Psychological Bulletin
2010, Vol. 136, No. 5, 740–767

0033-2909/10/$12.00 DOI: 10.1037/a0020473

The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools

Min Yang and Stephen C. P. Wong
University of Nottingham

Jeremy Coid
University of London

Actuarial risk assessment tools are used extensively to predict future violence, but previous studies comparing their predictive accuracies have produced inconsistent findings as a result of various methodological issues. We conducted meta-analyses of the effect sizes of 9 commonly used risk assessment tools and their subscales to compare their predictive efficacies for violence. The effect sizes were extracted from 28 original reports published between 1999 and 2008, which assessed the predictive accuracy of more than one tool. We used a within-subject design to improve statistical power and multilevel regression models to disentangle random effects of variation between studies and tools and to adjust for study features. All 9 tools and their subscales predicted violence at about the same moderate level of predictive efficacy with the exception of Psychopathy Checklist—Revised (PCL-R) Factor 1, which predicted violence only at chance level among men. Approximately 25% of the total variance was due to differences between tools, whereas approximately 85% of heterogeneity between studies was explained by methodological features (age, length of follow-up, different types of violent outcome, sex, and sex-related interactions). Sex-differentiated efficacy was found for a small number of the tools. If the intention is only to predict future violence, then the 9 tools are essentially interchangeable; the selection of which tool to use in practice should depend on what other functions the tool can perform rather than on its efficacy in predicting violence. The moderate level of predictive accuracy of these tools suggests that they should not be used solely for some criminal justice decision making that requires a very high level of accuracy such as preventive detention.

Keywords: risk assessment, violent outcome, meta-analysis, multilevel models

Violence and its control are significant social, political, criminal justice, mental health, and international security issues. It is a major public health issue as well, affecting perpetrators, victims, and witnesses, and influencing the general population through fear of crime. Violence has been identified as one of many hazards that should be minimized through risk assessment and appropriate management; some have argued that risk is to be avoided at all cost (Adams, 1995). The prediction of future violence has been one of the most complex and controversial issues in the behavioral sciences (Borum, 1996; Grisso & Appelbaum, 1993; Litwack, 1993; Poythress, 1992). Courts have increasingly relied on mental health professionals for assistance in civil and criminal cases to assess dangerousness or risk of future violence. The premium placed on prediction is evidenced by policy changes that reflect the growth of a culture emphasizing risk aversion, with the increasing implementation of policies, such as zero tolerance, hard targeting, surveillance, selective incapacitation (Haapanen, 1990), long-term incarceration (Kemshall, 2003; Kemshall & Maguire, 2001), and so forth.

The past 20 years have witnessed the development of specialized tools for the prediction and management of violence for use with a variety of populations (Heilbrun et al., 2009). The increasingly severe sanctions for those identified as high risk for violence together with dire career consequences for professional who made erroneous clinical judgments (Maden, 2007) have attracted extremely close scrutiny on the accuracy of risk prediction from both research and policy perspectives. Answers to the question of which risk assessment instrument should be applied to whom and under what circumstances have major implications for routine clinical practice, criminal justice work, teaching and training, and the commercial development of new instruments. The consequences of inaccurate predictions raise a host of legal and ethical issues as well. The identification of the most accurate violence prediction tool or tools therefore deserves the highest priority.

Violent Individuals and Violent Situations

Evidence exists that a disproportionate amount of violent crime is committed by the most persistent adult male offenders, who account for a relatively small proportion of the offender population. For example, it is estimated that about 50% of all crimes are committed by 5%–6% of the offender population (see Farrington, Ohlin, & Wilson, 1986, for a review). However, even violence-

Min Yang, School of Community Health Sciences, University of Nottingham, Nottingham, United Kingdom; Stephen C. P. Wong, Institute of Mental Health, University of Nottingham; Jeremy Coid, Queen Mary's School of Medicine and Dentistry, University of London, London, United Kingdom.

The support of the National Institute for Health Research (NIHR) United Kingdom Grant Programme (PP-PG-6407-10500) to the third author is acknowledged.

Correspondence concerning this article should be addressed to Min Yang, School of Community Health Sciences, Room B21 Sir Colin Campbell Building, Jubilee Campus, Nottingham, United Kingdom, NG7 2TU. E-mail: min.yang@nottingham.ac.uk

prone individuals are not always violent; they commit violence only under certain conditions. For example, the likelihood of violence for a spouse abuser increases when the individual is in contact with a partner (Dearwater et al., 1998) or, for a pedophile, when given access to children (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2004). Even in seemingly random violent acts, such as school shootings, retrospective investigation reveals the perpetrator to have acted only under exceptional personal circumstances (FBI Academy, National Center for the Analysis of Violent Crime, Critical Incident Response Group, n.d.). Thus, by identifying a relatively small number of individuals, understanding the cause of their violence, and effectively managing these individuals, it is theoretically possible to reduce the incidence of violence significantly. It follows that predicting who and under what conditions violence is more likely to occur, followed by effective management or intervention for those identified as at high risk for violence, could be an effective violence prevention strategy.

This model of violence reduction has been applied successfully in the reduction of future violence among offender populations (Andrews, 1995; Andrews, Bonta, & Hoge, 1990) and high-risk youths (Lipsey & Wilson, 1998) and should be equally applicable to many other types of violent behavior. For example, the government of the United Kingdom has committed significant resources to develop a program, termed the Dangerous and Severe Personality Disorder (DSPD) treatment program, to provide treatment and management services for a relatively small number of persons who are deemed to be at very high risk for future violent and sexual offending and also suffer from severe personality disorders, in particular, those with psychopathy (Maden & Tyrer, 2003). The assessment–prediction–intervention model for violence prevention is therefore based on the accurate assessment of risk and prediction of future violence. However, this model inevitably raises the question as to what type of violence risk assessment and prediction is the most accurate.

Issues in the Assessment and Prediction of Violence

There are several major hurdles to overcome in violence prediction, in particular, the problems inherent in trying to predict low-frequency events, vis-à-vis who will be the perpetrator of violence and when he or she will act violently. Predicting any low-frequency event is difficult and error prone (e.g., consider predicting who will be the next perpetrator of a school shooting and when he or she is likely to act). Making such predictions tends to overidentify suspected perpetrators, that is, committing many false positive errors. Even with a moderately accurate method of prediction, predicting low- or very-low-frequency events, such as serious violence (e.g., mass murder, serial killing, or predatory child sexual abuse) will inevitably result in a high false-positive error rate, that is, identifying many people who are deemed violent but, in fact, are not (see Meehl & Rosen, 1955, and Monahan, 1981, for more detailed discussion). The financial and human costs of such errors are very significant if individuals so identified are detained for preventive purposes. However, the human cost is less if therapeutic or rehabilitative services are offered instead to those identified as at risk.

Another issue is the identification of valid predictors of violent behaviors. In recent years, theoretical developments in risk prediction research have begun to tackle this issue with some success (e.g. Andrews & Bonta, 1998, 2003, 2006; Bonta, Law & Hanson, 1998; Hanson & Bussière, 1998; Hare, 1991, 2003; Monahan & Steadman, 1994). It is probable that the most significant advancement in the technology of risk assessment is the development of structured and standardized risk assessment tools, that is, actuarial tools, to complement, if not replace, the use of unstructured clinical judgments (sometimes referred to as the first-generation of risk assessment approaches) that are prone to error and biases (see Andrews, Bonta, & Wormith, 2006; Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Harris, Rice, & Quinsey, 1993; Monahan & Steadman, 1994).

The use of actuarial risk assessment tools has now become an accepted standard of forensic risk assessment practice (Monahan et al., 2001, pp. 134–135). In most cases, actuarial tools are designed by combining empirically or theoretically derived constructs that are predictive of violence or antisocial activities to guide the forecasting of future antisocial or violent acts. These constructs can be historical (e.g., criminal history), clinical (e.g., personality disorder), or situational (e.g., community support) in nature. They can be further classified as either static/unchangeable, such as criminal history, or dynamic/changeable, such as community support. Some constructs are theoretically derived (e.g., psychopathic personality), whereas others are purely empirically derived (e.g., victim age). Some constructs are more relevant to certain subgroups, such as youths (e.g., peer group influences), whereas others are more typically applicable to adults (e.g., employment history). The "rules" for combining predictor variables in forecasting violence can be quite specific, such as following guidelines in rating predictors and in summing and interpreting the ratings (vis-à-vis the actuarial approach), as opposed to being left to the assessor to use his or her clinical judgment to arrive at a decision (vis-à-vis the structured clinical judgment approach). It should be noted that the term *actuarial* refers to the specified rules that risk predictors are combined and results interpreted and not to the static nature of the risk predictors. Regardless of the approach taken, the predictive efficacies of all tools must be eventually subjected to repeated empirical validation with client groups that differ in demographic characteristics (e.g., age, gender, socioeconomic status, ethnicity), level and type of past violence (e.g., criminal histories, sexual vs. nonsexual offenders), psychiatric diagnosis (e.g., presence of personality disorder, psychosis), intervention received (e.g., treated vs. untreated), the specific criterion being predicted (e.g., violent vs. nonviolent behavior or different types of violent behavior), environmental setting (e.g., clients residing in institutions vs. the community), countries of origin of the research, and so forth.

Since the late 1970s, a range of actuarial risk assessment and risk prediction instruments have been developed in many countries and jurisdictions, all of which have been validated as demonstrating acceptable predictive efficacies for various types of antisocial and/or violent behaviors. With such a wide range of tools, it is reasonable to question which is best to use clinically for predicting violence (see also Campbell, French, & Gendreau, 2009). The answer has important theoretical and practical implications besides the political and legal implications highlighted above. From a theoretical perspective, it is important to know whether an actuarial approach or structured clinical judgment approach is better in violence prediction. Furthermore, how does the predictive efficacy

of theoretically derived violence prediction constructs compare with that emanating from empirically derived ones? How do the predictive efficacies of tools with only static constructs compare with tools that include both static and dynamic constructs? On the practical side, practitioners naturally want to use instruments that give them the best prediction possible, given that major criminal justice and forensic mental health decisions could hinge on the accuracy of such predictions. The predictive efficacies of these tools have been the focus of a number of traditional and meta-analytic reviews. However, there are significant methodological issues with a number of previous meta-analytic reviews, making the interpretation of the results problematic (see section, "Previous Meta-Analyses Conducted With Random-Effects Models and Rationale for the Present Study").

Selection of Risk Prediction Instruments for the Present Study

We selected nine tools for comparison in this meta-analysis. All were used in an actuarial manner in the sense of computing a "risk score" for prediction. All instruments were structured, standardized, and designed to predict antisocial behaviors or violence as their major objectives. Because the use of actuarial tools is now an accepted standard of forensic risk assessment practice (see Monahan et al., 2001, pp. 134–135), it makes sense to compare the predictive efficacies of such tools. They are also instruments designed for assessing nonsexual offenders, contrasting with purposely designed sexual offender risk assessment tools such as the Static 99 (Hanson & Thornton, 1999), the Sexual Violence Risk–20 (SVR-20; Boer, Hart, Kropp, & Webster, 1998), and the Violence Risk Scale—Sexual Offender version (VRS-SO; Olver, Wong, Nicholaichuk, & Gordon, 2007).

The tools included in this study differ along important dimensions often used to categorize risk tools (see Andrews, Bonta, & Wormith, 2006, and Campbell et al., 2009, for detailed discussion of the different "generations" of risk tools). Some are regarded as second-generation tools with mostly static/unchangeable risk predictors (Violence Risk Assessment Guide [VRAG]; Harris, Rice, & Quinsey, 1993; General Statistical Information for Recidivism [GSIR]; Bonta, Harman, Hann, & Cormier, 1996; Risk Matrix 2000 for Violence [RM2000V]; Thornton, 2007); and the Offender Group Reconviction Scale (OGRS; Copas & Marshall, 1998), whereas others are regarded as third-generation tools with mostly dynamic or potentially changeable risk predictors (Level of Service Inventory and revised version [LSI/LSI-R]; Andrews & Bonta, 1995; Historical, Clinical, and Risk Management Violence Risk Assessment Scheme [HCR-20]; Webster, Douglas, Eaves, & Hart, 1997; and the Violence Risk Scale [VRS]; Wong & Gordon, 2006). Although some second-generation tools (including all of the ones selected) demonstrate fairly good predictive validity (Gendreau, Little, & Goggin, 1996; Glover, Nicholson, Hemmati, Bernfeld, & Quinsey, 2002), the sole reliance on static factors for risk assessment has been criticized because these factors do not reflect the complexity of individual functioning and cannot measure changes in risk over time or identify areas for intervention (Andrews et al., 1990; Campbell, French, & Gendreau, 2009; Wong & Gordon, 2006). So-called third-generation tools were designed to overcome these problems. The tools selected for inclusion also differ according to whether their risk predictors have been largely theoretically derived (Psychopathy Checklist—Revised [PCL-R]; Hare, 2003; HCR-20; LSI-R; and VRS), identified empirically (GSIR, RM2000V, and OGRS), or represent a mixture of both approaches (VRAG). Theoretically based tools, unlike atheoretical ones, can also be used to test the validity of the theories on which they are based, can be informed by changing theoretical formulations, and can inform theoretically based clinical activities. For example, as discussed later, the predictive validity of Factors 1 and 2 of the PCL-R may be highly relevant to the treatment of psychopathy. Although we attempted to compare the selected risk tools like for like (all are actuarial, designed to predict risk, and used widely in forensic practice), the results of the study, in addition to answering the key question of which tool has the highest predictive efficacy, can potentially inform other relevant issues, such as the relative performance of second- vs. third-generation instruments and theoretically based vs. empirically based instruments.

Violence Prediction: What Is Being Predicted?

There is no universally accepted definition of violence. Definitions have changed over time and with technological developments. For example, cyber-bullying or bullying over the Internet, with no direct physical or even visual contact, can be deemed a form of violent behavior (Kowalski, Limber, Patricia, & Agatston, 2007). For researchers, a definition of violence such as "behaviors that can or are expected to lead to significant physical or psychological harm" (see Wong & Gordon, 2003, p. 76; see also Wong & Gordon, 2006, p. 288) would probably suffice as a working definition to guide research and theoretical discussions. However, the definition of the criterion or outcome variable for prediction, that is, what is being predicted, is more complex, as it has to withstand tests of validity, reliability, and practicality. The range of possible criterion variables for violence is wide: It includes self-reports to third-party reports of incidents of violence, informal social service or police contact, formal contact or police charges, formal adjudication and court convictions, and incarceration. The frequency or base rate of occurrence also varies: It is generally higher for self-reported incidents and lowest for measures of convictions and incarceration because many police contacts do not result in convictions. The level at which violence is defined can therefore be set according to the goal of the prediction and the practicality of data collection.

The selection of a criterion measure for violence should be guided by the goal of the research and the reliability and construct validity of the variable of choice, as well as the ease and cost of data access and collection. It would be ideal if there were a common metric to assess the level of violence assumed by various criterion variables such that between-study comparisons could be made. To our knowledge, none is available. For the purpose of the present review, the criterion variables, of necessity, were the ones chosen by the various investigations we reviewed. In general, studies usually focused on violent recidivism in the community or violence in institutions, such as assaults against staff. In a recent meta-analysis of the efficacy of risk assessment tools, all violent outcomes in 88 studies could be coded as either institutional violence or violent criminal recidivism (Campbell et al., 2009).

Psychopathy: Assessment, Links to Violence, and Implications

Psychopathy is a psychological construct underpinned by a number of personality traits that, taken together, can be described as a personality disorder. As a point of departure, researchers (e.g., Hare, 2003; Hare et al., 1990) have often referred to Cleckley's (1941, 1976) definition of *psychopathy* in the operationalization of the construct of psychopathy. The personality traits generally considered germane to psychopathy include affective deficits, such as shallow affect, lack of remorse and shame, callousness, and lack of empathy, as well as dysfunctional personality traits related to social functioning, such as egocentricity, manipulativeness, unwillingness to accept responsibility, insincerity, and lying (Cleckley, 1941, 1976; Hare, 2003).

One of the most widely used assessment tools for psychopathy is the PCL-R (Hare, 1991, 2003), a 20-item symptom construct rating scale. The PCL-R is broadly conceptualized as comprising two correlated factors, with Factor 1 tapping the interpersonal and affective personality traits similar to that indicated above and Factor 2 indexing chronic antisocial and unstable behaviors, including impulsivity, a persistent pattern of antisocial and criminal behaviors, and poorly regulated and unstable lifestyle. The number of factors indicative of psychopathy continues to be debated; both three-factor (Cooke & Michie, 2001) and four-factor (Hare, 2003) models (with Factor 1 and Factor 2 each subdivided into two facets) have been proposed. Still, there is much more research on the two-factor as compared with the three- or four-factor models. The debate centers on whether the chronic antisocial characteristics captured by Factor 2 should be part of the conceptualization of psychopathy. The debate is relevant both theoretically and with respect to violence prediction and violence reduction interventions for psychopathy. A major issue is the equivalency of the psychopathy constructs assessed with the PCL-R and the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [*DSM–IV*], American Psychiatric Association, 1994) as well as the *International Statistical Classification of Diseases and Related Health Problems* (10th Rev. [ICD-10]; World Health Organization, 1990) diagnoses of antisocial personality disorder and dyssocial personality disorder, respectively. Such discussion is beyond the scope of this article, but see Hare (2003, pp. 87–92) and Ogloff (2006) for further details.

