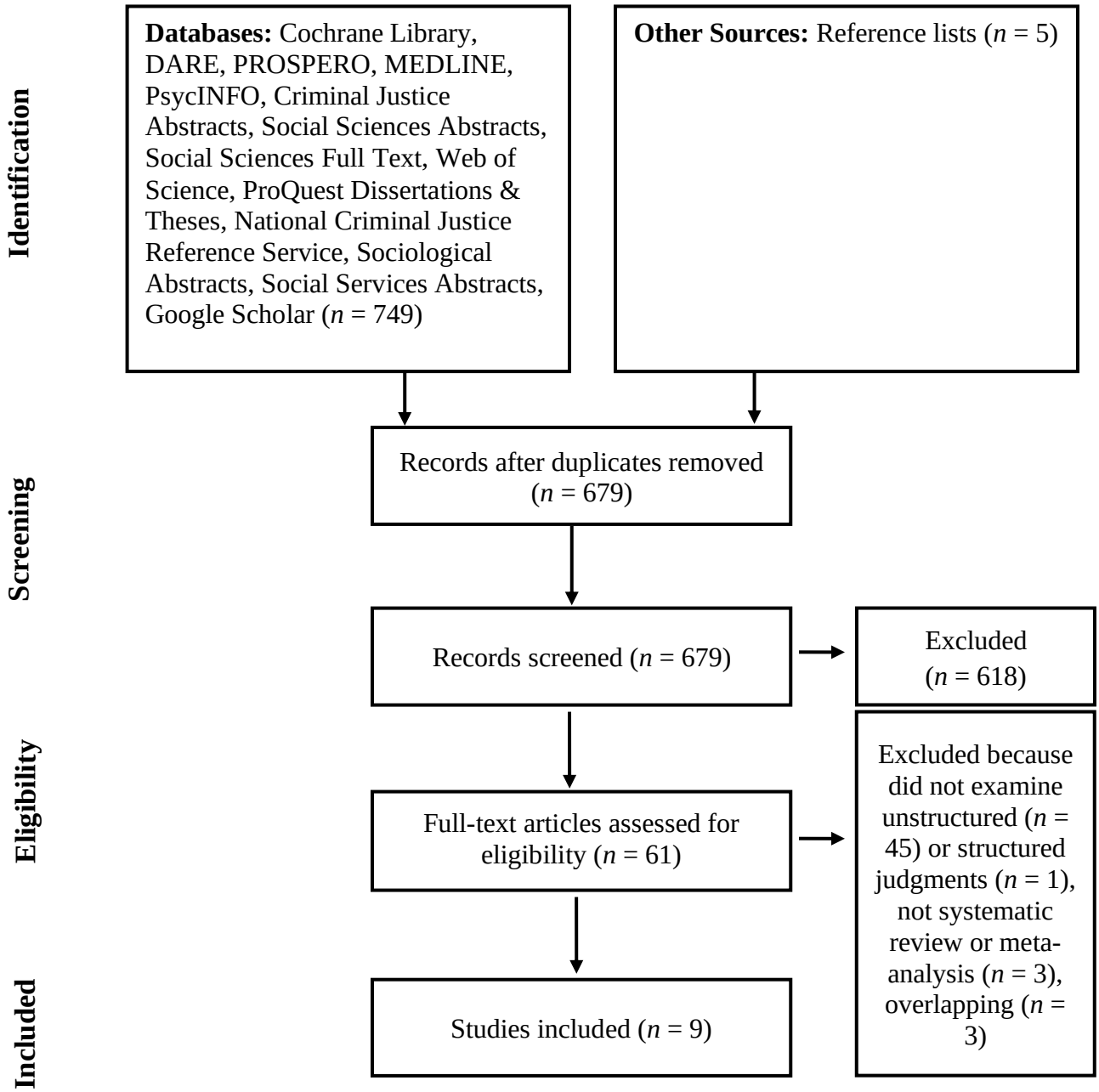
