Corrections Research:
User Report

# The RRASOR, Static-99R and Static-2002R All Add Incrementally to the Prediction of Recidivism among Sex Offenders

**2011-02**

Kelly M. Babchishin, R. Karl Hanson, & Leslie Helmus

Public Safety Canada

## Abstract

Empirically derived actuarial tools are increasingly being used in applied psychology, particularly for the assessment of risk for crime and violence. Although evaluators commonly use more than one scale, it is unclear how evaluators should interpret divergent findings. The current study examined the predictive accuracy and incremental validity of three risk assessment scales (RRASOR, Static-99R, and Static-2002R) in twenty distinct samples of sex offenders ($N$ = 7,491). Static-99R and Static-2002R outperformed the RRASOR in the prediction of sexual, violent, and any recidivism. No differences in predictive accuracy were found between Static-99R and Static-2002R. Nevertheless, almost all the scales provided incremental validity to the prediction of all types of recidivism. The direction of the incremental effects, however, was not consistently positive. When controlling for the other measures, high scores on the RRASOR were associated with lower risk for violent and general recidivism. Consequently, decisions concerning the interpretation of multiple risk scales must be informed by the construct validity of the measures. When scales measure the same domain of risk factors, an averaging approach can be justified. If the selected scales are not sampling the same types of risk factors, then evaluators need a defensible model concerning (1) the latent constructs measured by the scales and (2) empirical evidence concerning how the constructs should be weighted and combined.

## Authors' Note

The views expressed are those of the authors and not necessarily those of Public Safety Canada. Correspondence concerning this report should be addressed to: R. Karl Hanson, Corrections Research, Public Safety Canada, 340 Laurier Avenue West, Ottawa, ON, Canada, K1A 0P8. E-mail: karl.hanson@ps.gc.ca

## Acknowledgements

We would like to thank the following researchers for granting us permission to use their data and for being patient with our ongoing questions: Alfred Allan, Tony Beech, Susanne Bengtson, Jacques Bigras, Sasha Boer, Jim Bonta, Sébastien Brouillette-Alarie, Franca Cortoni, Margretta Dwyer, Reinhard Eher, Doug Epperson, Randolph Grace, Andy Haag, Leigh Harkins, Andreas Hill, Steve Johansen, Ray Knight, Niklas Långström, Terry Nicholaichuk, Kevin Nunes, Jean Proulx, Martin Rettenberger, Rebecca Swinburne Romine, Daryl Ternowski, Robin Wilson, and Annie Yessine.

## Product Information:

January 2011

Cat. No.: PS3-1/2011-2E-PDF

ISBN No.: 978-1-100-18295-7

Ottawa

**The RRASOR, Static-99R and Static-2002R All Add Incrementally to the Prediction of Recidivism among Sex Offenders**

Most psychological tests are designed to assess latent constructs and their results have practical importance to the extent that the latent constructs are related to outcomes of interest. Although desirable, it is not always necessary to fully understand the latent psychological constructs being assessed for a measure to have practical utility. In fact, complete understanding is rare (Cronbach & Meehl, 1955). Experts can continue to argue about the nature of major psychological constructs (e.g., positive mental health, intelligence, sexual deviance) while agreeing on the practical utility of existing measures for applied decision-making (e.g., discharge from treatment, school placement, risk assessment). Measures can have importance based simply on their empirical relationships with the outcome of interest (e.g., Meehl, 1956). Such an empirical prediction is particularly relevant when the evaluator's primary concern is predicting a discrete (i.e., yes/no) outcome (e.g., depression relapse, school failure, sexual recidivism).

One domain in which empirical prediction has gained prominence in recent years is in the evaluation of risk for crime and violence (Hanson, 2005, 2009; Quinsey, Harris, Rice, & Cormier, 2006). In the United States, the *Daubert* criteria (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993) is the most commonly used legal standard to determine whether scientific evidence (e.g., risk factors) is admissible in court (Monahan & Walker, 2010). The *Daubert* criteria requires that testimony has empirical support but the expert does not need to convince the court of a "cosmic understanding" (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993, para. 43) of the issues at hand. Using the *Daubert* criteria, US courts routinely accept empirical evidence on risk factors for crime and violence without necessarily understanding the causal mechanisms involved.