There is considerable empirical evidence, including a number of meta-analyses, linking psychopathy assessed by the PCL-R with criminality and violence (Edens, Campbell, & Weir, 2007; Walters, 2003a, 2003b). A previous meta-analysis of 18 studies reported the pooled raw effect sizes as 0.79 (95% confidence interval [*CI*] = 0.42–1.18) or area under the curve (AUC) value of 0.71 for the PCL/PCL-R, plus a somewhat larger value of 1.92 for the Psychopathy Checklist: Screening Version (PCL:SV) (Salekin, Rogers, & Sewell, 1996). A subsequent meta-analysis of 10 studies reported the point-biserial correlation between the PCL-R and institutional adjustment (mostly aggression and violence) of 0.25–0.27, which converted to AUC values of 0.64–0.66 (Walters, 2003b). Additional meta-analyses have also investigated the links of Factor 1 and Factor 2 separately to criminality and violence. The two factors are correlated in the range of .5 to .6 (Hare, 2003), and there are important conceptual differences between them. The PCL-R, originally developed to assess disordered personality, has become one of the most widely used instruments for assessing risk and predicting violence in the areas of criminal justice and forensic mental health. That the PCL-R can predict violence has received extensive empirical support (see Hare, 2003, for a review of the evidence). However, it is less clear as to whether its predictive efficacy should be attributed more to Factor 1 or to Factor 2. Aside from theoretical debates over what really constitutes psychopathy, clarifying the links of Factor 1 and Factor 2 with violence has important implications for risk assessment, violence prediction, and interventions to reduce violence. If Factor 2 has stronger links with violence than Factor 1, then it is the criminality and chronic patterns of antisocial behaviors that should be targeted in violence prediction. However, if Factor 1 has stronger links to violence than Factor 2, then violence risk predictions should focus more on assessing core psychopathic personality traits.

In parallel, interventions to reduce the likelihood of violence should be directed toward the factor or factors with significant causative links with violence. Correlational links between a factor and violence are a necessary, but not a sufficient, condition to indicate causation. However, intervention directed toward factors with few or no links to violence would not be effective in reducing violence (Coie et al., 1993).

Interventions aimed to change personality traits represented by Factor 1 would require therapeutic approaches effective in altering egocentricity, callousness, lack of guilt or empathy, and so forth. Personality traits are, by definition, resistant to change (e.g. *DSM-IV*; American Psychiatric Association, 1994) and, as of yet, there is no empirically supported effective intervention that can be used to change Factor 1 traits (see O'Donohue, Fisher, & Hayes, 2003). This is not to say that psychopathy is not treatable. Quite the contrary, a recent review of the evidence did not support the contention that treatment can make those with psychopathy worse (D'Silva, Duggan, & McCarthy, 2004). As well, there is increasing evidence to suggest that treatment can have a positive impact on psychopathic offenders (see Olver & Wong, 2009). However, if Factor 2 is the causative link with violence, then interventions toward antisocial behaviors should be effective in reducing violence.

There is an extensive literature (generally referred to as the "what works" literature) that addresses interventions effective in reducing antisocial and criminal behaviors, essentially Factor 2 characteristics. The risk, need, and responsivity principles have been set forth as guidelines for the delivery of risk reduction treatment and have received considerable empirical support, including meta-analyses (see Andrews & Bonta, 1998, 2003, 2006, 2010; McGuire, 2008). Within this context, the present study examined the efficacy of both Factor 1 and Factor 2 in predicting violence because of the theoretical, policy, and practical implications for violence risk assessment and prediction as well as violence reduction interventions.

Comparison of the Predictive Efficacy of Violence Prediction Instruments

To answer the question of which is the best tool for predicting violence, a proper index for comparison must be used. Two approaches are most frequently used when comparing the predictive efficacies of different risk assessment tools: (a) comparison of two or more tools, with indices of predictive efficacy such as AUC or

correlational statistics and (b) meta-analysis of a fixed-effects model to pool data from different studies for comparison. Studies conducted with the first approach have compared the PCL-R (Hare et al., 1990), the Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 1998), the Violence Risk Assessment Scheme (HCR-20; Webster et al., 1997), the Level of Service Inventory—Revised (LSI-R; Andrews & Bonta, 1995), the Psychopathy Check List: Screening Version (PCL:SV; Hart, Cox, & Hare, 1995), the Lifestyle Criminality Screening Form (LCSF; Walters, White, & Denney, 1991), General Statistical Information on Recidivism (GSIR; Nuffield, 1982), Sexual Violence Risk–20 (SVR-20; Boer et al., 1998), and Static 99 (Hanson & Thornton, 2000). However, these studies have produced inconsistent results, varying from no difference (e.g., Edens, Poythress, & Lilienfeld, 1999; Kroner & Mills, 2001) to large but inconsistent differences in favor of one or more instruments (e.g., Douglas, Ogloff, Nicholls, & Grant, 1999; Hilton, Harris, & Rice, 2001; Gendreau, Goggin, & Smith, 2002; Loza & Green, 2003; Stadtland et al., 2005). Such inconsistencies may be attributable, in part, to variations between the studies, including sample characteristics (e.g., age, gender, size of sample, length of follow-up) and criterion variables (general vs. violent recidivism vs. institutional infractions) and sample (mental health vs. criminal justice vs. a mixture of both), not to mention potential proprietary, biases that were unaccounted for in the studies. Meta-analyses conducted with random-effects models are intended to overcome some of these limitations and should yield more reliable results, as explained below.

Previous Meta-Analyses Conducted With Random-Effects Models and Rationale for the Present Study

Meta-analyses conducted with random-effects models are now considered to be a standard approach for dealing with heterogeneity among studies and have, in many cases, superseded fixed-effects models (see Hunter & Schmidt, 2000). According to Hunter and Schmidt, this is particularly true in social sciences, where studies of effect sizes on certain interventions are most likely based on observational investigation rather than randomized experimental design. The assumption that effect size is the same in all of the studies is not tenable, and the random-effects model is arguably preferable. Walters (2003b) reported a meta-analytic study that compared the effect sizes of (a) the PCL-R, PCL:SV, and PCL:Youth Version (PCL:YV) with (b) that of the LCSF in predicting institutional adjustment and general recidivism. Analyses conducted with inmate samples generated 48 separate effect sizes from 41 studies for the PCL family of tools and 14 separate effect sizes from nine studies for the LCSF. These studies were carried out from 1989 to 2001. Significant heterogeneity in effect sizes across studies was reported. Weighted effect sizes were then calculated to take into account the heterogeneity or significantly different variability in the outcome measures between studies. The 95% confidence intervals (*CI*s) were used to compare the overall weighted effect sizes between the two types of instruments for each of the two outcomes. No significant difference between the two types of instruments in predicting institutional adjustment and general recidivism was found, as their *CI*s overlapped. To examine sources of heterogeneity among studies, Walters then conducted a stratified analysis and found that differences in prediction were related to various study characteristics, such as country of origin of the report, retrospective versus prospective designs, follow-up time, and sample characteristics such as gender, age, and type of participants (mentally disordered vs. prisoners). However, this analysis did not consider violent outcome.

A subsequent meta-analysis by the same author (Walters, 2006) compared effect sizes between professional-rated and self-report risk assessment tools for institutional violent infractions and for general recidivism. The rated tools included the HCR-20, the LCSF, the LSI/LSI-R, the PCL/PCL-R, and the VRAG; there were 13 self-report measures. In all, 25 studies of adult male offenders published between 1986 and 2005, with one or more measures in the two groups of instruments, were included. Using the same random-effects model, weighted analysis suggested moderately larger effect sizes for rated tools compared with self-reported tools, but only for general recidivism. No comparison of effect sizes was made between the five professional-rated instruments, and no attempt was made to adjust for possible moderator effects, such as study features, when comparing instruments.

Edens, Skeem, and Douglas (2006) reported a meta-analysis of 21 studies that compared the effect sizes of the PCL-YV and the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002) in predicting general and violent recidivism among young offenders, with a simple random-effects model. They found that the predictive efficacy of the two measures was comparable and cautioned that there was considerable heterogeneity among the effect sizes, which should be addressed in further studies. No attempt was made to assess or account for study or sample characteristics in this analysis.

Following the study of Edens et al. (2006), Schwalbe (2007) conducted a meta-analytic study to compare a large number of instruments for youths, with similar outcomes in similar populations, based on 42 AUC values from 28 studies of youth recidivism. To address the issue of study heterogeneity in comparing risk instruments, the author used a different approach from previous meta-analyses by means of a two-step process: first, using restrictive inclusion criteria to minimize heterogeneity by including only prospective or longitudinal studies that were carried out with youths and, second, adjusting for potential moderators using a weighted least square (WLS) regression model that took into account random effects of studies while comparing instruments. Potential moderators were labeled as methodological and interval level. Instruments were broadly grouped as second and third generation. The methodological moderators were publication status (published or not), sampling frame (probation or institutional), information source (file review or direct interview), and cross validation (yes or no). Interval-level moderators included sample size, percentage female, percentage minority, and length of follow-up. The WLS analysis indicated significantly larger effects of studies on construct samples than validation samples, third-generation as compared with second-generation measures, and studies with smaller samples; smaller effects occurred with studies utilizing institutional samples as opposed to probation samples. The last three moderators together accounted for 42% of the total variation (based on AUC values), whereas instrument type accounted for only 17%. This study did not provide comparison by individual instruments because there were only 42 AUC values in the analysis. The WLS analysis was able to identify some key

moderators and adjust for them simultaneously while comparing two groups of instruments by second or third generation. This approach has yet to be applied in other settings, for example, efficacy of risk instruments for violence among adults.

In predicting adult violence, the recent meta-analysis by Campbell et al. (2009) was probably one of the most comprehensive comparisons of multiple instruments in predicting institutional violence (76 effect sizes) and violent recidivism (185 effect sizes). These authors pooled 88 independent studies from 1980 to 2006 and compared effect sizes of the HCR-20, the LSI/LSI-R, the PCL/PCL-R, the PCL:SV, the GSIR Scale, and the VRAG for each of the two violent outcomes. The weighted effect sizes showed no differences among instruments for institutional violence, but a somewhat larger effect size of the VRAG compared with the HCR-20 and the GSIR Scale for violent recidivism. The authors applied a conventional random-effects model and weighted effect size analysis. Their study again demonstrated significant heterogeneity among studies for most instruments in the comparison except for the HCR-20, but the sources of the heterogeneity remained unexamined.

In sum, most of the previous meta-analyses reviewed found inconsistent to no difference among instruments they compared. However, the authors of these studies recognized the presence of heterogeneity among studies and attempted to account for them by using random-effects models to calculate weighted effect sizes and by examining the effects of one moderator at a time by a stratified analytic approach. On the basis of subsample data, such analysis has two obvious drawbacks: (a) reduced statistical power to detect differences in predictive efficacy and (b) unexplained variation in effect sizes due to differences in moderators that could not be included in the stratification, which, in practice, usually involves no more than two moderators at a time. Both drawbacks could lead to large standard errors and wide confidence intervals in effect sizes and, hence, could potentially obscure moderate differences between two instruments. The WLS regression analysis reported by Schwalbe (2007) with restrictive study selection criteria could be effective in estimating effects of multiple moderators by using all available data. However, whether the findings could be generalized to studies with larger heterogeneity based on less restrictive inclusion criteria is debatable.

Another source of study heterogeneity, rarely acknowledged in previous meta-analyses, was large individual differences embedded in differences between risk instruments because the comparisons of the tools were based on different studies with different individuals.

Our study objectives were to make a number of improvements on the extant literature, in light of the above methodological issues and conceptual considerations. First, we compared the efficacy of nine widely used instruments to predict violent behavior, including the PCL-R, the PCL:SV, the HCR-20, the VRAG, the OGRS, the RM2000V, the LSI/LSI-R, the GSIR, and the VRS, as well as seven subscales: PCL-R Factor 1 and Factor 2, the 10-item Historical subscale, the five-item Clinical subscale, and the five-item Risk Management subscale of the HCR-20; and the Static and Dynamic scales of the VRS. The PCL-R subscales were included for key conceptual reasons, elucidated above. The HCR-20 subscales are often reported in the literature together with the total score. The VRS is the only tool that has separately identified static and dynamic predictors, and their comparison should also be informative for reasons discussed earlier. Second, in an attempt to minimize sampling error between individuals, we used a within-group design by including only independent studies that compared the predictive efficacy of more than one risk tool on the same individual. Third, we used multilevel regression models (Goldstein, 2003) to estimate the magnitude of heterogeneity or random effects to compare weighted effect sizes among instruments, taking into account random effects, and to examine and adjust for impacts of study features on the differences of effect sizes between the risk instruments. Indeed, our position is that the multilevel regression model can improve on the WLS used by Schwalbe (2007) in several ways. It decomposes total variance by the natural layers in the data structure, such as between studies and between instruments within study. It tests for random effects as the conventional Q statistic does, estimates weighted effect sizes for instruments, and adjusts for moderators or study features simultaneously all within the same model. The model also measures variation of effect sizes among studies that are attributable to different study features and sample characteristics (see Method section for more detail).

Overall, our primary objective was to determine which, among the instruments included in the study, is the most effective violence prediction tool after addressing the methodological issues of earlier meta-analyses. Furthermore, we aimed to evaluate the predictive efficacy of second-generation (static) and third-generation (dynamic) tools, together with comparisons between theoretically derived and empirically derived tools. The study also investigated the links between the PCL-R Factor 1 and Factor 2, and violence.

Method

Selection of Studies

There were four study selection criteria: (a) more than one risk assessment instrument must have been evaluated in the same sample; (b) the reported outcome measures must have clearly involved some form of violent behavior, including violent charges or convictions as well as noncriminal violence against persons or objects; (c) reported statistics must have been reported in sufficient detail for the computation of the instruments' effect sizes; and (d) published or unpublished studies reported since 1999 to capture recent work as most comparative studies of actuarial instruments were reported during the last decade. On the basis of the above selection criteria, key words *risk assessment, violence prediction,* and *comparing risk assessment instruments* were used in literature searches. The databases included PsycINFO, Embase, and Medline, from 1999 to 2008. Authors who were known contributors to the risk assessment literature were added to the searches. Keyword search was also applied to specific criminal justice and behavioral sciences journals. The abstracts were independently read and selected by the first and third authors. Full versions of articles were obtained if the abstract indicated compliance with inclusion criteria a and b. At this point, cross-reference reviews of reference lists of all papers were used to identify any other relevant papers missed in the original search. The first author then read all papers to decide whether sufficient statistics were presented in the article (using tables, figures or text) to calculate effect sizes for subsequent analyses. Unpublished papers identified were solicited from authors by mail or e-mail. A final source of relevant studies came

from recommendations of anonymous journal reviewers, who read an early version of the manuscript. Initial selection of the meta-analysis sample included four Canadian studies with overlapping samples but with different follow-up periods. Advice from journal reviewers prompted us to exclude all but the one with the largest sample size. Two studies (Mills & Kroner, 2006; Mills, Kroner, & Hemmati, 2007) included some individuals who participated in both studies but were assessed by different instruments over different follow up periods. They were coded in the analysis as one study.

These procedures yielded 28 independent studies, published or unpublished, from 1999 to 2008, which compared between two and nine risk assessment instruments, including subscales of the HCR-20, the PCL-R/PCL:SV, and the VRS, and which had sufficient data to be included in the meta-analysis.