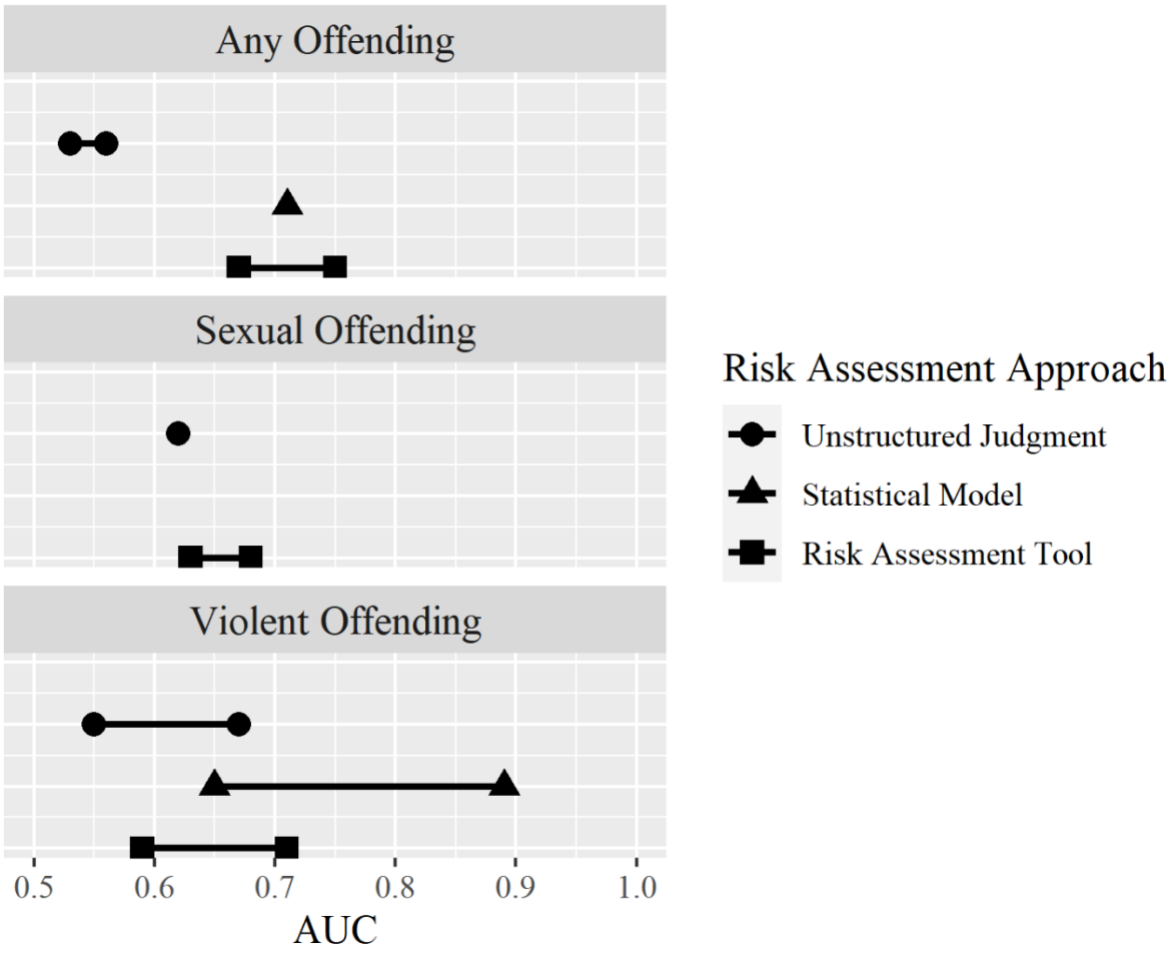