Although there is consensus that risk factors need to be empirically established (e.g., Kraemer et al., 1997), evaluators disagree on the best way of combining risk factors into an overall evaluation. Research has consistently found that structured risk assessments are more accurate than unstructured professional opinion (Gendreau, Goggin, & Law, 1997; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson & Morton-Bourgon, 2009); there is no consensus on how they should be structured.

In the violence risk assessment field, most evaluators use some form of structured professional judgement (SPJ; Archer, Buffington-Vollum, Stredny, & Handel, 2006). In this form of evaluation, the risk factors are selected in advance based on their relationship with the outcome of interest. The combination of these items into an overall evaluation, however, is left to the judgement of the evaluator (Douglas & Kropp, 2002). In contrast to SPJ, *mechanical* prediction tools specify in advance the items and provide explicit methods for combining the items into a total score (Grove et al., 2000). When mechanical prediction tools also provide empirically derived probability estimates for a particular outcome of interest, they are called *actuarial* (Dawes, Faust, & Meehl, 1989; Meehl, 1954).

The use of actuarial risk tools is common in certain high-stakes risk evaluations. In sexual civil commitment trials, for example, 95% of civil commitment evaluators report using Static-99 (an actuarial risk tool for sexual recidivism) always or most of the time (Jackson & Hess, 2007). This contrasts with decision-making methods in general clinical psychology, where the majority of psychologists (68%) rely on unstructured, clinical prediction (Vrieze & Grove, 2009).

Although the use of mechanical and actuarial risk tools has clear strengths (e.g., reduced bias, high reliability; Garb, 2003), there are barriers to their routine use. For many applied decisions, validated prediction tools are simply not available (Vrieze & Grove, 2009). The current study, however, addresses the opposite problem: What should evaluators do when there are several different risk predictions tools available?

## Use of Multiple Measures

For the prediction of recidivism among sexual offenders, evaluators have choice. A number of different tools are available including the Minnesota Sex Offender Screening Tool – Revised (MnSOST–

R; Epperson et al., 1998), Rapid Risk Assessment for Sex Offence Recidivism (RRASOR; Hanson, 1997), Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997), Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 2006), Static-99 (Hanson & Thornton, 2000) and Static-2002 (Hanson & Thornton, 2003). There is considerable overlap in their items (demographics, prior criminal history) and each tool has shown similar levels of predictive accuracy, defined in terms of their ability to differentiate sexual recidivists from non-recidivists (Hanson & Morton-Bourgon, 2009).

Although evaluators often use more than one measure (Jackson & Hess, 2007), it is not clear how to interpret the results when the measures disagree, and unfortunately, divergent results are common (e.g., Mills & Kroner, 2006). Barbaree, Langton, and Peacock (2006) found that less than 8% ($n$ = 20) of sex offenders sampled ($N$ = 262) were consistently identified as high risk or as low risk by five commonly used actuarial risk tools (i.e., Violent Risk Appraisal Guide [VRAG; Quinsey et al., 2006], SORAG, Static-99, RRASOR, and MnSOST–R). Consequently, evaluators interested in actuarial risk prediction with sexual offenders must decide which measures to use, and, if they use more than one, how to interpret divergent results.

The use of multiple measures is standard practice in many fields of applied assessment, such as neurological and cognitive assessments (Brooks, Strauss, Sherman, Iverson, & Slick, 2009; Malloy et al., 1997). For example, the use of multiple (versus single) instruments has been shown to improve the accuracy of decisions concerning cancer patients' self-reported health status (Cella et al., 1995), job performance (Avis, Kudisch, & Fortunato, 2002), and smoking behaviour outcomes (Sledjeski et al., 2006). When multiple measures are used, certain general psychometric principles inform their use and interpretation (Weiner, 2003). In general, evaluators should privilege measures that (1) can be coded reliably (adequate level of interrater reliability), (2) have relevant normative data, and, (3) make valid inferences (adequate predictive accuracy).