Criterion or Outcome Measures

Generally, outcome measures reported in the literature are based on violent criminal reconvictions extracted from official records after the individual has been released and followed up for some time in the community (oftentimes referred to as community violence), or some form of physical aggression or violence toward others based on staff observations documented in institutional case files when the individual (often a forensic psychiatric patient) was in custody in an institution (oftentimes referred to as institutional violence). To address the potential concern raised by an anonymous reviewer that the many criterion variables of violence reported in the literature could represent qualitatively different types of violent behaviors, we created a covariate consisting of four violence categories based on the study outcome descriptions. The following four categories were developed because, first, they appear to be able to best sort the articles into mutually exclusive categories and, second, they are sufficiently conceptually different as to potentially represent different types of violent behaviors. The four categories are the following: (a) specific mention of actual physical aggression or violent acts toward institutional staff and others (excluding threats or attempts) perpetrated within an institution (see Morrissey et al., 2007); (b) actual, attempted, or threats of harm to others primarily determined with the HCR-20 definition of violence as per Webster et al. (1997; see de Vogel, de Ruiter, de Hildebrand, Bos, & van de Ven, 2004); (c) broadly defined violent criminal recidivism from official records which included sexual offense and robbery (e.g., Wong & Gordon, 2006); and (d) violent criminal recidivism from official records excluding sexual offenses and robbery (see Coid et al., 2009). All of the original articles were reviewed and coded into one of four categories by the first author. When there was overlap in the criteria in the article, the predominant category that best represented the outcomes was selected. Ten articles were selected and retrospectively reviewed by the second author. There were agreements on the categorization on 7 of 10. On further discussion, all disagreements were resolved in favor of the categorization of the first author: One was misread by the second author, and two were agreed to after further reviewing of the criteria use. It was not possible to develop even more precise categories to cover the broad literature because of the lack of detailed descriptions in the studies, and outcomes of convenience were often used. The outcome categories also overlapped according to type of participant and country of studies. However, according to this categorization of outcome, 62.5% of studies of forensic psychiatric patients used category a; 62.5% of studies of mixed offender and psychiatric patients used category c, and 25% used category b. A total of 50% of studies of prisoners used category d, with 33.3% of these studies also using category c. Some studies reported multiple outcomes that included both violent and nonviolent acts. Nonviolent acts, such as general criminal recidivism and behavior that involved only verbal aggression, were excluded from the meta-analyses. In view of the importance of outcome criteria in this type of research, we make some specific suggestions in the Recommendations section regarding future attempts to resolve these issues.

The question of whether sexual offenses should be considered as violent offending is open to interpretation. Our position that they should be is in line with the views of the authors of a number of risk assessment tools, such as the HCR-20: "All sexual assaults should be considered violent behaviour" (see Webster et al., 1997, p. 25; see also the VRAG, Quinsey et al., 1998, p. 142; and the VRS, Wong & Gordon, 2006, p. 288). However, if a certain study author chose to exclude sex offenders from his or her study for specific reasons, then we accepted such reasoning in our choice of studies to review. The complexity of the issue is illustrated by the following: An offender with a long history of nonsexual offending but with an index sexual offense may be deemed a sex offender for the purpose of his or her current sentence management; on the other hand, an offender who committed a minor sex offense many years ago but more recently was convicted of a serious, nonsexual violent offense is likely to be deemed a nonsexual violent offender.

Study Features Included in the Analyses

Differences in study characteristics and sample variables must be taken into account, as they could act as covariates or moderators in the estimation of instrument efficacy. The effects of study characteristics and sample variables should be quantified and adjusted in order to obtain independent estimates of the effect sizes in violence prediction. In addition, it is well established that risk of violence is strongly associated with sex, age, and certain forms of mental disorder, for example, antisocial personality disorder. Other potential contributing factors include retrospective compared with prospective study design, different operationalizations of the criterion variable (as discussed above), and country of origin of studies. Although there is an inevitable lack of uniformity in the use and/or reporting of such factors, we endeavored to extract as much information as possible from all studies to include in our analyses.

Sample variables used in the study were as follows: (a) mean age, (b) percentage of male participants, (c) study type (retrospective vs. prospective), (d) country where the study was carried out, (e) type of participants (nonsexual offenders or prison inmates vs. forensic mental health patients vs. mixed samples), (f) type of violence, and (g) average follow-up time in months. For studies reporting a number of follow-up times (e.g. Craig, Beech, & Browne, 2000; Dahle, 2006; Snowden, Gray, Taylor, & MacCulloch, 2007; Wong & Gordon, 2006), we used data at the time point for which the maximum sample size was reported.

Effect Size Measure

There are three commonly used measures of effect size for predictive accuracy: Cohen's *d*, receiver–operating characteristics area under the curve (AUC), and the correlation coefficient. Cohen's *d* is well established for meta-regression analysis with covariates or mediators; it is used particularly to deal with random effects (Goldstein, Yang, Tuner, Omar, & Thompson, 2000). It can easily be converted into the other two measures for comparison purposes (M. E. Rice & Harris, 2005). Cohen's *d* values have been calculated directly for eight studies in which the risk assessment instrument's means and standard deviations of scores for groups, with and without the defined violent outcomes, were available. For one study (Grann, Belfrage, & Tengström, 2000), the Cohen's *d* effect size value was approximated on the basis of medians and quartiles observed in graphs. For another 13 studies that reported various correlation coefficients, we converted the correlation coefficient *r* to Cohen's *d* using the formula $d \approx r[pq(1 - r^2)]^{-0.5}$ (Hunter & Schmidt, 2004; M. E. Rice & Harris, 2005), where *p* was the base rate of the outcome and $q = 1 - p$. When the base rate was close to 50%, the formula was reduced to $d \approx 2r(1 - r^2)^{-0.5}$ or $d \approx [(n - 2)/n]^{0.5}\,[2r(1 - r^2)^{-0.5}]$ for small samples. If the study reported the correlation coefficient separately for men and women, the *d* value was computed for men and women separately. The sample size reported for assessing each instrument was used as a weighting factor in the meta-regression model for the effect size analysis. For eight studies that reported only the AUC values, a direct conversion of the *AUC* value to the *d* value was carried out on the basis of the table of M. E. Rice and Harris (2005).

The effect size as *d* value was calculated for each risk instrument assessed for each study. In total, 174 effect size values from 28 studies were included in the analysis.

Multilevel Regression Models

Multilevel regression models developed from educational effectiveness assessment have been shown to provide optimal flexibility both to disentangle random effects by sources of variation (Goldstein, 2003) and to estimate effects of sample characteristics simultaneously in meta-analysis with complex data structure (Goldstein et al., 2000). This approach has followed the principle of meta-regression methods (Greenland, 1987) for observational data with measurable moderators in epidemiology. It is advanced by building in random parameters to identify and quantify sources of variation or heterogeneity. The merits of multilevel models in comparison with standard statistical approaches to meta-analysis of effect sizes and odds ratios have been explicitly demonstrated (Tuner, Omar, Yang, Goldstein, & Thompson, 2000). Applications of multilevel models can be found in health (Leyland & Goldstein, 2001; N. Rice, Carr-Hill, Dixon, & Sutton, 1998; Von Korff, Koepsell, Curry, & Diehr, 1992) and in educational (Goldstein, 2003), as well as social, political, and behavioral studies (Sampson, Raudenbush, & Earls, 1997; Yang, Heath, & Goldstein, 2000). Software tools for multilevel analysis models are now available in many major statistical packages, such as SAS, Stata, SPSS, and SPlus.

The hierarchical features in many published risk assessment studies are well suited for multilevel regression analysis. Hierarchical features pertain to the condition that several risk assessment instruments are applied to the same sample of individuals within a study; they also afford analysis of the marked heterogeneity or random effects between studies, including differences in sample characteristics, such as sex and age, and differences in study characteristics, such as country of origin, follow-up time, prospective versus retrospective designs, and outcome categories. Multilevel models consist of two parts: (a) random parameter estimates for random effects at the level of variation sources, and (b) fixed parameter estimates for mean effects of covariates or moderators. We used random parameter estimates to quantify and disentangle total variation in effect sizes to the level of study (between-study variation) and level of instrument (within-study variation), and fixed parameter estimates to examine independent effects of study features or moderators mentioned above. Weighted effect sizes of instruments adjusted for random effects and effects of moderators were estimated and compared within the framework of multilevel models.

Design and Analytic Strategy

Compared with a between-subjects design, a within-subject design yields smaller random sampling error and thus provides better statistical power to detect differences of interest. We chose a within-subject design for this study, meaning that only studies evaluating more than one risk assessment instrument on the same sample of subjects with the same outcome variable were included in the analyses. Such a model provides a natural three-level hierarchical structure: that is, risk instruments nested within studies and participants nested within instruments. Three-level multilevel regression models were therefore applied.

When comparing effect sizes among risk instruments, the PCL-R was used as the reference category because it was one of the most widely used tools and was reported by most studies included in our meta-analysis. The mean differences in effect sizes between other tools and the PCL-R were estimated in the multilevel regression model and tested by the generalized Wald test after fitting a model. The program MLwiN (Rasbash et al., 2000) was used to perform multilevel regression analysis. All models were weighted by the inverse of sample size.

Three nested models were fitted. Model A, a three-level variance component model, provided one estimate for the overall mean effect size in the regression, together with two variances of random effects segregated into study and instruments within study. This model was fitted to quantify the heterogeneity of effect sizes by the sources and to test the presence of random effects between studies and within study. If the two variances of random effects were no more than chance or sampling error, Model A was reduced to a simple fixed-effects model that estimated a pooled mean estimate of all studies, with the meta-analysis sample considered homogeneous.

Model B, an elaboration of Model A, provided a mean estimate of effect size for each instrument and its subscales, with PCL-R as the reference. It explored the magnitude of the total variance attributable to different risk assessment instruments. Changes in the study-level variances between the two nested models are measures of the contributions of the instruments to the variance of effects sizes. For example, Model A may estimate a variance of effect sizes at study level as 0.5, and Model B may estimate a

variance as 0.3, which is smaller than the Model A estimate. If the instrument effects in Model B are significant, the difference between 0.5 and 0.3 is the variance component in the total effect size explained by the differences in instruments. For comparison purposes, we also estimated a fixed-effects only regression model (B1) for the predictive efficacy of all instruments by removing variance components from Model B, that is, ignoring random effects in the effect sizes. Comparison of models with and without taking into account random effects between studies can indicate the contributions of random effects on the estimates of efficacy of risk instruments.

In Model C, study characteristics, such as type of participants, country of origin, type of study, age and sex of participants, time of follow-up, type of criterion measures, and so forth, were included to examine the extent to which study characteristics contributed to the effect size variation across studies. Model C provided estimates of the contributions of study characteristics to the effect size of violence prediction, such that comparisons of the effect sizes among risk instruments could be made independent of study characteristics. Interactive effects between some study features, such as sex and efficacy of instruments, were considered in order to understand the reason for the differences between instruments.

The adequacy of models was assessed by the goodness of fit, with the likelihood ratio test of chi-square statistic, that is, the difference in the −2log-likelihood values between any two nested models. The difference in predictive efficacy estimates between instruments was tested with the generalized Ward test in MLwiN. All models in the study were fitted by MLwiN. Conversions of Cohen's *d* values from the ROC AUC values were carried out following the table in M. E. Rice and Harris (2005) when appropriate.

It is well established that correlation coefficients and AUC values based on a smaller sample can be inflated and may have a direct impact on the effect size measures. In our multilevel regression models for aggregated data, the raw sample size for the calculation of each effect size value was used as a weighting factor to address this issue; it was applied in all models presented.

Results

Features of Studies

A summary of studies included in the meta-analysis is presented in Table 1. The majority ($k = 11$) were carried out in the United Kingdom, followed by Canada ($k = 9$), with three in Sweden, two in Holland, three in the United States, and one in Germany. Nine studies were prospective; 19 were retrospective. In the latter, participants were identified using archival information and were then followed up to assess their violent outcomes. The total sample size in the meta-analysis was in a range of 6,348–7,221 by different instruments and a range of 34–1,650 by study. Only those original studies in which some form of violence was identified as the outcome variable were included in the meta-analysis. As such, some samples in the present study may be lower than those in the original reports.

The mean age of participants in the sample was 33.3 years (range = 24–44 years), with 17 studies consisting of male participants only, 9 of mixed sex, and 2 of women only. Participants in the studies were mostly prisoners ($k = 12$); others were psychiatric patients residing in forensic hospitals ($k = 8$) or offenders with mental disorders ($k = 8$). Specific categories of mental disorder were not evaluated in the present study. The overall mean follow-up time was 43.8 months, varying from 3 to 133 months. For two studies that did not report the follow-up time, the average follow-up time of all studies was used as an estimation.

In total, 18 risk assessment tools, including subscales of the instruments, were evaluated. These included the VRAG ($k = 17$), the HCR-20 ($k = 16$), HCR-20 subscales Historical, Clinical, and Risk Management (ks = 18, 14, and 12, respectively), the PCL-R ($k = 16$), the PCL:SV ($k = 8$), PCL-R/PCL:SV Factors 1 and 2 (ks = 13 and 13), the RM2000V ($k = 3$), the GSIR ($k = 3$), the LSI/LSI-R ($k = 5$), the OGRS ($k = 2$), the VRS ($k = 4$), and the VRS Static and Dynamic scales (ks = 3 and 3). For one study (Craig et al., 2000), the SVR-20 and the Static99 were evaluated for violence of nonsexual offenders. The overall base rate of violent outcomes was 24.9%, ranging from 4.8% of violence recidivism of patients in a 5-year follow-up to 100% of physical aggression by female patients with mental disorders in a nearly 2-year follow-up. The raw effect size varied from −0.187 to 1.34.

Pooled Effect Size and Its Random Effects

The raw effect sizes were symmetrically distributed with a mean of 0.65 (variance = 0.096). The pooled effect size and its distribution of variance based on Model A (Table 2) demonstrated that, of the total estimated variance (0.0923), 48.2% (0.0445/0.0923) was due to the difference or random effects across studies and 51.7% (0.0478/0.0923) to the different instruments within the study. Both variances were statistically significant, Wald test $\chi^2(1) = 8.26$, $p = .004$, and $\chi^2(1) = 43.15$, $p < .0001$, respectively. The results suggested significant heterogeneity of effect sizes across studies as well as across instruments. On the basis of this model with the pooled effect size estimation as 0.66, a 95% distribution range of such effect sizes among all studies was estimated to vary from 0.25 to 1.08 and among all instruments, from 0.23 to 1.09, respectively.

Effect Sizes of Instruments From Fixed- and Random-Effects Models

We first fitted a single level or fixed effect regression model (Model B1 in Table 2) with weighting factor to compare effect sizes of instruments. By ignoring the heterogeneity among studies on the outcome measure, the simple meta-regression analysis suggested that effect sizes for eight instruments and their subscales, including the VRAG, the HCR-20, the PCL:SV, the OGRS, the GSIR, the RM2000V, the VRS Static subscale, the VRS Dynamic subscale, and PCL-R Factor 2, were significantly larger than that of PCL-R total, whereas the effect sizes of PCL-R Factor 1, the five-item Clinical subscale and the five-item Risk Management subscale in HCR-20, and other sexual risk assessment instruments were significantly smaller than that of PCL-R. However, by allowing effect sizes to vary randomly among different studies and estimating a variance at study level for such difference, the two-level random effect model (Model B2 in Table 2) showed considerably improved goodness of fit over Model B1, with the likelihood ratio test $\chi^2(1) = 800.9$, $p < .0001$. Furthermore, by

Table 1
Description of Studies Included in the Meta-Analysis

Authors of study	Country[a]	Type of participants[b]	% Male participants	Mean age of participants (years)	Study type[c]	Mean follow-up (months)	Sample size[d]	Outcomes	Rate of violence (%)	Instruments assessed	Effect size[e]	Source data of effect size
Belfrage et al. (2000)	2	1	100.0	34.1	2	8	41	Institutional violent acts vs. nonviolent acts	19.5	HCR-20 H10 C5 R5 PCL:SV Part 1 Part 2	1.20 0.36 0.73 0.95 0.78 0.289 1.065	Mean, standard deviation
Coid et al. (2009)	4	1	100	30.7	2	24	1271 1281 1339 1337 1337 1350 1345 1353 1353 1343	Violent re-offending vs. nonoffending and other offending	13.2	HCR-20 H10 C5 R5 RM2000V OGRS PCL-R Factor 1 Factor 2 VRAG	0.622 0.585 0.500 0.325 0.707 0.825 0.500 0.141 0.665 0.740	AUC value
Coid et al. (2009)	4	1	0	28.2	2	25.2	302 302 302 302 303 290 303 302 302 302	Violent re-offending vs. nonoffending and other-offending	8.2	HCR-20 H10 C5 R5 RM2000V OGRS PCL-R Factor 1 Factor 2 VRAG	0.745 0.675 0.795 0.320 0.545 0.141 0.870 0.545 0.780 0.545	AUC value
Cooke et al. (2002)	4	1	100	26.8	1	Missing	190	Any violent reconviction vs. other	20.0~30.0	HCR-20 H10 C5 R5 PCL-R VRAG	0.707 0.665 0.522 0.160 0.545 0.622	Mean, standard deviation AUC value
Craig et al. (2006)	4	1	100	33.9	1	120	131	Violent recidivism vs. nonviolent/non-sex recidivism	16.0	RM2000V SVR-20 Static-99	0.972 0.159 0.426	Correlation coefficient (Spearman's)
Dahle (2006)	6	1	100	29.8	2	120	86	Violent reconviction vs. other	32.6	PCL-R HCR-20 LSI-R	0.668 0.644 0.467	Correlation coefficient (Pearson's)

(*table continues*)

Table 1 (*continued*)