Figure 2

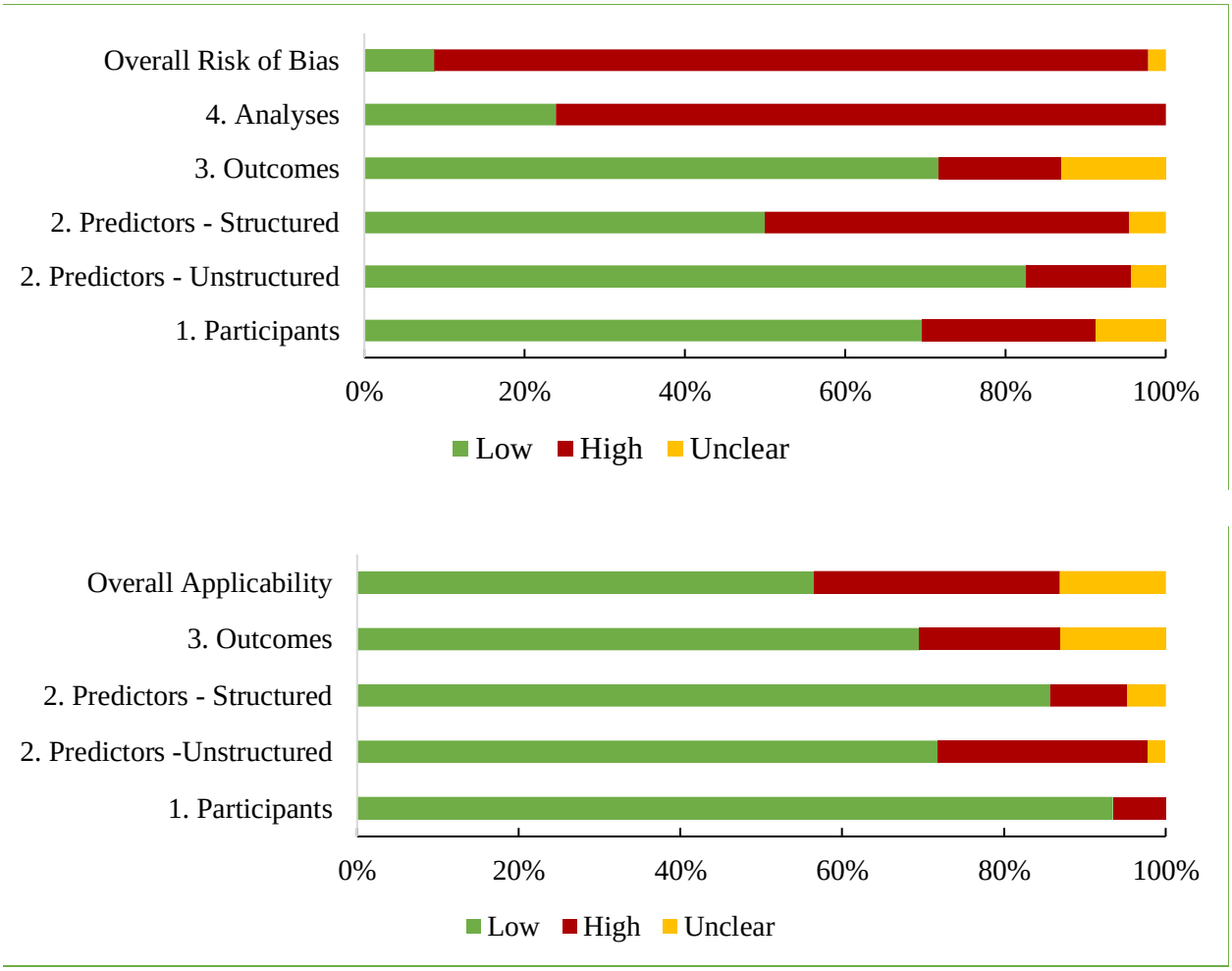
Aggregated AUC Scores from Meta-Analyses



Note. Each square, triangle, and circle represent the lower and upper range of aggregated AUCs from the four meta-analyses that provided effect sizes for unstructured and structured risk judgments (i.e., Bonta et al., 1998; Guy, 2008; Hanson & Morton-Bourgon, 2009; Mossman, 1994). The x-axis ranges from AUC scores of 0.50 (which represents chance) to 1.00 (perfect predictive validity). The numbers shown in the range are derived from Table 4.