When these general criteria are applied to sexual risk assessment, however, no one instrument is identified as superior. Specifically, all actuarial risk tools for sex offenders have acceptable and similar levels of interrater reliability (Barbaree, Seto, Langton, & Peacock, 2001; G. T. Harris et al., 2003), and there are minimal differences in their overall predictive accuracy (Hanson & Morton-Bourgon, 2009; Rettenberger, Matthes, Boer, & Eher, 2010).

In the absence of a clear winner, psychometric theory supports the use of multiple instruments. Classical test theory holds that test error can be minimized by increasing the item pool ("the more, the better"). Specifically, an observed score on a test (or item) has two components: the true score (or under item response theory, the examinee's ability or trait parameter) and measurement error (see Rust & Golombok, 2009, for a review). As such, increasing items or instruments should reduce the amount of measurement error. Because error is theorized to be random, the errors are expected to cancel themselves out across observations (Nunnally & Bernstein, 1994). Consequently, adding items to prediction tools should result in increased predictive accuracy. Of course, if the additional items are substantially worse (less predictive) than the items already considered, the accuracy of the overall prediction would deteriorate.

## Incremental Validity

When using multiple scales in applied risk assessment, a central concern is incremental validity. Specifically, incremental validity is the extent to which new information improves the accuracy of a prediction above and beyond that of the previous instrument(s) used. Conceptually, if an instrument provides new information to better understand an offender's risk, it provides incremental information. For example, additional information about antisociality would aid in understanding an offender's risk to reoffend above that provided by a particular risk instrument that only considered mental health problems.

That certain items or domains of risk factors add incrementally to the prediction of violence or crime is uncontroversial. Indeed, the construction of most actuarial tools considered the incremental validity of the items retained in the final scale (e.g., the Level of Service Inventory-Revised [LSI-R; Andrews & Bonta, 1995], Static-99, Static-2002, and VRAG). Having these established risk assessment tools, the question

then becomes how effectively the final measures have sampled and weighted the relevant variables. Namely, of the measures that are currently in use and are intended to be global assessments of risk, to what extent is it possible to identify other variables or scales that add incrementally to these measures?

Research on the incremental validity of commonly used risk instruments is mixed. Seto (2005) found that routinely used scales (i.e., RRASOR, Static-99, SORAG, and VRAG) did not add incrementally to one another in the prediction of sexual recidivism. Such findings suggest that the use of multiple instruments is an unnecessary hassle. The study, however, was limited by a small sample size of sex offenders ($N$ = 215). In addition, of the risk instruments sampled, Seto (2005) found the RRASOR to be the most predictive of sexual recidivism. Most available studies, however, have found the RRASOR to be inferior to other available risk instruments, such as Static-99 (Hanson & Morton-Bourgon, 2009). In short, Seto's (2005) recommendation to choose the best instrument is difficult to apply because, as yet, there is no scientific consensus concerning which is the "best" instrument for the prediction of sexual recidivism, and different instruments may be better or worse for specific decisions in specific jurisdictions.

Lloyd (2008) examined a large set of actuarial instruments (MNSOST-R, Risk Matrix 2000 [Thornton et al., 2003], RRASOR, SORAG, Static-99), structured clinical guidelines (Structured Risk Assessment - Need Assessment [SRA; Thornton, 2002], SVR-20) and other variables hypothesized to predict sexual recidivism (e.g., number of male victims) in a group of sex offenders ($N$ = 391). Lloyd (2008) found that a combination of risk scales best predicted sexual recidivism and added incremental validity to one another (including the SORAG, MNSOST-R, the Social-Affective score of the SRA, and the SVR-20). Although there may be some question of overfitting due to a large number of variables entered into the regression equation, the study demonstrates the possibility that existing scales can add incrementally to one another in the prediction of sexual recidivism.