Authors of study	Country[a]	Type of participants[b]	% Male participants	Mean age of participants (years)	Study type[c]	Mean follow-up (months)	Sample size[d]	Outcomes	Rate of violence (%)	Instruments assessed	Effect size[e]	Source data of effect size
de Vogel et al. (2004)	5	3	89.0	26.0	1	72.5	119	Violent reconviction	36.1	HCR-20 H10 C5 R5 PCL-R Factor 1 Factor 2	1.27 1.11 1.08 1.11 0.49 0.469 1.056	Correlation coefficient (Pearson's)
De Vries Robbe et al. (2006)	5	3	100.0	30.0	1	112.8	50	Violent reconviction	Missing	VRS PCL-R HCR-20	1.10 0.905 1.10	AUC value
Dolan & Fullam (2007)	4	1	100.0	35.5	2	12	136 80	Violence against others	53.7 48.8	VRS VRS-Static VRS-Dynamic HCR-20 H10 C5 R5	0.72 0.42 0.75 0.80 0.62 0.85 0.60	Cohen's *d* effect size
Douglas et al. (1999)	1	2	61.0	38.1	1	20.6	193	Any violence	38.0	HCR-20 H10 C5 R5 PCL:SV Factor 1 Factor 2	1.00 0.825 0.470 0.870 0.660 0.430 0.660	AUC value
Douglas et al. (2005)	1	1	100.0	38.3	1	92.4	188	Violent recidivism vs. nonviolent recidivism	49.5	HCR-20 H10 C5 R5 VRAG PCL-R Factor 1 Factor 2 PCL:SV Part 1 Part 2 PCL:SV VRAG	1.19 0.76 1.16 1.13 1.04 0.94 0.377 1.113 0.67 0.410 1.009 1.02 0.71	Mean, standard deviation
Doyle & Dolan (2006)	4	3	67.0	40.6	2	5.5	112	Violent recidivism vs. nonviolent recidivism	18.8	H10 PCL:SV Part 1 Part 2 VRAG	0.56 0.68 0.65 0.559 0.50	Mean, standard deviation

Table 1 (*continued*)

Authors of study	Country[a]	Type of participants[b]	% Male participants	Mean age of participants (years)	Study type[c]	Mean follow-up (months)	Sample size[d]	Outcomes	Rate of violence (%)	Instruments assessed	Effect size[e]	Source data of effect size
Doyle et al. (2002)	4	2	97.0	35.5	1	3	87	Violent physical assault vs. nonviolent assault	51.7	H10	0.69	Mean, standard deviation
Edens et al. (2006)	3	2	59.0	30.0	2	4.6	741	Violent behavior/crime	18.6	VRAG PCL:SV	0.865 1.11	AUC value
Glover et al. (2002)	1	1	100	28.0	1	23.5	78	Violent recidivism vs. nonrecidivism	43.6	GSIR PCL-R Factor 1 Factor 2 VRAG	1.06 0.58 −0.139 0.432 0.94	Mean, standard deviation
Grann et al.[f] (2000)	2	3	92.8	32.5	1	24	404	Violent crime vs. nonviolent crime	22.5	H10 VRAG	0.72 0.71	Median, quartile
Gray et al. (2003)	4	2	76.5	33.0	1	3	34	Physical aggression vs. other	32.4	HCR-20 H10 C5 PCL-R Factor 1 Factor 2	1.34 1.02 1.20 0.80 0.396 0.749	Correlation coefficient (Spearman's)
Gray et al. (2007)	4	3 (Offenders with intellectual disabilities)	81.4	31.6	1	60	115 132 124 129 139 139 139 138	Violence re-offending (including robbery, rape, kidnap) vs. other	4.8	VRAG PCL:SV Factor 1 Factor 2 HCR-20 H10 C5 R5	0.870 0.870 0.608 0.471 0.785 0.791 0.780 0.505	AUC value
Gray et al. (2007)	4	3 (Offenders without intellectual disabilities)	85.6	32.0	1	60	420 775 667 762 898 889 893 894	Violence re-offending (including robbery, rape, kidnap) vs. other	11.2	VRAG PCL:SV Factor 1 Factor 2 HCR-20 H10 C5 R5	0.910 0.700 0.470 0.780 0.660 0.700 0.190 0.470	AUC value
Grevatt et al. (2004)	4	2	100.0	44.0	1	6	44	Physical aggression vs. other	29.5	H10 C5 VRS VRS-Static VRS-Dynamic	0.116 0.362 −0.12 0.077 −0.154	Mean, standard deviation
Hilton et al. (2001)	1	1	100.0	24.0	1	82.5	88	Violent recidivism vs. nonviolence	21.6	PCL-R VRAG	0.91 1.13	Correlation coefficient (unspecified)

(*table continues*)

Table 1 (*continued*)

Authors of study	Country[a]	Type of participants[b]	% Male participants	Mean age of participants (years)	Study type[c]	Mean follow-up (months)	Sample size[d]	Outcomes	Rate of violence (%)	Instruments assessed	Effect size[e]	Source data of effect size
McDermott et al. (2008)	3	2	84.0	45.6	2	30	108	Physical violence aggression to others	28.0	PCL-R Factor 1 Factor 2 VRAG HCR-20 H10 C5 R5	0.283 0.212 0.354 0.141 0.540 0.396 0.396 0.585	AUC value
Mills & Kroner (2006)	1	1	100.0	29.9	1	14	209	Violent recidivism vs. nonviolence	29.0	GSIR LSI-R PCL-R VRAG	0.69 0.59 0.40 0.59	Correlation coefficient (unspecified)
Mills et al. (2007)	1	1	100.0	27.9	1	55	83	Violent recidivism vs. nonviolence	35.0	HCR-20 H10 C5 R5 VRAG	0.84 0.64 0.96 0.76 0.61	Correlation coefficient (unspecified)
Morrissey et al. (2007)	4	3	100.0	38.0	2	12	54 60	Physical aggression vs. other	59.3	HCR-20 PCL-R Factor 1 Factor 2	0.94 0.08 −0.187 0.319	Correlation coefficient (Spearman's)
Nicholls (2004)	1	2	100	36	1	22.7	117 117 117 117 146 146 146	Any violent behavior/ crime vs. other	79.3	HCR-20 H10 C5 R5 PCL:SV Part 1 Part 2	0.82 0.608 0.396 0.865 0.47 0.354 0.396	AUC value
Nicholls et al. (2004)	1	2	0	42	1	22.7	75 75 75 75 90 90 90	Any violent behavior/ crime vs. other	100	HCR-20 H10 C5 R5 PCL:SV Part 1 Part 2	0.779 0.95 0.545 0.78 0.622 0.431 0.622	AUC value
Snowden et al. (2007)	4	2	100.0	37.7	1	60.0	320	Violent reconviction	14.7 13.0	OGRS VRAG	0.830 0.976	AUC value
Tengström (2001)	2	3	100.0	33.0	1	86	106	Violent crime	29.0	H10 VRAG PCL-R Factor 1 Factor 2	1.00 0.665 1.150 −0.139 0.432	
Warren et al. (2005)	3	1	0.0	30.0	1	Missing	132	Violent crime/behavior vs. other	43.2	HCR-20 H10 C5 R5 PCL-R	0.28 −0.03 −0.11 0.00 0.28	Mean, standard deviation

Table 1 (*continued*)

Authors of study	Country[a]	Type of participants[b]	% Male participants	Mean age of participants (years)	Study type[c]	Mean follow-up (months)	Sample size[d]	Outcomes	Rate of violence (%)	Instruments assessed	Effect size[e]	Source data of effect size
Wong & Gordon (2006)	1	1	100.0	38.8	1	52.8	918	Violent reconviction	31.3	VRS	0.872	Correlation coefficient (Pearson's)
						52.8	918			VRS-Static	0.651	
						52.8	918			VRS-Dynamic	0.872	
						50.7	696			PCL-R	0.872	
						42.4	701			GSIR	0.675	
Wormith et al. (2007)	1	1	100	25.7	1	133.2	61	Violent reconviction	55.0	PCL-R	0.618	Correlation coefficient (unspecified)
										YLS/CMI	0.687	

Note. HCR-20 = Historical, Clinical, and Risk Management Violence Risk Assessment Scheme; H10 = 10-item Historical subscale of the HCR-20; C5 = 5-item Clinical subscale of the HCR-20; R5 = 5-item Risk Management subscale of the HCR-20; PCL:SV = Psychopathy Checklist: Screening Version; RM2000V = Risk Matrix 2000 for Violence; OGRS = Offender Group Reconviction Scale; PCL–R = Psychopathy Checklist—Revised; VRAG = Violence Risk Assessment Guide; SVR-20 = Sexual Violence Risk–20; LSI-R = Level of Service Inventory—Revised; VRS = Violence Risk Scale; GSIR = General Statistical Information for Recidivism; YLS/CMI = Youth Level of Service/Case Management Inventory.

[a] 1 = Canada; 2 = Sweden; 3 = United States; 4 = United Kingdom; 5 = Holland; 6 = Germany. [b] 1 = prisoner; 2 = psychiatric patients; 3 = mixed sample. [c] 1 = retrospective follow-up; 2 = prospective follow-up. [d] No. of participants used to calculate the effect size. [e] Effect size $d = \sqrt{(n-2)/n} \times 2 \times r/\sqrt{(1-r^2)}$ if r is Pearson's or Spearman's correlation coefficient, and $d = r/\sqrt{(pq(1-r^2)}$ if r is point-biserial correlation coefficient or unspecified. The Cohen effect size d was calculated if mean scores and standard deviations of the violent and nonviolent groups were provided for each instrument scale. The pooled SD for Cohen's effect size d was defined as $\sqrt{((n_1-1)SD_1 + (n_2-1)SD_2)/(n_1+n_2-2)}$. [f] The study by Grann et al. (2000) did not present means and standard deviations in numbers but plots of distribution by outcome categories. The median and 25% percentiles were read off from the plots to approximate means and standard deviations.

allowing effect sizes to vary among instruments and by disentangling variance components for studies and for instruments respectively, the three-level model (Model B3) further significantly improved the goodness fit of Model B3 over Model B2, with the likelihood ratio test $\chi^2(1) = 398.6$, $p < .0001$. Among the three models, the single-level model B1 had the smallest standard errors of efficacy estimates for all instruments, hence, a greater chance for Type I error (i.e., more false significant findings). The considerably larger log-likelihood value of Model B1 than the other two nested models indicated the worst fit of the model to the data.

Results of Model B3 suggested that only HCR-20 had a larger effect size than PCL-R, $\chi^2(1) = 12.86$, $p = .0003$, and only PCL-R Factor 1 had a significantly smaller effect size than PCL-R, $\chi^2(1) = 21.36$, $p < .0001$. No significant differences were found between the remaining risk assessment instruments compared with the PCL-R. The goodness of fit of Model B3, after estimating instrument differences, was a significant improvement to Model A, with the likelihood ratio test, $\chi^2(16) = 64.20$, $p < .0001$. The differences among instruments reduced the total variance in effect sizes by 22.6% between Models A and B3 and, in particular, reduced the variance within study by 48.0%, from 0.0445 in Model A to 0.0249 in Model B3. This finding signifies that a large proportion of variation in the mean effect sizes between studies was related to the correlation between instruments within studies. Failing to take into account such correlation in comparing the predictive efficacy, such as in Model B1 and B2, can lead to underestimations of standard errors of parameters and could produce false positive findings.

Based on Model B3, we still observed a considerable amount of variation among studies, $\chi^2(1) = 10.29$, $p < .0001$, and among instruments, $\chi^2(1) = 35.00$, $p < .0001$. It was reasonable to hypothesize that differences in effect sizes between studies could be related to study features that tended to vary from study to study. Taking such variability into account may reduce the effect size estimates for instruments. In Model C1, we accounted for study and sample characteristics that were not accounted for in Model B3. The likelihood ratio test suggested that Model C1 represented a significant improvement in the goodness of fit over Model B3, $\chi^2(11) = 30.55$, $p = .001$, and a marked reduction of the study level variance by 70.3% (from 0.0445 in Model A to 0.0132 to Model C1). The results strongly supported our hypothesis that study and sample characteristics could be major contributors to differences between studies. On the basis of this model, the major study features that contributed to the effect size were the origin of study, type of study, time of follow-up, and participants' sex. After controlling for such difference in study features in Model C1, the effect size of HCR-20 was still significantly larger than that of PCL-R, $\chi^2(1) = 12.45$, $p = .0004$, and the effect size of PCL Factor 1 was still significantly smaller than that of PCL-R, $\chi^2(1) = 20.89$, $p < .0001$, and that of Factor 2, $\chi^2(1) = 31.79$, $p < .0001$. The rest of the instruments were no different from the PCL-R in their predictive efficacy. Raw effect sizes of risk instruments and their estimated efficacy determined with Model C1 are presented in Table 3. It can be seen that after taking into account the data structure, the country of study, participants' sex, mean age of participants, follow-up time to the outcome, and type of study, the predictive efficacy of the risk instruments all fall between a range of 0.56 and 0.71 in terms of the AUC value, with the majority falling within a narrow range of 0.65–0.69. According to the general rule of the effect size, d values = 0.2, 0.5, and 0.8 are small, medium, and

Table 2
Effect Sizes of Risk Instruments to Predict Violent Behavior From Multilevel Regression Analysis

Variable	Model B1 Estimate (*SE*)	Model A Estimate (*SE*)	Model B2 Estimate (*SE*)	Model B3 Estimate (*SE*)
Overall	0.636 (0.016)	0.664 (0.046)	0.666 (0.047)	0.637 (0.064)
Instrument: PCL-R	Reference		Reference	Reference
OGRS	0.088 (0.028)***		0.134 (0.030)***	0.130 (0.141)
VRAG	0.129 (0.022)***		0.105 (0.024)***	0.089 (0.071)
RM2000V	0.072 (0.029)*		0.132 (0.031)***	0.204 (0.153)
HCR-20	0.095 (0.022)***		0.136 (0.024)***	0.243 (0.068)***
H10	0.021 (0.022)		0.061 (0.023)**	0.059 (0.067)
C5	−0.128 (0.022)***		−0.080 (0.024)***	0.038 (0.071)
R5	−0.152 (0.023)***		−0.107 (0.024)***	0.051 (0.073)
PCL:SV	0.184 (0.026)***		0.094 (0.030)**	0.068 (0.087)
PCL-R/PCL:SV Factor 1	−0.315 (0.023)***		−0.310 (0.024)***	−.335 (0.073)***
PCL-R/PCL:SV Factor 2	0.084 (0.022)***		0.090 (0.024)***	0.061 (0.073)
LSI/LSI-R	−0.058 (0.055)		−0.042 (0.062)	−0.023 (0.129)
GSIR	0.073 (0.036)*		−0.069 (0.040)	0.063 (0.119)
VRS	0.013 (0.033)		−0.135 (0.041)**	−0.025 (0.117)
VRS Static	0.148 (0.034)***		0.009 (0.042)	−0.047 (0.129)
VRS Dynamic	0.188 (0.034)***		0.041 (0.042)	0.013 (0.129)
Others[a]	−0.341 (0.064)***		−0.525 (0.106)***	−0.424 (0.242)
Level of variance				
Between study		0.0445 (0.015)**	0.0521 (.014)**	0.0500 (0.015)**
Within study		0.0478 (0.007)***		0.0249 (0.004)***
−2 log-likelihood	1,400.91	265.60	599.96	201.39
χ^2 for goodness of fit	1,135.30*** (Model B1 vs. A)		800.95*** (Model B1 vs. B2)	398.57*** (Model B2 vs. B3)

Note. $N = 174$. See text for description of the models. PCL-R = Psychopathy Checklist—Revised; OGRS = Offender Group Reconviction Scale; VRAG = Violence Risk Assessment Guide; RM2000V = Risk Matrix 2000 for Violence; HCR-20 = Historical, Clinical, and Risk Management Violence Risk Assessment Scheme; H10 = 10-item Historical subscale of the HCR-20; C5 = 5-item Clinical subscale of the HCR-20; R5 = 5-item Risk Management subscale of the HCR-20; PCL:SV = Psychopathy Checklist: Screening Version; LSI/LSI-R = Level of Service Inventory/Revised version; GSIR = General Statistical Information for Recidivism; VRS = Violence Risk Scale.
[a] Others included the Sexual Violence Risk–20 and the Static 99.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

large effects, respectively (Cohen, 1988). Thus the instruments included in this study demonstrated medium effects for predicting violence risk. As PCL-R Factor 1 showed no predictive effect (*CI* overlaps with 0), the efficacy of the PCL-R and the PCL:SV were mainly explained by Factor 2 (or Part 2 for PCL:SV). The three subscales of the HCR-20 were all predictive, with medium effects respectively. The larger effect size in the HCR-20 total seemed to suggest some incremental effects among the subscales. For the VRS, the Dynamic scale appeared to contribute more to the total than to the Static scale, but there was no significant difference between them as a result of the limited number of studies of this instrument.