Figure 3

PROBAST Ratings for Primary Studies



Supplemental Materials

List of Excluded Studies

Article Citation	Reason for Exclusion A = Not a systematic review or meta-analysis B = Did not examine UCJ C = Did not examine structured risk assessment O = Overlap
1. Babchishin, K. M., Hanson, R. K., & Helmus, L. M. (2012). Even highly correlated measures can add incrementally to predicting recidivism among sex offenders. *Assessment, 19*(4), 442-461. https://doi.org/10.1177/1073191112458312	B
2. Bechtel, K., Lowenkamp, C. T., & Holsinger, A. (2011). Identifying the predictors of pretrial failure: A meta-analysis. *Federal Probation, 75*(2), 78-87.	B
3. Blacker, J. E. (2009). *The assessment of risk in intellectually disabled sexual offenders* [Unpublished doctoral dissertation]. University of Birmingham.	B
4. 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*(6), 567-590 https://doi.org/10.1177/0093854809333610	B
5. Castelletti, L., Rivellini, G., & Straticò, E. (2014). Efficacia predittiva degli strumenti di Violence Risk Assessment e possibili ambiti applicativi nella psichiatria forense e generale italiana. Una revisione della letteratura [Predictive efficacy of violence risk assessment tools, implications for forensic and general psychiatry in Italy: A literature review]. *Journal of Psychopathology*, *20*(2), 153-162.	A
6. Chevalier, C. S. (2017). *The association between structured professional judgment measure total scores and summary risk ratings: Implications for predictive validity* [Unpublished doctoral dissertation]. Sam Houston State University.	B
7. Cottle, C. C., Lee. R. J., & Heilbrun, K. (2001). The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior*, *29*(3), 367-394. https://doi.org/10.1177%2F0093854801028003005	B
8. Dickens, G. L., & O'Shea, L. E. (2018). Protective factors in risk assessment schemes for adolescents in mental health and criminal justice populations: A systematic review and meta-analysis of their predictive efficacy. *Adolescent Research Review*, *3*, 95-112. https://doi.org/10.1007/s40894-017-0062-3	B
9. Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *BMJ*, 1-12. https://doi.org/10.1136/bmj.e4692	B
10. Forrester, Y. (2005). *The quality of expert judgment: An interdisciplinary investigation* [Unpublished doctoral dissertation]. University of Maryland.	B

11. Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, *34*(4), 575-607. https://doi.org/10.1111/j.1745-9125.1996.tb01220.x	B
12. Geraghty, K. A., & Woodhams, J. (2015). The predictive validity of risk assessment tools for female offenders: A systematic review. *Aggression and Violent Behavior*, 21, 25-38. https://doi.org/10.1016/j.avb.2015.01.002	A, O
13. Geraghty, K. A. (2015). *Assessing risk in female offenders* [Unpublished doctoral dissertation]. University of Birmingham.	A
14. Green, L. (2006). *Gender influences and methodological considerations in adolescent risk-need assessment: A meta-analysis* [Unpublished master's thesis]. University of New Brunswick.	B
15. Guo, B., & Harstall. C. (2008). *Risk assessment tools for predicting recidivism of spousal violence.* Institute of Health Economics.	B
16. Hawes, S. W., Boccaccini, M. T., & Murrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist – Revised. *Psychological Assessment*, 25(1), 233-243. https://doi.org/10.1037/a0030391	B
17. Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior*, 39(9), 1148-1171. https://doi.org/10.1177/0093854812443648	B
18. Hogan, N. R., & Olver, M. E. (2018). A prospective examination of the predictive validity of five structured instruments for inpatient violence in a secure forensic hospital. *International Journal of Forensic Mental Health*, 17(2), 122-132. doi: 10.1080/14999013.2018.1431339	B
19. Horcajo-Gil, P. J., Dujo-López, V., Andreu-Rodríguez, J. M., & Marín-Rullán, M. (2019). Valoración y Gestión del Riesgo de Reincidencia Delictiva en Menores Infractores: una Revisión de Instrumentos [Assessment and management of the risk of criminal recidivism in juvenile offenders: A review of instruments]. *Anuario de Psicología Jurídica*, *29*(1), 41-53. https://doi.org/10.5093/apj2018a15	B
20. Judge, J. (2012). *The clinical practice of risk assessment of sexual violence* [Unpublished doctoral dissertation]. University of Edinburgh, Edinburgh.	B
21. Judges, R. C. (2016). *An exploration into the value of protective factors in violence risk assessment of psychiatric inpatients* [Unpublished doctoral dissertation]. University of Nottingham.	B
22. Kingston, D. A., Yates, P. A., Firestone, P., Babchishin, K., & Bradford, J. M. (2008). Long-term predictive validity of the Risk Matrix 2000: A comparison with the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research and Treatment*, *20*(4), 466-484. https://doi.org/10.1177/1079063208325206	B
23. Lam, H. P. (2014). *A meta-analysis of the prediction of violence among adults with mental disorders* [Unpublished doctoral dissertation]. City University of New York.	B
24. Leven, L. (2019). *Violence risk assessment through a gendered lens: Is there a need to develop gender-specific risk assessment tools?* (Unpublished Masters thesis). Malmö University, Malmö, Sweden.	B
25. Lofthouse, R. (2016). *Assessing and managing risk with adults with intellectual disabilities (ID)* [Unpublished doctoral dissertation]. University of Liverpool.	B

26. Lofthouse, R., Golding, L., Totsika, V., Hastings, R., & Lindsay, W. (2017). How effective are risk assessments/measures for predicting future aggressive behaviour in adults with intellectual disabilities (ID): A systematic review and meta-analysis. *Clinical Psychology Review, 58*, 76-85. https://doi.org/10.1016/j.cpr.2017.10.001	B, O
27. Neil, C. (2015). *Assessment of protective factors for violence risk* [Unpublished doctoral dissertation]. University of Edinburgh.	B
28. Nguyen, K. D. (2018). *Evaluating aleatory uncertainty assessment* [Unpublished doctoral dissertation]. University of Southern California.	B
29. Nunn, L. K. (2018). *An investigation into risk assessment and staff coping with patient perpetrated violence in inpatient forensic psychiatric settings* [Unpublished doctoral dissertation]. University of Edinburgh.	B
30. O'Shea, L. E., & Dickens, G. L. (2014). Short-Term Assessment of Risk and Treatability (START): Systematic review and meta-analysis. *Psychological Assessment, 26*(3), 990-1002. https://doi.org/10.1037/a0036794	B
31. O'Shea, L. E., & Dickens, G. L. (2016). Performance of protective factors assessment in risk prediction for adults: Systematic review and meta-analysis. *Clinical Psychology: Science and Practice, 23*(2), 126-138. https://doi.org/10.1111/cpsp.12146	B
32. O'Shea, L. E., Mitchell, A. E., Picchioni, M. M., & Dickens, G. L. (2013). 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. https://doi.org/10.1016/j.avb.2012.11.016	B
33. Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior, 36*(4), 329-353. https://doi.org/10.1177/0093854809331457	B
34. Ramesh, T., Igoumenou, A., Montes, M. V., & Fazel, S. (2018). Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: a systematic review and meta-analysis. *European Psychiatry, 52*, 47-53. https://doi.org/10.1016/j.eurpsy.2018.02.007	B
35. Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31*(5), 449-462. https://doi.org/10.1007/s10979-006-9071-7	B
36. Schwalbe, C. S. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice and Behavior, 35*(11), 1367-1381. https://doi.org/10.1177/0093854808324377	B
37. Singh, J. P. (2011). *Forensic risk assessment: A metareview, novel meta-analysis, and empirical study developing a violence screening tool for schizophrenia* [Unpublished doctoral dissertation]. University of Oxford.	A, O
38. Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice and Behavior, 37*(9), 965-988. https://doi.org/10.1177/0093854810374274	A
39. Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499-513. https://doi.org/10.1016/j.cpr.2010.11.009	B
40. Singh, J. P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other psychiatric disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin, 37*(5), 899-912. https://doi.org/10.1093/schbul/sbr093	B
41. Singh, J. P., Grann, M., Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS ONE*, 8(9), e:72848. https://doi.org/10.1371/journal.pone.0072484	B