Mills and Kroner (2006) expanded the examination of incremental validity by examining the impact of discordance among the risk instruments. They examined the incremental validity of the General Statistical Information on Recidivism Scale (GSIR; Nuffield, 1982), the LSI-R, and the VRAG for the prediction of general and violent recidivism for offenders (approximately 3/4 violent offenders). Further, they divided offenders into those with low discordance among risk instruments (i.e., the average standardized differences between instruments were small, suggesting consistency across instruments in relative risk estimates) and high discordance (i.e., the average standardized differences between instruments were large, suggesting inconsistency across instruments in relative risk estimates). Mills and Kroner (2006) found that the scales added incrementally to the prediction of general and violent recidivism for offenders with low discordance ($n$ = 140), but not those with high discordance ($n$ = 69). Given the small sample size of the discordant group, a plausible explanation for the null finding is lack of statistical power required to test such hypotheses.

Welsh, Schmidt, McKinnon, Chattha, and Meyers (2008) examined the incremental validity of the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002), Structured Assessment of Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2002) and Psychopathy Checklist: Youth Version (PCL:YV; Forth, Kosson, & Hare, 2003) in a sample of juvenile offenders ($N$ = 105), for predicting general and violent recidivism. Even with a small sample size, Welsh and colleagues (2008) found that the SAVRY added incrementally to the PCL:YV and the YLS/CMI for both violent and general recidivism. In addition, the PCL:YV was found to add incrementally to the YLS/CMI, whereas the YLS/CMI did not add incremental validity to the other two scales.

In summary, there are relatively few studies examining the incremental validity of unmodified (e.g., no items removed) risk scales for crime and violence, and most available studies are limited by small sample sizes. Overall, the research suggests that multiple risk instruments may add incremental validity to one another. Further research with larger samples is required, however, to better understand whether there is practical utility in using several risk instruments.

Current Study

The purpose of the present study was to compare the predictive validity of three commonly used measures for the prediction of recidivism among sexual offenders: RRASOR, Static-99R, and Static-2002R. Specifically, we examined (1) whether the RRASOR, Static-99R, or Static-2002R predicted sexual, violent, and any recidivism more accurately than the others and (2) whether the three instruments added incremental validity to one another in the prediction of the three types of recidivism. All three scales included in the current study are similar to each other in that they have the same purpose (predicting sexual recidivism) and are based on similar demographic and criminal history variables. If one of the instruments was clearly superior in terms of predictive accuracy and no other scales added incrementally to it, evaluators would be justified in using only the "best" measure. The choice of instruments would be less clear, however, if none of the measures had superior predictive accuracy or if they were found to add incrementally to one another.

## Method

Measures

*Rapid Risk Assessment for Sex Offence Recidivism (RRASOR)*

The RRASOR (Hanson, 1997) is an actuarial instrument designed to measure risk of sexual recidivism. Scores range from 0 to 6, with a higher score indicating greater risk of sexual recidivism. It has four items: (1) prior sexual offenses, (2) any unrelated victims, (3) any male victims, and (4) offender is less than 25 years of age. For the current study, the items of Static-99 were used to compute the RRASOR. The coding rules for the items of the RRASOR and Static-99 are identical with the exception of prior sexual offences. Specifically, unlike the RRASOR, the coding rules of Static-99 do not count pseudo-recidivism as prior sexual offences. Pseudo-recidivism is estimated to affect approximately 5% of offenders (Phenix, Doren, Helmus, Hanson, & Thornton, 2009), and hence, the difference between using the item scoring of Static-99 rather than RRASOR is expected to be minimal.