Association of Study Characteristics With Predictive Effect Size

Results in Model C1 demonstrated that country of study, mean time of follow-up, type of study, and sex of participants significantly affected predictive efficacy for violent outcomes. In general, the U.S. studies reported smaller effect sizes by a mean of −0.513 compared with studies conducted in Canada, $\chi^2(1) = 20.99$, $p < .001$. Prospective studies reported a larger effect size by a mean of 0.156 compared with retrospective studies, $\chi^2(1) = 4.82$, $p = .028$. Longer follow-up time was associated with larger effect size, $\chi^2(1) = 7.73$, $p = .0005$, and studies on women and mixed samples reported larger effect sizes by a mean of 0.045, $\chi^2(1) = 5.68$, $p = .017$, and 0.245, $\chi^2(1) = 9.03$, $p = .0038$, respectively, compared with studies utilizing only men.

Model C2 tested interactive effects between study origin and sex, between instruments and sex for differentiated effect sizes. Including these interactions in Model C2 significantly improved the goodness of model fit over Model C1, $\chi^2(7) = 179.34$, $p < .0001$, as shown in Table 4.

Compared with Model C1, in Model C2 the efficacy of the OGRS for men became significantly larger than that for the PCL-R, $\chi^2(1) = 5.25$, $p = .022$, by a mean of 0.315, whereas for women, the effect size was considerably reduced by −0.81, $\chi^2(1) = 132.9$, $p < .0001$. There was a significantly reduced efficacy of the RM2000V for women, $\chi^2(1) = 13.22$, $p = .003$, and an increased efficacy of the PCL-R/PCL:SV Factor 1 for women, $\chi^2(1) = 14.53$, $p = .0001$. The overall effect size for women in the U.S. studies was significantly lower than that of others by −0.48, $\chi^2(1) = 4.44$, $p = .035$, whereas the effect size for men in the U.S. studies was no different from that in other studies, $\chi^2(1) = 3.69$, $p = .055$. Furthermore, the difference between prospective and retrospective studies now became nonsignificant, $\chi^2(1) = 1.43$, $p = .230$.

Other consistent findings in both models C1 and C2 were as follows: (a) better efficacy of the HCR-20 compared with the PCL-R total, (b) poorer efficacy of PCL-R/PCL:SV Factor 1 (for men) compared with the PCL-R total, and (c) larger effect sizes for

Table 3
Efficacy of Risk Instruments in Predicting Violent Outcomes

Instrument	No. reports	No. participants	Raw effect size (minimum, maximum)	Model C1 estimates (weighted and adjusted)	
				Effect size (95% *CI*)	*AUC* (r_{pb})
PCL-R	16	3,854	0.64 (0.08, 1.15)	0.55 (0.37, 0.74)	0.65 (0.27)
PCL:SV	8	2,506	0.76 (0.47, 1.11)	0.65 (0.40, 0.90)	0.68 (0.31)
PCL-R/PCL:SV Factor 1	13	3,895	0.34 (−0.19, 0.65)	0.22 (0.00, 0.45)	0.56 (0.11)
PCL-R/PCL:SV Factor 2	13	3,995	0.71 (0.32, 1.11)	0.61 (0.38, 0.84)	0.67 (0.30)
OGRS	2	1,955	0.60 (0.14, 0.83)	0.78 (0.45, 1.11)	0.71 (0.36)
RM2000V	3	1,784	0.75 (0.58, 0.97)	0.76 (0.41, 1.11)	0.70 (0.35)
VRAG	17	4,894	0.74 (0.14, 1.13)	0.68 (0.44, 0.92)	0.68 (0.32)
HCR-20	16	4,161	0.85 (0.28, 1.34)	0.79 (0.56, 1.02)	0.71 (0.37)
H10	18	4,725	0.66 (−0.03, 1.11)	0.61 (0.38, 0.84)	0.67 (0.29)
C5	14	4,078	0.64 (−0.11, 1.20)	0.59 (0.40, 0.78)	0.66 (0.29)
R5	12	3,998	0.63 (0.00, 1.13)	0.60 (0.37, 0.83)	0.66 (0.29)
GSIR	3	988	0.81 (0.68, 1.06)	0.67 (0.37, 0.97)	0.68 (0.25)
LSI-R	3	355	0.58 (0.47, 0.69)	0.51 (0.20, 0.82)	0.65 (0.25)
VRS	4	1,148	0.59 (−0.12, 1.10)	0.53 (0.22, 0.83)	0.65 (0.25)
VRS-Static	3	1,098	0.46 (0.08, 0.87)	0.51 (0.21, 0.84)	0.65 (0.25)
VRS-Dynamic	3	1,098	0.49 (−0.15, 0.87)	0.57 (0.27, 0.89)	0.66 (0.28)

Note. PCL-R = Psychopathy Checklist—Revised; PCL:SV = Psychopathy Checklist Screening Version; OGRS = Offender Group Reconviction Scale; RM2000V = Risk Matrix 2000 for Violence; VRAG = Violence Risk Assessment Guide; HCR-20 = Historical, Clinical, and Risk Management Violence Risk Assessment Scheme; H10 = 10-item Historical subscale of the HCR-20; C5 = 5-item Clinical subscale of the HCR-20; R5 = 5-item Risk Management subscale of the HCR-20; GSIR = General Statistical Information for Recidivism; LSI-R = Level of Service Inventory—Revised; VRS = Violence Risk Scale.

female participants (except for a U.S. study) and mixed-sex studies than for male-only studies.

Sex-differentiated predictive efficacy of the risk instruments is presented in Figure 1. The mean effect size and its 95% confidence interval (*CI*) for each of seven instruments and certain subscales were derived from the estimates of Model C2. The OGRS and the RM2000V demonstrated considerably better efficacy in predicting violence for men than for women, whereas PCL-R/PCL:SV Factor 1 had a larger effect size for women than for men. No sex difference was found for PCL-R and PCL:SV total scores, HCR-20, or PCL-R/PCL:SV Factor 2.

Variance of Effect Sizes Explained

Through multilevel models, the total variance of random effects in the effect sizes was decomposed into variance between and within studies. If a variable is known to contribute to a source of variance component, adding such a variable to the model will result in a substantial reduction in variance attributed to that component. Table 5 shows variances of both between and within study in four nested models: A, B3, C2, and C3. Model A is the "empty" model without any covariate effects; Model B3 is the elaborated model with instrument effects only; Model C2 is a further elaboration with both instruments and study features effects; and, finally, Model C3 includes effects of outcome categories in addition to all variables in Model C2. The reduction of 48.1% of the within-study variance in Model B3 compared with Model A was related to differences between instruments. However, 51.9% of within-study variation remained significant and unexplained, $\chi^2(1) = 35.0, p < .0001$. Compared with Model A, the marked 76.6% reduction of between-study variance in Model C2 was mainly related to study features, such as mean age of participants, follow-up time, proportion of women participants, sex-differentiated efficacy between countries and in risk instruments.

However, the study-level variance of Model C2 was still significantly larger than zero ($p < .05$), which could relate to the use of different criterion measures by different studies. To test our hypothesis, we added the outcome criterion category as another moderator in Model C2, to form Model C3. With four outcome categories, three variables were entered into the model to estimate differences in effect sizes between violent official criminal recidivism, excluding sexual offenses and robbery (the reference category) and (a) physical aggression within an institution; (b) actual, attempted, or threat of harm to others (as defined by HCR-20); and (c) broadly defined violent official criminal recidivism, including sexual offense and robbery. The new model, Model C3, estimated a moderately larger effect size from studies with the broadly defined violence by a mean 0.239, $\chi^2(1) = 4.64$, $p < .05$, than that of the reference category but no difference among the other three. The between-study variance was reduced further by 31.9% to 0.0068 compared with that of Model C2, indicating the absence of any study differences, $\chi^2(1) = 3.27$, $p = .071$. All other significant findings in Model C2 remained unchanged.

Considering the impact of heterogeneity of study on effect sizes reported in literatures, we compared effect sizes of studies between models without and with adjustment of study features in Figure 2. Without taking into account heterogeneity among studies, Model B1 yielded a wide range of effect sizes across studies, from 0.08 to 1.03, with an overall mean of 0.64 (AUC = 0.67). With adjustments in Model C1, the effect sizes of most studies were significantly reduced, with a mean estimate of 0.55 (AUC = 0.65). Overlapping confidence intervals of the estimates in the studies indicated no substantive differences.

Discussion

The critical importance of violence assessment, prediction, and reduction in forensic mental health and criminal justice practices has resulted in the rapid research and development of violence

Table 4
Effect Sizes of Risk Instruments and Associations of Study Features to Predict Violent Behavior

Instrument	Model C1 Estimate (*SE*)	Model C2 Estimate (*SE*)
Overall	0.554 (0.094)	0.629 (0.097)
Instrument: PCL-R	Reference	Reference
OGRS	0.231 (0.139)	0.315 (0.138)*
VRAG	0.123 (0.071)	0.116 (0.071)
RM2000V	0.204 (0.149)	0.245 (0.147)
HCR-20	0.240 (0.068)***	0.249 (0.068)***
H10	0.054 (0.068)	0.061 (0.068)
C5	0.032 (0.072)	0.040 (0.071)
R5	0.043 (0.073)	0.050 (0.073)
PCL:SV	0.095 (0.087)	0.094 (0.086)
PCL-R/PCL:SV Factor 1	−.335 (0.074)***	−0.341 (0.073)***
PCL-R/PCL:SV Factor 2	0.061 (0.073)	0.067 (0.073)
LSI-R	−.045 (0.129)	−.063 (0.128)
GSIR	0.120 (0.121)	0.106 (0.120)
VRS	−.021 (0.116)	−.002 (0.115)
VRS-Static	−.040 (0.127)	−.022 (0.126)
VRS-Dynamic	0.019 (0.127)	0.038 (0.126)
Others[a]	−.425 (0.228)	−.381 (0.221)
Country		
Sweden vs. Canada	−.015 (0.119)	−.025 (0.111)
United Kingdom vs. Canada	−.172 (0.086)	−.167 (0.081)
United States vs. Canada	−.513 (0.112)***	−.299 (0.156)
Holland vs. Canada	0.114 (0.155)	0.058 (0.146)
Germany vs. Canada	−.490 (0.201)*	−.361 (0.198)
Study type		
Prospective vs. retrospective	0.156 (0.071)*	0.094 (0.078)
Type of participants		
Psychiatric patients vs. prisoners	0.022 (0.097)	0.017 (0.096)
Mixed vs. prisoners	−.132 (0.109)	−.086 (0.102)
Participant gender		
Women only vs. men only	0.045 (0.019)*	0.108 (0.027)***
Mixed gender vs. men only	0.245 (0.081)***	0.192 (0.081)*
Mean age of participants	−.003 (0.005)	−.013 (0.005)**
Mean time at risk (months)	0.0028 (0.0011)**	0.0020 (0.001)*
Women-only study in United States		−.476 (0.226)*
OGRS for Women		−.807 (0.070)***
PCL:SV for Women		0.082 (0.136)
PCL-R/PCL:SV Factor 1 for Women		0.240 (0.063)***
PCL-R/PCL:SV Factor 2 for Women		0.025 (0.063)
RM2000V for Women		−.251 (0.069)***
HCR-20 for Women		−.011 (0.061)
Level of variance		
Between study	0.0132 (0.0056)*	0.0104 (0.0048)*
Within study	0.0257 (0.0043)***	0.0253 (0.0042)***
−2 log-likelihood	170.85	−8.50
χ^2 for goodness of fit	30.55** (Model B3 vs. C1)	179.35*** (Model C1 vs. C2)

Note. $N = 174$. PCL-R = Psychopathy Checklist—Revised; OGRS = Offender Group Reconviction Scale; VRAG = Violence Risk Assessment Guide; RM2000V = Risk Matrix 2000 for Violence; HCR-20 = Historical, Clinical, and Risk Management Violence Risk Assessment Scheme; H10 = 10-item Historical subscale of the HCR-20; C5 = 5-item Clinical subscale of the HCR-2-; R5 = 5-item Risk Management subscale of the HCR-20; PCL:SV = Psychopathy Checklist: Screening Version; LSI-R = Level of Service Inventory—Revised; GSIR = General Statistical Information for Recidivism; VRS = Violence Risk Scale.
[a] Others included the Sexual Violence Risk–20 and the Static 99.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

prediction methodologies. Clinicians and researchers now have available to them an assortment of well-constructed and well-validated tools that purport to assess and predict violence to various degrees. Which tool or tools can provide the most accurate prediction of violence is an important theoretical and practical question. Recent attempts to answer this question by way of meta-analytic reviews of the literature have produced inconsistent results, in part because of various methodological issues. In the present study, we attempted to answer the question of which tool can best predict violence by comparing the predictive efficacy of nine commonly used risk assessment tools with multilevel regression models based on a within-study design that addressed many

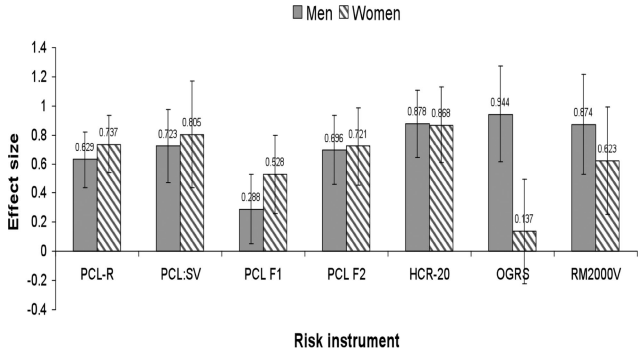

Figure 1. Effect sizes and 95% confidence intervals by gender from Model C2. PCL-R = Psychopathy Checklist—Revised; PCL:SV = Psychopathy Checklist: Screening Version; PCL F1 = Psychopathy Checklist—Revised and Psychopathy Checklist: Screening Version Factor 1; PCL F2 = Psychopathy Checklist—Revised and Psychopathy Checklist: Screening Version Factor 2; HCR-20 = Historical–Clinical–Risk Management–20; OGRS = Offender Group Reconviction Scale; RM2000V = Risk Matrix 2000 for Violence.

key methodological issues. Overall, our results showed that all of the nine tools predicted violence at above-chance levels, with medium effect sizes, and no one tool predicting violence significantly better than any other. In sum, all did well, but none came first.

Comparison of the Predictive Efficacies of the Tools and Subscales

Only the OGRS (when applied to men) and the HCR-20 were found to predict significantly better than the PCL-R; all other instruments predicted better than chance at about a medium level of efficacy (AUC range from .65 to .70). PCL-R/PCL:SV Factor 1 was significantly worse compared with total PCL-R scores. We discuss the tools individually and then the implications for violence risk prediction, assessment, and management.

OGRS. For the OGRS, both the construction and the validation/prediction samples consisted of United Kingdom prisoners. It is to be expected that the OGRS would enjoy some predictive advantage because of the similarity of the two samples. Schwalbe's (2007) meta-analyses also found similar effects. Until the predictive efficacy of the OGRS can be compared with samples different to its construction sample, it is premature to conclude that such a predictive advantage will generalize.

HCR-20. Consistent with the literature, we also found the HCR-20 predicted violence better than the PCL-R/PCL:SV. However, PCL-R/PCL:SV scores are used to rate one of its 20 items and are thus already embedded in the HCR-20. The additional Historical, Clinical, and Risk Management variables in the tool would be expected to improve on violence prediction. Removal of the psychopathy item in the HCR-20 may remove the prediction advantage of the HCR-20 over the PCL-R, and this was indeed shown to be the case by de Vogel et al. (2004). We were not able to disentangle this confound in our analyses. For research purposes, the total HCR-20 scores are often derived from summing individual HCR-20 item scores, a practice the developers of the HCR-20 specifically cautioned against in the clinical use of the tool (Webster et al., 1997, p. 22). It is, therefore, unclear to what extent the present findings, based entirely on summing of the scores, could be generalized to the clinical use of the tools. For the above reasons, it is premature to conclude that the HCR-20 predicted violence better than the PCL-R. We also found that each of the three HCR-20 subscales demonstrated similar predictive effects compared with other risk instruments. The three subscales also appeared to have a synergistic effect: The overall predictive efficacy appeared higher when the subscales were combined, which is the way the tool was developed. As the present results indicate, this is how it should be used.