42. Tully, R. J., Chou, S., & Browne, K. D. (2013). A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review, 33*(2), 287-316. https://doi.org/10.1016/j.cpr.2012.12.002	B
43. van den Berg, J. W., Smid, W., Schepers, K., Wever, E., van Beek, D., Janssen, E., & Gijs, L. (2018). The predictive properties of dynamic sex offender risk assessment instruments: A meta-analysis. *Psychological Assessment, 30*(2), 179-191. https://doi.org/10.1037/pas0000454	B
44. Walter, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice and Behavior, 33*(3), 279-304. https://doi.org/10.1177/0093854805284409	B
45. Whittington, R., Hockenbull, J. C., McGuire, J., Leitner, M., Barr, W., Cherry, M. G.,…Dickson, R. (2013). A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002–8. *Health Technology Assessment, 17*(50), 1-128. https://doi.org/10.3310/hta17500	B
46. Witt, K., van Dorn, R., & Fazel, S. (2013). Risk factors for violence in psychosis: Systematic review and meta-regression analysis of 110 studies. *PloS ONE, 8*(2), e:55942, https://doi.org/10.1371/journal.pone.0055942	B
47. Rossdale, S-V., Tully, R. J., & Egan, V. (2020). The HCR-20 for predicting violence in adult females: A meta-analysis. *Journey of Forensic Psychology Research and Practice, 20*(1), 15-52. https://doi.org/10.1080/24732850.2019.1681875	B
48. van der Put, C. E., Gubbels, J., & Assink, M. (2019). Predicting domestic violence: A meta-analysis on the predictive validity of risk assessment tools. *Aggression and Violent Behavior, 47*, 100-116, https://doi.org/10.1016/j.avb.2019.03.008	B
49. Viljoen, J. L., Cochrane, D. M., & Jonnson, M. R. (2018). Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review. *Law and Human Behavior, 42*(3), 181-214. https://doi.org/10.1037/lhb0000280	B
50. Wheller, L., & Wire, J. (2014). *Domestic abuse risk factors and risk assessment: Summary of findings from a rapid evidence assessment*. College of Policing. https://www.college.police.uk/News/College-news/Documents/DA_ROR_Summary_14-12-15.doc	B
51. Wand, T. (2011). Investigating the evidence for the effectiveness of risk assessment in mental health care. *Issues in Mental Health Nursing, 33*(1), 2-7. https://doi.org/10.3109/01612840.2011.616984	C
52. Welsh, J. L., Schmidt, F., McKinnon, L., Chattha, H. K., & Meyers, J. R. (2008). A comparative study of adolescent risk assessment instruments: Predictive and incremental validity. *Assessment, 15*(1), 104-115. https://doi.org/10.1177/1073191107307966	A

Risk of Bias of Primary Studies Included in Prior Meta-Analyses and Systematic Reviews