PCL-R, PCL-SV, Factors 1 and 2. The average PCL-R effect size (0.64) was smaller than, but still within, the 95% *CI* of Salekin's meta-analysis and close to Walters's (2003b) findings. The PCL:SV effect size was larger than that for the PCL-R. However, this difference in the effect sizes did not exceed chance after adjusting for study characteristics and other random effects. The effect size (.55, AUC = .65) of the PCL-R is comparable to that of the other tools. However, Factor 1, which assesses the core psychopathic personality features, demonstrated practically no predictive efficacy (effective size = .22, AUC = .56); it was the only scale among the 16 investigated with effect size overlapping with 0. Recent meta-analyses on institutional adjustment and recidivism, on youth recidivism, and among civil psychiatric patients also produced similar findings (Edens et al., 2006; Skeem & Mulvey, 2001; Walters, 2003a, 2003b). Together, these findings

Table 5
Variance of Effect Sizes Estimated and Attributors

Variable	Model A	Model B3	Model C2	Model C3
Variation level				
Between study	0.0445**	0.0495**	0.0104*	0.0068
Within study	0.0478***	0.0248***	0.0253***	0.0253***
		Model B3 vs. A	Model C2 vs. A	Model C3 vs. A
% Reduction of variance				
Between study			76.6	84.7
Within study		48.1	47.1	47.1
Total		19.5	61.3	65.1
Attributes to variance reduction		Differences in instruments	Differences in instruments and study features	Differences in instruments, study features, and outcome criterion

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

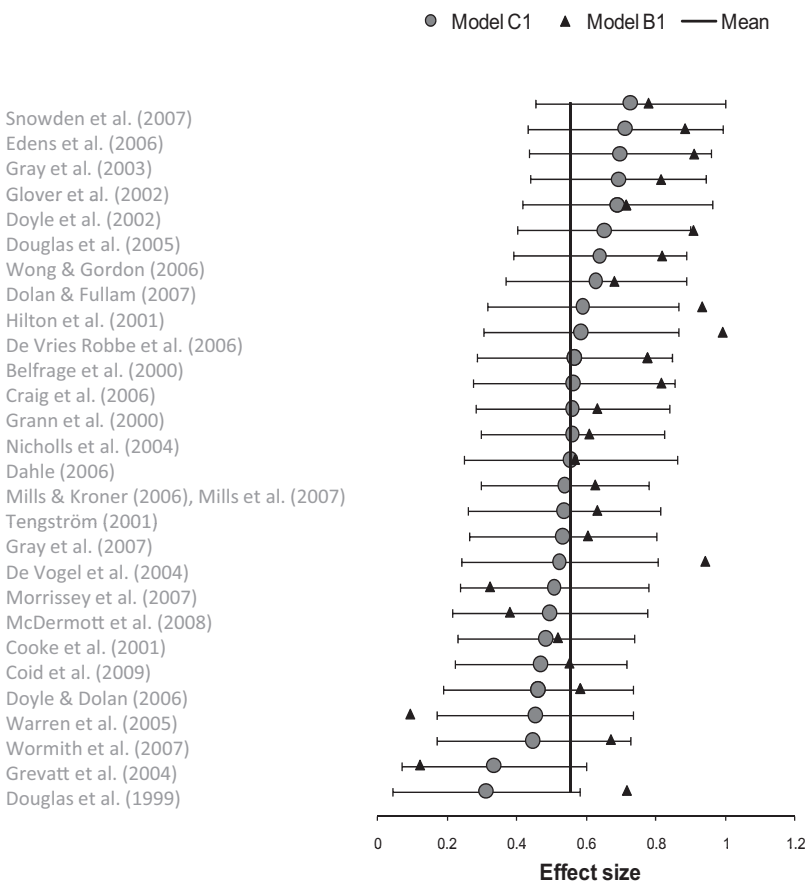

Figure 2. Effect sizes of studies with 95% confidence intervals estimated from Model C.

suggest that Factor 1 personality features, the core personality features of psychopathy, are not linked to violence. The predictive efficacy of the PCL-R appeared to be attributable almost entirely to Factor 2 (effect size of .61, AUC = .67), which is essentially a measure of previous criminality and antisocial behavior, such as impulsivity, criminal versatility, and irresponsibility (often termed *criminogenic characteristics*). Previous violence and criminality are powerful predictors of future violence and criminality, which may explain why the predictive efficacy of Factor 2 is similar to the RM2000V and OGRS, tools that also rely heavily on past criminality to predict violence.

VRAG. We found an effect size of 0.68 for the VRAG based on 4,894 participants in 17 studies, which is comparable to the AUC value in a recent meta-analysis of 14 effect sizes (Campbell et al., 2009) but smaller than that found in the construction sample (AUC of 0.76; Harris et al., 1993). The mean follow-up for the construction study was 6.80 years compared with only 3.07 years in this meta-analysis; our finding may therefore not be unexpected given that we found larger effect sizes with longer follow-up time.

Past studies comparing predictive effects between the VRAG and the PCL-R or PCL:SV were either inconsistent (Campbell et al., 2009; Coid et al 2009; Douglas, Yeomans, & Boer, 2005; Glover et al., 2002; Hilton, Harris, & Rice, 2001; Loza & Loza-Fanous, 2001; Mills & Kroner, 2006) or reported higher effect sizes of the PCL-R/PCL:SV than of the VRAG (Doyle & Dolan 2006; Kroner & Loza, 2001; Kroner & Mills, 2001; Loza & Green, 2003). The present results indicated that any difference in effect sizes between the two measures was due to chance after adjusting for study characteristics and correlations between instruments.

LSI-R. The effect size of the LSI-R in the present study (AUC = 0.65) was identical to that from a previous large-scale meta-analysis (Gendreau et al., 2002) and close to that from a meta-analysis by Walters (2003a; AUC = 0.67) in predicting institutional adjustment and recidivism (see also a recent meta-analysis by Campbell et al., 2009, which reported an AUC of 0.61). Our finding of similar predictive effects between the LSI-R and the PCL-R is consistent with previous findings.

Risk Matrix 2000 for Violence, OGRS, and sex effects. Whereas there was no difference for either the RM2000 V or the OGRS compared with the PCL-R for the combined sample of men and women (Model C1), we showed for the first time that predictive efficacies for both tools were significantly better for men than for women when men and women were considered separately (Model C2). The sex effect may be due to the fact that both tools were developed with male offenders in mind. For example, predictors such as offense history, which is a good risk predictor for men in the United Kingdom, and was selected for that purpose, did not perform as well for women in the United Kingdom (Coid et al., 2009). Furthermore, female participants in this study had a significantly higher prevalence of Axis I clinical syndromes, such as

affective disorder, psychotic illness, and substance use dependence (Coid et al., 2009); similar variables demonstrated smaller predictive ability with the HCR-20 among women compared with men. The sex-differential effects of the two instruments in predicting violence in the present study, with United Kingdom prison samples only, require further research with non–United Kingdom offender samples.

VRS. The VRS, which has both static and dynamic factors in separate domains, allows for within-subject comparison of the predictive efficacies of these domains. In contrast with static predictors, dynamic predictors are useful in guiding treatment intervention by identifying treatment targets linked to violence and measuring treatment change. The VRS and the PCL-R have similar predictive efficacies (effect sizes of .53 and .55, respectively), whereas the VRS dynamic domain performed slightly, but not significantly, better than the static domain (effect size of .57 vs. .51 respectively). The results highlight that static and dynamic predictors appeared to perform equally well in predicting violence recidivism. However, the clinical usefulness of dynamic variables outweighs the static ones in risk reduction treatment and management of forensic clients.

Conclusions and Implications

The HCR-20 and the OGRS showed statistically significantly larger effect sizes than the PCL-R, but such findings are tentative at best and did not exceed the other instruments by a large amount. This level of difference, even if replicated, is not likely to be translated into a meaningful level of difference in clinical practice. In contrast, a recent meta-analysis that compared the efficacy of five risk assessment tools (the HCR-20, the LSI-R, the PCL-R, the GSIR, and the VRAG) in predicting violence recidivism revealed that the HCR-20 and the PCL-R had similar predictive efficacies (overlapping 95% sample-adjusted *CI*s), whereas the VRAG performed better than the HCR-20 (with nonoverlapping but a small separation in their *CI*s). However, the *CI* of the VRAG overlapped with that of the LSI-R and PCL-R (Campbell et al., 2009). In essence, when differences in predictive efficacies for violence between instruments are found in different meta-analytic studies, they are usually not large and are inconsistent.

In all, based on our overall findings and the literature, such as Campbell et al. (2009), we conclude that there is no appreciable or clinically significant difference in the violence-predictive efficacies of the nine tools after accounting for differences in study features and other unexplained random effects with multilevel regression analysis. If prediction of violence is the only criterion for the selection of a risk assessment tool, then the tools included in the present study are essentially interchangeable. It would follow that the choice of using any one of the nine tools over another should be based largely on what additional relevant clinical, criminal justice, or management functions the tool of choice can perform, rather than on how well it can predict violence in comparison with other tools. Furthermore, predictive efficacy is essentially very similar when we contrasted second-generation (e.g., VRAG) with third-generation (e.g., LSI-R) tools, theoretically derived (e.g. PCL-R) with empirically derived tools (e.g., GSIR), or tools consisting of only static/unchangeable predictors (e.g., RM2000) with tools with both static and dynamic/changeable predictors (e.g., VRS). The instruments' confidence intervals overlap to such an extent that it is not possible to say that any one tool predicts violence consistently and significantly better than any others.

We are not implying that all of the tools are equivalent in all respects; different tools are designed for different functions in addition to risk prediction. Tools with dynamic risk predictors can assess change in risk (see Olver et al., 2007) while those with static predictors cannot. The PCL-R was designed for assessing a personality construct, whereas the LSI Case Management Inventory can inform on case management processes and the VRS, on treatment readiness and change. The knowledgeable assessor needs to select the appropriate tool from his or her toolbox for the purpose at hand. The nine tools have similar efficacy in violence predictions, but they have other important differences.

Despite the many conceptual and theoretical differences in the tools, why are they so similar when it comes to predicting violence? We can only speculate, but we posit first that, for the purpose of making violence predictions, the risk factors in the different tools could have been drawing from the same pools of variance that reflect a long-standing pattern of dysfunctional and aggressive interpersonal interactions and antisocial and unstable lifestyle that are common to many perpetrators of violence. The risk factors are probably different labels we use to tap into these common variances. The results of the study by Kroner, Mills, and Reddon (2005), which revealed that risk factors in many tools are essentially interchangeable, nicely illustrates this point.

After almost five decades of developing risk prediction tools, the evidence increasingly suggests that the ceiling of predictive efficacy may have been reached with the available technology. Other approaches such as tree modeling (Steadman et al., 2000) and Neural Networks (Price et al., 2000) require further exploration, but it is unlikely that a very high level of predictive accuracy is achievable because of theoretical constraints. Violent behavior is the result of the individual interacting with the immediate environment. Although it may be possible to improve on our understanding and predicting what an individual may do in hypothetical situations, it will be much more difficult to predict the situation that an individual actually encounters in the open community. Even predicting violence within an institutional environment is difficult, where the assessor has much more information about that environment.

From Risk Assessment to Risk Management

Building a better model of violence prediction should not be the sole aim of risk prediction research, which is just one link in the risk assessment–prediction–management triad that aims to achieve violence reduction and improved mental health. Risk management could be achieved by providing better treatment and continuity of care, but it must rely on good risk assessment. The risk, need and responsivity principles derived from the theory of the psychology of criminal conduct (see Andrews & Bonta, 1998, 2003, 2006, 2010; Andrew et al., 1990) provide a useful theoretical framework for risk reduction intervention. Appropriate risk assessment can identify high-risk individuals in need of more intensive management and intervention, by means of the risk principle. Using tools with dynamic risk predictors to assess risk can identify appropriate changeable treatment targets linked to violence (the need principle) in particular for treatment-resistant clients who require more specialized intervention (the responsivity principle). Assessment tools with dynamic or changeable

predictors, such as the HCR-20, the VRS, and the LSI-R can accomplish some of these tasks provided that the dynamic predictors are, in fact, causal predictors according to criteria set forth by Kraemer et al. (1997). A causal risk predictor is one that can be manipulated and, when it is manipulated, results in corresponding changes in the outcome measures (Kraemer et al., 1997). For example, criminal attitude is a causal risk predictor if reduction in criminal attitude with intervention in a treatment program could be linked to reduction in recidivism.

Prediction research, as typically undertaken, with tools and correlational methodologies illustrated in the present review, can elucidate the links between two variables, but it cannot establish the causal nexus between them: Correlations do not imply causation (see Arboleda-Florez & Stuart, 2000; Kraemer et al., 1997; Mullen, 2000). Risk management and violence reduction interventions require the clear understanding of causation (see Buchanan, 2008; Douglas & Skeem, 2005; Wong & Gordon, 2006); we can only intervene with confidence if we know that A causes B and that reducing A would lead to reducing B. Prediction research has identified many potential causes of violence, such as substance abuse, acute mental disorder, and criminal lifestyle. However, research is only scratching the surface of identifying causal predictors (see Andrews, Bonta, & Wormith, 2009; Hanson, Harris, Scott, & Helmus, 2007; Hudson, Wales, Bakker, & Ward, 2002; Olver & Wong, 2009; Olver et al., 2007). Understanding the causal relationships should also sharpen our predictive power. Much more research is required to identify causal risk predictors.

Our finding that Factor 1 interpersonal and affective traits of psychopathy are not linked to future violence can have important clinical and treatment implications. Treatment interventions that focus on changing these core psychopathy traits, based on the previous findings, will not have any significant impact on reducing future violence in men, even if the treatment is successful and the psychopathic traits are substantially modified. For example, The Dangerous and Severe Personality Disordered treatment program, established about ten years ago in the United Kingdom, aims at treating individuals who are dangerous or at high risk for violence and have one or more severe personality disorders, such as psychopathy, that are functionally linked to violence (Maden & Tyrer, 2003). The present results suggest that in such treatment programs, reducing the risk of violence should focus on reducing criminogenic factors rather than on reducing the core psychopathic traits.

To reduce propensity for violence among psychopathic individuals, treatment must target causal links to violence or criminogenic characteristics, such as Factor 2 characteristics, with "what works" approaches (Wong, Gordon, & Gu, 2007; Wong & Hare, 2005). However, Factor 1 core personality traits are still important clinical considerations because they interfere with treatment delivery as a result of conning, manipulative characteristics, lack of responsibility for actions, and low motivation to change. In addition, affective deficits and interpersonally exploitative behaviors could be significant impediments to the formation of a functional working alliance (Wong & Hare, 2005, p. 20). However, these are responsivity issues rather than criminogenic factors, and such responsivity issues must be appropriately managed in order for treatment to proceed.

There are a number of caveats here: lack of predictive effect of PCL-R Factor 1 for violent risk was only observed among men. Factor 1 has a small but significant effect size for women even after adjusting for study features; however, more research is required to validate these findings. Although Factor 1 did not appear to have direct links to future violence, it could interact with other risk factors, such as sexual deviance, to increase the risk of sexual violence in moderate- to high-risk sex offenders (see Hildebrand, de Ruiter, & de Vogel, 2004; Olver & Wong, 2006). Such possibilities were not tested in this meta-analysis. The present findings also point to the need to further assess the violence-predictive efficacy of Factor 1 and its derivatives (Facet 1 and Facet 2; see Hare, 2003) within the four- and the three-factor structure of the construct of psychopathy.

The Unpacking of Study Heterogeneity

It is widely accepted in meta-analysis that study heterogeneity originating from differences in study settings can be controlled for, but similar heterogeneity that originates from other sources may not be measurable or controlled for. It is convenient to use the term *random effects* to include all sources of differences attributable to heterogeneity without clearly identifying the specific attributes. Most researchers nowadays routinely apply Q statistic to test for overall random effects between studies and use weighted mean effect size to adjust for them (e.g., see Edens et al., 2006; Guy, Edens, Anthony, & Douglas, 2005; Walters, 2003a). In contrast to previous reports, the present study disentangled the total variation in effect sizes into two components: between-study variation or heterogeneity, accounting for 48%, and within-study variation, accounting for 52% of the total variance (Model A). Of the between-study heterogeneity, 85% was attributable to the age of participants, follow-up time, sex, sex–country interaction, sex–tool interaction, and outcome criteria, whereas 47% of within-study variation was attributable to instrument differences (Model C3). In sum, only about 25% of the total variance was attributable to instrument differences. Using a different type of regression analysis, a previous meta-analysis of predictive efficacy of risk tools for juvenile recidivism also showed that only 17% of the total variance in the AUC values was accounted for by type of risk tools, whereas 42% of the total variance was contributed by several methodological moderators (Schwalbe, 2007). Moderator effects in effect sizes of certain risk instruments were previously examined in some meta-analytic studies (Campbell et al., 2009; Guy et al., 2005; Walters, 2006) for different outcomes. The lack of significant effects of most moderators reported in those studies could be due to limitations of standard statistical procedures described in the early sections of this report.

By applying multilevel regression analysis that combines multivariate regression model and a random-effects model into a single model to preserve the maximum statistical power afforded by the data, we uncovered significant mean and differential effects of key moderators that were major sources of study heterogeneity. After controlling for these sources of study heterogeneity by explicitly modeling the effects of moderators, we were able to compare the predictive efficacies of risk instruments in a more effective and less biased manner based on homogeneous study samples. In essence, we created a statistically level playing field on which to compare the risk tools, a strategy that has many obvious advantages.