		Risk of Bias					Applicability Problems				Overall	
	First Author	Participants	Predictors – Unstructured	Predictors – Structured	Outcomes	Analyses	Participants	Predictors – Unstructured	Predictors – Structured	Outcomes	ROB	Applicability
1.	Bengston	Low	High	Low	Low	Low	Low	Low	Low	Low	High	Low
2.	Bloom	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
3.	Cocozza	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
4.	De Vogel	Low	High	Low	Low	Low	Low	Low	Low	Low	High	Low
5.	Dix	Low	Low	-	Unclear	High	Low	Unclear	-	Low	High	Unclear
6.	Enebrink	Low	Low	Low	Low	Low	Low	Low	Low	Unclear	Low	Unclear
7.	Florida	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
8.	Gardner	Low	Low	High	High	Low	Low	Low	Low	Low	High	Low
9.	Glaser [a]	High	Low	High	Unclear	High	High	High	Low	Unclear	High	High
10.	Glaser [b]	Low	Low	High	Unclear	High	Low	High	Low	High	High	High
11.	Hall	Unclear	Low	High	Low	High	Low	High	Low	Low	High	High
12.	Holland	Low	Low	Unclear	Low	Low	Low	Low	Low	Low	Unclear	Low
13.	Hood	Low	High	Low	Low	High	Low	High	Low	Low	High	High
14.	Janofsky	Low	High	-	Low	High	Low	Low	-	Low	High	Low
15.	Johansen	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
16.	Kirk	Low	Low	-	Unclear	High	Low	Low	-	Low	High	Low
17.	Kolko	Unclear	Unclear	-	Low	High	Low	Low	-	Low	High	Low
18.	Kozol	High	Low	-	Unclear	High	Low	Low	-	High	High	High
19.	Kropp	Low	Low	Low	Low	High	Low	Low	Low	Low	High	Low
20.	Langton	Low	Low	High	Low	Low	Low	Low	Low	Low	High	Low
21.	Levinson	Unclear	Low	-	High	High	High	Low	-	High	High	High
22.	Lodewijks	Low	High	Low	Low	Low	Low	Low	Low	Low	High	Low
23.	McNeil [c]	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
24.	McNeil [d]	Low	Low	-	High	High	Low	Low	-	Low	High	Low
25.	Mullen	High	Low	-	Unclear	High	Low	Low	-	Unclear	High	Unclear
26.	Perez	High	Low	High	Low	High	Low	High	Unclear	High	High	High
27.	Philipse	High	Low	Low	Low	High	Low	Low	Low	Low	High	Low
28.	Phillips	Low	Low	-	High	High	Low	High	-	High	High	Unclear
29.	Polvi	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
30.	Rofman	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
31.	Sacks	High	Low	High	High	High	Low	High	High	Unclear	High	High
32.	Schram	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
33.	Sepejak	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
34.	Shergill	High	Unclear	-	High	High	Low	Low	-	High	High	High

35. Smith	Low	Low	High	Low	High	Low	High	Low	High	High	High
36. Steadman	High	Low	-	Low	High	Low	High	-	Low	High	High
37. Storment	High	Low	-	High	High	High	High	-	High	High	High
38. Sturgeon	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
39. Thompson	High	Low	Low	Low	High	Low	Low	Low	Unclear	High	Unclear
40. Van Emmerick	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
41. Werner	Unclear	Low	High	Low	High	Low	High	High	Low	High	High
42. Wieand	Low	Low	High	Low	High	Low	Low	-	Unclear	High	Unclear
43. Wormith [e]	Low	Low	Low	Low	High	Low	Low	Low	Low	High	Low
44. Wormith [f]	Low	High	-	Low	Low	Low	High	-	Low	High	High
45. Yesavage	Low	Low	-	Low	High	Low	Low	-	Low	High	Low
46. Zeiss	Low	Low	-	Low	Low	Low	Low	-	Low	Low	Low

Note. All ratings were made using the PROBAST (Prediction model Risk of Bias ASsessment Tool; Wolff et al., 2019). Low = low risk of bias or low concern regarding applicability; High = high risk of bias or high concern regarding applicability; Unclear = unclear risk of bias or unclear concern about applicability. [a] Glaser (1955); [b] Glaser & Hangren (1958); [c] McNeil & Binder (1987); [d] McNeil & Binder (1991); [e] Wormith & Goldstone (1984); [d] Wormith & Ruhl (1986).