That age and follow-up time are significant moderators is not surprising. The association of increase age with a decrease in

prevalence in offending—the age–crime curve—is a well-established finding in criminology (Hirschi & Gottfredson, 1983). Violent offending occurs much less frequently than nonviolent offending. For example, the offense histories in a sample of over 900 Canadian federal offenders in their mid-30s showed that nonviolent convictions were more than five times more likely than violent convictions (Wong & Gordon, 2006). As such, longer follow-up time is expected to be associated with larger effect size as it takes more time to accumulate a substantial number of violent infractions for assessing predictive efficacy.

The variation of predictive efficacy for women in terms of instruments, country, and clinical characteristics (Factor 1) is complex. Needless to say, more studies are required to unravel these relationships, and it is important not to overinterpret the present results, as they are based on relatively few studies from a limited number of countries.

Predictive efficacy of risk assessment instruments may differ depending on the use of different outcome criteria. We found predictions of the broadly defined criterion of violence to have larger effect sizes than those of other three categories of violent outcome, contributing 8.2% to study heterogeneity and independent of the effects of other study features or moderators. A previous meta-analysis (Campbell et al., 2009) also found differences in predictive efficacy for institutional violence compared with violence recidivism. Most studies that used the broadly defined criterion of violence drew from samples of Canadian prisoners, and seven out of nine tools we compared were developed in Canada using Canadian forensic samples. It is possible that the larger effect size of the one outcome criterion could be associated with the similarity of the study samples with the construction samples on which the tools were developed. We did not model differential effects between outcome category and individual risk instruments, as such analyses might cause overfitting of our model. Further research is required to assess the replicability of the findings and the validity of our hypothesis.

Unpacking study heterogeneity with multilevel regression analyses has important implications. In validating risk assessment tools, one must take into account, either in the study design or in the statistical analyses, the various potential sources of heterogeneity.

Limitations

First, the literature search may not have included all published and unpublished papers that met our inclusion criteria, and some systematic biases may be introduced into article selection. However, these biases were minimized by using two persons to select and review the articles.

Second, a range of outcomes were used as criterion variables in the reviewed studies, and prediction efficacies vary with types of outcome Violence also varied in quality (type of violence), severity (harm inflicted), and frequency of occurrence (base rate). To truly compare the predictive efficacy of the tools, one needs to equate the outcomes or criterion variables of the predictions. Most, if not all, of the studies reviewed used prediction of the first occurrence of violence rather than prediction of a pattern of violence as the criterion variable; the latter has just as much, if not more, relevance to violence prediction, management, and reduction. The present study, as in other meta-analytic studies, is inevitably limited by the criterion reported in the studies.

A caveat that is common to many meta-analyses is that there is no control over the quality of the study and the data, nor proprietary interests; no study was excluded on the basis of quality considerations in the present analyses. We did not code for study quality, although some meta-analysts do so, and we did not code for proprietary interests. We also did not investigate the "operation" of the subscales within the mother tool (such as that of the HCR-20, the PCL-R, and the VRS), as doing so would involve major factor- analytic studies that were beyond the scope of this analysis. Findings on lower effect sizes of predictive efficacy in studies on U.S. women must be considered tentative as a result of the small numbers of studies included in the analysis.

Recommendations

On the basis of the results of present meta-analyses and review of the literature, we put forth the following recommendations.

1. All risk assessment instruments (excluding subscales) included in the study predicted violent recidivism moderately well, and their predictive efficacies were not significantly different. Because of their moderate level of predictive efficacy, they should not be used as the sole or primary means for clinical or criminal justice decision making that is contingent on a high level of predictive accuracy, such as preventive detention.

2. The selection of a tool for clinical or research purposes should be determined more by what other functions the tool can perform than its violence prediction efficacy per se.

3. Efforts should be directed toward investigating situational contingencies that precipitate violence. Little research has been carried out in this area, in contrast to individual variability.

4. The efficacy in identifying risk predictors and extracting prediction information from them based on the current methodology of summing ratings of predictors (exemplified by all the tools under study) may have reached a plateau. Future research should explore other novel means of identifying and combining risk predictors, for example, the tree method and neural network approaches, including all aspects of the risk assessment process, such as different categorizations of violent offender groups, criteria of violence, and additional situational or dynamic predictors that might be specific for violent prediction (Yang, Liu, & Coid, 2010).

5. More research should be carried out to identify causal predictors of violence to inform violence reduction interventions and to improve the accuracy of prediction.

6. The present results suggest that Factor 2 rather than Factor 1 of the PCL-R predict violence. It is hypothesized that when intervening to reduce violence among psychopathic individuals, efforts directed at changing Factor 2 (criminality) characteristics should be more effective than those directed at Factor 1 (personality) characteristics. Future research should test this hypothesis directly.

7. More studies of violence prediction should be undertaken with female participants as the pattern of prediction results for women appeared significantly different, in many instances, from those of men. Most prediction tools have been developed for use with men.

8. A common metric to assess different dimensions of violence, such as quality, severity, and frequency should be developed to facilitate between-study comparisons of the criterion variables.

9. Multilevel regression model analysis may be the preferred tool for meta-analysis where common methodological issues include (a) presence of random effects in effect size due to heterogeneity among studies, (b) lack of statistical power to draw meaningful conclusions due to small sample size, and (c) the need to adjust for characteristics of studies in order to estimate the pooled effect size. The nature of multilevel models in handling data with clustering effects and dependency also opens the door for meta-analysts to estimate effect sizes based on studies reporting efficacy measures at different follow-up times within study, effect sizes of correlated multiple outcomes, or effect sizes based on studies with individual data.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Adams, J. (1995). *Risk.* London: UCL Press.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Andrews, D. A. (1995). The psychology of criminal conduct and effective treatment. In J. McGuire (Ed.), *What works: Reducing reoffending, guidelines from research and practice* (pp. 3–34). Chichester, United Kingdom: Wiley.

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

Andrews, D. A., & Bonta, J. (1998). *The psychology of criminal conduct* (2nd ed.). Cincinnati, OH: Anderson.

Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson.

Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct* (4th ed.). Cincinnati, OH: Anderson.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Cincinnati, OH: Anderson.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19–52.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52,* 7–22. doi:10.1177/0011128705281756

Andrews, D. A., Bonta, J., & Wormith, J. S. (2009). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 199–225). Routledge, NY: Springer.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28,* 369–404. doi:10.1111/j.1745-9125.1990.tb01330.x

Arboleda-Florez, J., & Stuart, H. (2000). The future for risk research. *Journal of Forensic Psychiatry, 11,* 506–509. doi:10.1080/09585180010002696

*Belfrage, H., Fransson, G., & Strand, S. (2000). Prediction of violence using the HCR-20: A prospective study in two maximum-security correctional institutions. *Journal of Forensic Psychiatry, 11,* 167–175. doi:10.1080/095851800362445

Boer, D., Hart, S., Kropp, P., & Webster, D. R. (1998). *Manual for the Sexual Violence Risk–20: Professional guidelines for assessing risk of sexual violence.* Lutz, FL: Psychological Assessment Resources.

Bonta, J., Harman, W. G., Hann, R. G., & Cormier, R. B. (1996). The prediction of recidivism among federally sentenced offenders: A revalidation of the SIR scale. *Canadian Journal of Criminology, 38,* 61–79. Retrieved from http://www.ccja-acjp.ca/en/cjc.html

Borum, R. (1996). Improving the clinical practice of violence risk assessment: Technology, guidelines, and training. *American Psychologist, 51,* 945–956. doi:10.1037/0003-066X.51.9.945; PMID:8819363

Buchanan, A. (2008). Risk of violence by psychiatric patients: Beyond the "actuarial versus clinical" assessment debate. *Psychiatric Services, 59,* 184–190. doi:10.1176/appi.ps.59.2.184

Campbell, M., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36,* 567–590. doi:10.1177/0093854809333610

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

Cleckley, H. (1976). *The mask of sanity* (5th ed.). St Louis, MO: Mosby.

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

*Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., Farrington, D., & Rogers, R. D. (2009). Gender differences in structured risk assessment: Comparison of the accuracy of five instruments. *Journal of Consulting and Clinical Psychology, 77,* 337–348. doi:10.1037/a0015155; PMID:19309193

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., Ramey, S. L., Shure, M. B., & Long, B. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist, 48,* 1013–1022. doi:10.1037/0003-066X.48.10.1013; PMID:8256874

Cooke, D. J., & Michie, C. (2001). Refining the concept of psychopathy: Towards a hierarchical model. *Psychological Assessment, 13,* 171–188. doi:10.1037/1040-3590.13.2.171; PMID:11433793

*Cooke, D. J., Michie, C., & Ryan, J. (2002). Evaluating risk for violence: A preliminary study of the HCR-20, PCL-R and VRAG in a Scottish prison sample (Scottish Prison Service Occasional Papers No. 5/2001). Edinburgh, Scotland: Scottish Prison Service.

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

*Craig, L. A., Beech, A., & Browne, K. D. (2006). Cross-validation of the Risk Matrix 2000 Sexual and Violent scales. *Journal of Interpersonal Violence, 21,* 612–633. doi:10.1177/0886260506286876; PMID:16574636

D'Silva, K., Duggan, C., & McCarthy, L. (2004). Does treatment really make psychopaths worse? A review of the evidence. *Journal of Personality Disorders, 18,* 163–177.

*Dahle, K. P. (2006). Strengths and limitations of actuarial prediction of criminal reoffence in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. *International Journal of Law and Psychiatry, 29,* 431–442. doi:10.1016/j.ijlp.2006.03.001; PMID:16780950

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. doi:10.1126/science.2648573; PMidID:2648573

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

*De Vries Robbe, M., Weenink, A., & de Vogel, V. (2006, June). *Dynamic risk assessment: A pilot study comparing the VRS to the HCR-20.* Paper presented at the meeting of the International Association of Forensic Mental Health Services, Amsterdam, the Netherlands.

Dearwater, S. R., Coben, J. H., Campbell, J. C., Nah, G., Glass, N., McLoughlin, E., . . . Bekemeier, B. (1998). Prevalence of intimate

partner abuse in women treated at community hospital emergency departments. *Journal of the American Medical Association, 280,* 433–438. doi:10.1001/jama.280.5.433; PMID:9701078

*Dolan, M., & Fullam, R. (2007). The validity of the Violence Risk Scale second edition (VRS-2) in a British forensic inpatient sample. *Journal of Forensic Psychiatry and Psychology, 18,* 381–393. doi:10.1080/14789940701489390: PMCID:1443148

Douglas, K. S., & Reeves, K. A. (2009). Historical–Clinical–Risk Management–20 (HCR-20) violence risk assessment scheme: Rationale, application, and empirical overview. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 147–185). New York: Routledge.

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

*Douglas, K. S., Ogloff, J. R. P., Nicholls, T. L., & Grant, I. (1999). Assessing risk for violence among psychiatric patients: The HCR-20 violence risk assessment scheme and the Psychopathy Checklist: Screening version. *Journal of Consulting and Clinical Psychology, 67,* 917–930. doi:10.1037/0022-006X.67.6.917; PMID:10596513

*Douglas, K. S., Yeomans, M., & Boer, D. P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior, 32,* 479–510. doi:10.1177/0093854805278411

*Doyle, M., & Dolan, M. (2006). Predicting community violence from patients discharged from mental health services. *British Journal of Psychiatry, 189,* 520–526. doi:10.1192/bjp.bp.105.021204; PMID:17139036

*Doyle, M., Dolan, M., & McGovern, J. (2002). The validity of North American risk assessment tools in predicting in-patient violent behavior in England. *Legal and Criminological Psychology, 7,* 141–154. doi:10.1348/135532502760274756

Edens, J. F., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior, 31,* 53–75. doi:10.1007/s10979-006-9019-y; PMID:17019617

Edens, J. F., Poythress, N. G., & Lilienfeld, G. O. (1999). Identifying inmates at risk for disciplinary infractions: A comparison of two measures of psychopathy. *Behavioral Sciences and the Law, 17,* 435–443. doi:10.1002/(SICI)1099-0798(199910/12)17:4<435::AID-BSL356>3.0.CO;2-Z

*Edens, J. F., Skeem, J. L., & Douglas, K. S. (2006). Incremental validity analyses of the Violence Risk Appraisal Guide and the Psychopathy Checklist: Screening version in a civil psychiatric sample. *Assessment, 13,* 368–374. doi:10.1177/1073191105284001; PMID:16880286

Farrington, D. P., Ohlin, L., & Wilson, J. Q. (1986). *Understanding and controlling crime.* New York: Springer Verlag.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the "unparalleled" measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29,* 397–426. doi:10.1177/0093854802029004004

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works. *Criminology, 34,* 557–607. doi:10.1111/j.1745-9125.1996.tb01220.x

*Glover, A. J. J., Nicholson, D. E., Hemmati, T., Bernfeld, G. A., & Quinsey, V. L. (2002). A comparison of predictors of general and violent recidivism among high-risk federal offenders. *Criminal Justice and Behavior, 29,* 235–249. doi:10.1177/0093854802029003001

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Goldstein, H., Yang, M., Tuner, R. M., Omar, R. Z., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics, 49*(3), 399–412. doi:10.1111/1467-9876.00200

*Grann, M., Belfrage, H., & Tengström, A. (2000). Actuarial assessment of risk for violence—Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice and Behavior, 27,* 97–114. doi:10.1177/0093854800027001006

*Gray, N. S., Fitzgerald, S., Taylor, J., MacCulloch, M. J., & Snowden, R. J. (2007). Predicting future reconviction in offenders with intellectual disabilities: The predictive efficacy of VRAG, PCL-R, and HCR-20. *Psychological Assessment, 19,* 474–479. doi:10.1037/1040-3590.19.4.474; PMID:18085940

*Gray, N. S., Hill, C., McGleish, A., Timmons, D., MacCulloch, M. J., & Snowden, R. J. (2003). Prediction of violence and self-harm in mentally disordered offenders: A prospective study of the efficacy of HCR-20, PCL-R, and psychiatric symptomatology. *Journal of Consulting and Clinical Psychology, 71,* 443–451. doi:10.1037/0022-006X.71.3.443; PMID:12795569

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiology Review, 9,* 1–30. PMID: 3678409

*Grevatt, M., Thomas-Peter, B., & Hughes, G. (2004). Violence, mental disorder and risk assessment: Can structured clinical assessments predict the short-term risk of inpatient violence? *Journal of Forensic Psychiatry and Psychology, 15,* 278–292. doi:10.1080/1478994032000199095

Grisso, T., & Appelbaum, P. S. (1993). Structuring the debate about ethical predictions of future violence. *Law and Human Behavior, 17,* 482–485. doi:10.1007/BF01044381; PMid:11659733

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323. doi:10.1037/1076-8971.2.2.293

Guy, L. S., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting and Psychology, 73,* 1056–1064. doi:10.1037/0022-006X.73.6.1056; PMid:16392979

Haapanen, R. A. (1990). *Selective incapacitation and the serious offender: A longitudinal study of criminal career patterns.* New York: Springer-Verlag.

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66,* 348–362. doi:10.1037/0022-006X.66.2.348; PMID:9583338

Hanson, R. K., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sex offenders on community supervision* (User Report No. 2007–05). Ottawa, Ontario, Canada: Public Safety and Emergency Preparedness Canada.

Hanson, R. K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis* (User Report 2004–02). Ottawa, Ontario, Canada: Public Safety and Emergency Preparedness Canada.

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

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24,* 119–136. doi:10.1023/A:1005482921333; PMID:10693322

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

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

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The Revised Psychopathy Checklist: Reliability and factor structure. *Psychological Assessment, 2,* 338–341. doi:10.1037/1040-3590.2.3.338

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical predic-

tion instrument. *Criminal Justice and Behavior, 20,* 315–335. doi: 10.1177/0093854893020004001

Hart, S., Cox, D., & Hare, R. (1995). *The Hare Psychopathy Checklist: Screening Version (PCL:SV).* Toronto: Multi-Health Systems.

Heilbrun, K., Yasuhara, K., & Shah, S. (2009). Violence risk assessment tools: Overview and clinical analysis. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 1–18). New York: Routledge.

Hildebrand, M., de Ruiter, C., & de Vogel, V. (2004). Psychopathy and sexual deviance in treated rapists: Association with sexual and nonsexual recidivism. *Sexual Abuse: A Journal of Research and Treatment, 16,* 1–24. doi:10.1177/107906320401600101.97; PMID:15017823

*Hilton, N. Z., Harris, G. T., & Rice, M. E. (2001). Predicting violence by serious wife assaulters. *Journal of Interpersonal Violence, 16,* 408–423. doi:10.1177/088626001016005002

Hirschi, T., & Gottfredson, M. (1983). Age and the explanation of crime. *American Journal of Sociology, 89,* 552–584.

Hoge, R., & Andrews, D. (2002). *Youth Level of Service/Case Management Inventory.* Toronto: Multi-Health Systems.

Hudson, S. M., Wales, D. S., Bakker, L., & Ward, T. (2002). Dynamic risk factors: The Kia Marama evaluation. *Sexual Abuse: A Journal of Research and Treatment, 14,* 103–119. doi:10.1177/107906320201400203; PMID:11961886

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–298. doi: 10.1111/1468-2389.00156

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). London: Sage.

Kemshall, H. (2003). *Understanding risk in criminal justice.* Philadelphia, PA: Open University Press.

Kemshall, H., & Maguire, M. (2001). Public protection, partnership and risk penalty in the multiagency risk management of sexual and violent offenders. *Punishment & Society, 3,* 237–264. doi:10.1177/14624740122228311

Kowalski, R. M., Limber, S. P., Patricia, W., & Agatston, P. W. (2007). *Cyber bullying: Bullying in the digital age.* Malden, MA: Blackwell.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk, *Archives of General Psychiatry, 54,* 337–343. PMID:9107150

Kroner, D. G., & Loza, W. (2001). Evidence for the efficacy of self-report in predicting nonviolent and violent criminal recidivism. *Journal of Interpersonal Violence, 16,* 168–177. doi: 10.1177/088626001016002005

Kroner, D. G., & Mills, J. F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior, 28,* 471–489. doi:10.1177/009385480102800405

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry, 28,* 360–374. doi: 10.1016/j.ijlp.2004.01.011; PMID:15936077

Leyland, A., & Goldstein, H. (2001). *Multilevel modelling of health statistics.* New York: Wiley.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 313–345). Thousand Oaks, CA: Sage.

Litwack, T. R. (1993). On the ethics of dangerousness assessments. *Law and Human Behavior, 17,* 479–482. doi:10.1007/BF01044380; PMID: 11659732

Loza, W., & Green, K. (2003). The Self-Appraisal Questionnaire, a self-report measure for predicting recidivism versus clinician-administered measures: A 5-year follow-up study. *Journal of Interpersonal Violence, 18,* 781–797. doi: 10.1177/0886260503253240

Loza, W., & Loza-Fanous, A. (2001). The effectiveness of the self-appraisal questionnaire in predicting offenders' postrelease outcome, a comparison study. *Criminal Justice and Behavior, 28,* 105–121. doi: 10.1177/0093854801028001005

Maden, A. (2007). *Treating violence: A guide to risk management in mental health.* Oxford, United Kingdom: Oxford University Press.

Maden, T., & Tyrer, P. (2003). Dangerous and severe personality disorders: A new personality concept from the United Kingdom. *Journal of Personality Disorders, 17,* 489–496. doi:10.1521/pedi.17.6.489.25356; PMID:14744075

*McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., & Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable- and person-focused analysis. *Law and Human Behavior, 32,* 325–338. doi:10.1007/s10979-007-9094-8; PMID:17597388

McGuire, J. (2008). A review of effective interventions for reducing aggression and violence. *Philosophical Transactions of the Royal Society, 363,* 2483–2622. doi:10.1098/rstb.2008.0035; PMID:18467276; PMCID:2606715

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216. doi:10.1037/h0048070

*Mills, J. F., & Kroner, D. G. (2006). The effect of discordance among violence and general recidivism risk estimates on predictive accuracy. *Criminal Behaviour and Mental Health, 16,* 155–166. doi:10.1002/cbm.623

*Mills, J. F., Kroner, D. G., & Hemmati, T. (2007). The validity of violence risk estimates: An issue of item performance. *Psychological Services, 4,* 1–12.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques.* Beverly Hills, CA: Sage.

Monahan, J., & Steadman, H. (Eds.). (1994). *Violence and mental disorder: Developments in risk assessment.* Chicago: University of Chicago Press.

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

*Morrissey, C., Hogue, T., Mooney, C. A., Johnston, S., Hollin, C., Lindsay, W. R., & Taylor, J. L. (2007). Predictive validity of the PCL-R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry & Psychology, 18,* 1–15. doi:10.1080/08990220601116345; PMCID:1443148

Mullen, P. (2000). Forensic mental health. *British Journal of Psychiatry, 176,* 307–311. doi:10.1192/bjp.176.4.307; PMid:10827876

FBI Academy, National Center for the Analysis of Violent Crime, Critical Incident Response Group FBI Academy. (n.d.). *The school shooter: A threat assessment perspective.* Retrieved March 18, 2009, from http://www.fbi.gov/publications/school/school2.pdf

Nicholls, T. L. (2004). Violence risk assessments with female NCRMD acquittees: Validity of the HCR-20 and PCL:SV. *Dissertation Abstracts International: Section B. Sciences and Engineering, 64,* 8-B.

*Nicholls, T. L., Ogloff, J. R., & Douglas, K. S. (2004). Assessing risk for violence among male and female civil psychiatric patients: The HCR-20, PCL:SV, and VSC. *Behavioral Sciences and the Law, 22,* 127–158. doi:10.1002/bsl.579

Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines.* Ottawa, Ontario, Canada: Supply and Services Canada.

O'Donohue, W., Fisher, J. E., & Hayes, S. C. (2003). *Cognitive behavior therapy: Applying empirically supported techniques in your practice.* Hoboken, NJ: Wiley.

Ogloff, J. R. P. (2006). Psychopathy/antisocial personality disorder conun-

drum. *Australian and New Zealand Journal of Psychiatry, 40,* 519–528. PMID:16756576

Olver, M., & Wong, S. C. P. (2006). Psychopathy, sexual deviance, and recidivism among different types of sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 18,* 65–82. doi:10.1177/107906320601800105; PMID:16763759

Olver, M., & Wong, S. C. P. (2009). Therapeutic responses of psychopathic sexual offenders: Treatment attrition, therapeutic change, and long-term recidivism. *Journal of Consulting and Clinical Psychology, 77,* 328–336. doi:10.1037/a0015001; PMid:19309191

Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale—Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment, 19,* 318–329. doi:10.1037/1040-3590.19.3.318; PMID:17845123

Poythress, N. G., Jr. (1992). Avoiding negligent release: Contemporary clinical and risk management strategies. *American Journal of Psychiatry, 147,* 994–997. Retrieved from http://ajp.psychiatryonline.org/

Price, R. K., Spitznagel, E. L., Downey, T. J., Meyer, D. J., Risk, N. K., & El-Ghazzawy, O. G. (2000). Applying artificial neural network models to clinical decision making. *Psychological Assessment, 12,* 40–51. doi:10.1037/1040-3590.12.1.40; PMID:10752362

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

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., . . . Lewis, T. (2000). *A user's guide to MLwiN.* London: Institute of Education, University of London.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d,* and *r. Law and Human Behavior, 29,* 615–620. doi:10.1007/s10979-005-6832-7; PMID:16254746

Rice, N., Carr-Hill, R., Dixon, P., & Sutton, M. (1998). The influence of households on drinking behaviour: A multilevel analysis. *Social Science and Medicine, 46,* 971–979. doi:10.1016/S0277-9536(97)10017-X

Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist—Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice, 3,* 203–215. doi:10.1111/j.1468-2850.1996.tb00071.x

Sampson, R., Raudenbush, S., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*(5328), 918–924. doi:10.1126/science.277.5328.918; PMID: 9252316

Schwalbe, C. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31,* 449–462. doi:10.1007/s10979-006-9071-7; PMID:17211688

Skeem, J. L., & Mulvey, E. P. (2001). Psychopathy and community violence among civil psychiatric patients: Results from the MacArthur Violence Risk Assessment Study. *Journal of Consulting and Clinical Psychology, 69,* 358–374. doi:10.1037/0022-006X.69.3.358; PMID: 11495166

*Snowden, R. J., Gray, N. S., Taylor, J., & MacCulloch, M. J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine, 37,* 1539–1549. doi:10.1017/S0033291707000876; PMID:17537287

Stadtland, C., Hollweg, M., Kleindienst, N., Dietl, J., Reich, R., & Nedopil, N. (2005). Risk assessment and prediction of violent and sexual recidivism in sex offenders: Long-term predictive validity of four risk assessment instruments. *Journal of Forensic Psychiatry & Psychology, 16,* 92–108. doi:10.1080/1478994042000270247; PMCID:1443148

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., . . . Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24,* 83–100. doi:10.1023/A:1005478820425; PMID: 10693320

*Tengström, A. (2001). Long-term predictive validity of historical factors in two risk assessment instruments in a group of violent offenders with schizophrenia. *Nordic Journal of Psychiatry, 55,* 243–249. doi:10.1080/080394801681019093

Thornton, D. (2007). *Scoring guide for Risk Matrix 2000.9/SVC.* Retrieved from http://www.cfcp.bham.ac.uk/Extras/SCORING%20GUIDE%20FOR%20RISK%20MATRIX%202000.9-%20SVC%20-%20(ver.%20Feb%202007).pdf

Tuner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). Random effects meta-analysis with binary outcomes using multilevel models. *Statistics in Medicine, 19,* 3417–3432. Retrieved from http://www3.interscience.wiley.com/journal/2988/home

Von Korff, M., Koepsell, T., Curry, S., & Diehr, P. (1992). Multi-level analysis in epidemiological research on health behaviors and outcomes. *American Journal of Epidemiology, 135,* 1077–1082. PMID:1632420

Walters, G. D. (2003a). Predicting criminal justice outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences and the Law, 21,* 89–102. doi:10.1002/bsl.519

Walters, G. D. (2003b). Predicting institutional adjustment and recidivism with the Psychopathy Checklist factors scores: A meta-analysis. *Law and Human Behavior, 27,* 541–558. doi:10.1023/A:1025490207678PMid:14593797

Walters, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice and Behavior, 33,* 279–304. doi:10.1177/0093854805284409

Walters, G. D., White, T. W., & Denney, D. (1991). The Lifestyle Criminality Screening Form: Preliminary data. *Criminal Justice and Behavior, 18,* 406–418. doi:10.1177/0093854891018004003

*Warren, J. I., South, S. C., Burnette, M. L., Rogers, A., Friend, R., Bale, R., & Van Patten, I. (2005). Understanding the risk factors for violence and criminality in women: The concurrent validity of the PCL-R and HCR-20. *International Journal of Law and Psychiatry, 28,* 269–289. doi:10.1016/j.ijlp.2003.09.012; PMID:15923037

Webster, C. K., Douglas, D. E., Eaves, D., & Hart, D. (1997). *HCR-20 assessing risk for violence: Version II.* Burnaby, British Columbia, Canada: Mental Health, Law & Policy Institute, Simon Fraser University.

Wong, S., & Gordon, A. (2001). The Violence Risk Scale. *Bulletin of the International Society for Research on Aggression, 23,* 16–20. Retrieved from http://www.israsociety.com

Wong, S. C. P., & Gordon, A. (2003). *Violence Risk Scale.* Available from the authors, Department of Psychology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 5A5.

*Wong, S. C. P., & Gordon, A. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology, Public Policy, and Law, 12,* 279–309. doi:10.1037/1076-8971.12.3.279

Wong, S. C. P., Gordon, A., & Gu, D. (2007). The assessment and treatment of violence-prone forensic clients: An integrated approach. *British Journal of Psychiatry, 190,* 566–574. doi:10.1192/bjp.190.5.s66; PMid:17470945

Wong, S., & Hare, R. (2005). *Guidelines for a psychopathy treatment program.* Toronto: Multi-Health Systems.

World Health Organization. (1990). *International statistical classification of diseases and related health problems* (10th Rev.). Geneva, Switzerland: Author.

*Wormith, S., Olver, M., Stevenson, H., & Girard, L. (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services, 4,* 287–305. doi:10.1037/1541-1559.4.4.287

Yang, M., Heath, A., & Goldstein, H. (2000). Multilevel models for

repeated binary outcomes: Attitudes and vote over the electoral cycle. *Journal of Royal Statistical Society, 163A*(1), 49–62. http://www.rss.org.uk/main.asp?page=1711

Yang, M., Liu, Y. Y., & Coid, W. J. (2010). *Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism: Research summary.* London: Ministry of Justice, Great Britain. Retrieved from www.justice.gov.uk/publications/research.htm

Appendix

A Brief Description of Seven Risk Assessment Tools

The Violence Risk Appraisal Guide (VRAG; Harris et al., 1993) is a 12-item actuarial tool designed to assess risk of violent recidivism and can be used for men apprehended for criminal violence and with male mentally disordered offenders. The items assess early childhood problems, alcohol problems, criminal history, *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed. [*DSM-III*]; American Psychiatric Association, 1980) diagnoses of schizophrenia, personality disorder, and so forth. The items are differentially weighted as reflected by the score assigned to the item and can be rated on the basis of a comprehensive social history. The Psychopathy Checklist score is included as one of the items and has the largest weight. Each total score has been associated with one of nine categories with a known likelihood of violent recidivism based on data from the construction sample with 7 years of follow-up data. The VRAG has been extensively validated with an average area under the curve (AUC) of .72 for the prediction of violent recidivism (Rice & Harris, 2005).

The Violent Risk Scale (VRS; Wong & Gordon, 2001, 2006) uses six static and 20 dynamic variables derived primarily from the risk, need, and responsivity principles (Andrew & Bonta, 2003). The VRS dynamic variables (measuring violence-linked attitudes, cognition, emotional regulation, community support, etc.) are changeable; changes in the dynamic factors have been shown to be associated with changes in recidivism in the community (Olver, Wong, Nicholaichuk, & Gordon, 2007). The VRS dynamic and static variables are equally weighted and are all rated on 4-point Likert scales (0, 1, 2 or 3) based on file review and a semi-structured interview. For most variables, higher ratings indicate a closer link to violence. Dynamic variables closely linked to violence (rated 2 or 3) are appropriate targets for violence reduction treatment. The total VRS score indicates the level of violence risk; the higher the score, the higher is the risk. The VRS is appropriate for use with male offenders and forensic psychiatric patients. The AUC of .74 has been reported for the prediction of violent recidivism (Wong & Gordon, 2006).

The Historical–Clinical–Risk Management–20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) is a 20-item violence risk assessment tool based on the structured professional judgment model of risk assessment. This model relies on the assessor scoring the items and clinically combining the items to arrive at a risk estimate of low, medium, or high. The Historical domain assesses the presence of personality disorder, major mental illnesses, psychopathy (using formally assessed PCL-R or PCL:SV scores), history of violence, and so forth; the Clinical domain assesses insight, active symptoms of mental illness, impulsivity, and so forth; and the Risk Management domain assesses exposure to destabilizers, availability of support and stress, and so forth. Ratings of the items are based on file information and interview. (Formal PCL-R assessment of psychopathy and diagnosis of mental disorder based on *Diagnostic and Statistical Manual of Mental Disorders* or the International Classification of Diseases is required.) The median AUC value for the HCR-20 total score across 42 studies was .69 based on the summation of the numeric scores of the HCR-20 (see Douglas & Reeves, 2009).

The Level of Service Inventory—Revised (LSI-R; Andrews & Bonta, 1995) is a 54-item survey of indicators of risk and need across 10 components: Criminal History, Education/Employment, Financial, Family/Marital, Accommodation, Leisure/Recreation, Companions, Alcohol/Drug Problems, Emotional/Personal, and Attitude /Orientation. Some items are scored *absent* (0) or *present* (1); other items are rated 0 to 3, indicating very high risk) or very low risk, respectively, on the basis of file review and interview. A most recent meta-analyses of the LSI-R indicated a predictive validity for violent recidivism with an adjusted effect size of .28 (AUC = .61).

The Offender Group Reconviction Scale—Version 2 (OGRS-2; Copas & Marshall, 1998) is a 12-item rating tool based almost entirely on past offending history and demographic information, such as offence category; various offence history indicators, such as burglary, breach of an official order, offender's age at time of sentence and earliest possible release, gender, and a composite variable that measures the quantity and speed of past offending. Rating can be done based on file review alone, as all variables are either demographic or historical in nature. Predictive validity (AUC) on a large sample of male offenders has been found to be about .72 (Coid et al., 2009).

(*Appendix continues*)

The Risk Matrix 2000V (RM2000V) is a three-item rating tool designed to predict nonsexual violence in adult males serving a prison sentence. The items are age, number of sentencing occasions for nonsexual violence, and ever conviction for burglary. Scoring can be done from file information alone. Predictive validity determined with samples of prisoners ranged from AUCs of .78 to .80 depending on length of follow-up (see Thornton, 2007).

The General Statistical Information on Recidivism Scale (GSIR; Bonta, Harman, Hann, & Cormier, 1996), originally developed in 1982 by Nuffield, is 15-item rating scale designed to assess the risk of general re-offending. Items are all historical in nature and include criminal history, marital status, and employment status and are rated with weighted scores. Lower scores on the instrument are related to higher risk for recidivism. The instrument has been reliably associated with general recidivism (AUC = .76) and has been found to predict violent recidivism as well (Bonta et al., 1996).

Received July 29, 2008
Revision received May 6, 2010
Accepted May 14, 2010 ■