In the development study, the RRASOR differentiated sexual recidivists from nonrecidivists with an Area Under the Curve (AUC) of .71 (Hanson, 1997). A recent meta-analysis conducted by Hanson and Morton-Bourgon (2009) found that the RRASOR showed similar, although slightly smaller effects, when averaged across 34 diverse follow-up studies (weighted mean $d$ = 0.60, 95% CI = 0.54 to 0.65, $N$ = 11,031, $k$ = 34; which translates to an AUC of .66, 95% CI = .65 to .68).

*Static-99R*

Static-99R is a 10-item actuarial measure that assesses recidivism risk of adult male sexual offenders. The items are identical to Static-99 (Hanson & Thornton, 2000), with the exception of updated age weights (see Helmus, Thornton, Hanson, & Babchishin, 2010). In Canada and the United States, Static-99 is the most commonly used actuarial scale to predict sexual recidivism (Archer et al., 2006; Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). It is commonly used for treatment planning (McGrath et al., 2010; Jackson & Hess, 2007), community supervision (Interstate Commission for Adult Offender Supervision, 2007), and civil commitment evaluations (Jackson & Hess, 2007).

Static-99R contains all the RRASOR items as well additional items concerned with relationship history (1 item), sexual offence history (stranger victims, non-contact sexual offences), and general criminal history (number of prior sentencing occasions, index non-sexual violence, prior non-sexual violence; see Table 1). A recent meta-analysis found a moderate relationship between Static-99 and sexual recidivism (weighted mean $d$ = 0.67, 95% CI = 0.62 to 0.72, $N$ = 20,010, $k$ = 63; which translate to an AUC for ROC of .68, 95% CI = .67 to .70; Hanson & Morton-Bourgon, 2009). For an overview of research on Static-99, see Anderson and Hanson (2010).

Table 1. *Items Contained in the RRASOR, Static-99R and Static-2002R*

| Notes | RRASOR | STATIC-99/STATIC-99R | STATIC-2002/STATIC-2002R |
|---|---|---|---|
| a | Offender's age at release | Offender's age at release | Offender's age at release |
| b | Number of prior sexual offence charges and convictions | Number of prior sexual offence charges and convictions | Prior sentencing occasions for sexual offences |
| c | Any unrelated victims of sexual assaults | Any unrelated victims of sexual assaults | Any unrelated victims of sexual assaults |
| c | Any male victims of sexual assaults | Any male victims of sexual assaults | Any male victims of sexual assaults |
| d | | Convictions for non-contact sexual offences | Convictions for non-contact sexual offences |
| d | | Any stranger victims of sexual assaults | Any stranger victims of sexual assaults |
| a | | Number of prior sentencing dates | Prior sentencing occasions for anything |
| e | | Conviction for non-sexual violence prior to the Index Offence | Prior violent non-sexual sentencing occasion |
| f | | Conviction for non-sexual violence at the time of the Index Offence | Any prior involvement with the criminal justice system |
| f | | Ever lived with an intimate partner for two consecutive years | Any young, unrelated victims |
| f | | | Rate of sexual offences |
| f | | | Any community supervision violation |
| f | | | Arrests for sexual offences as both an adult and a juvenile |
| f | | | Years free prior to Index |

*Note*. Adapted from A. J. R. Harris and Hanson (2010). Static-99 and Static-2002 are identical with their "R" versions, with the exception of the cut-points and weights accorded to age.
[a]Same definition, but different cut-points and weights.
[b]Static -99 and RRASOR have the same definitions and same weights for prior sex offences, but Static -99 scoring includes the concept of "pseudo-recidivism" whereas RRASOR does not. Static-2002 has a different definition than the other measures.
[c]Identical item across all three measures.
[d]Identical item for Static -99 and Static -2002.
[e]Similar concepts, different definitions.
[f]Different items (no equivalent on the other scale).

### *Static-2002R*

Static-2002 (Hanson & Thornton, 2003) was created with the aim of improving Static-99. Static-2002R is a 14-item actuarial measure that assesses recidivism risk of adult male sexual offenders. The items are identical to Static-2002 (Hanson & Thornton, 2003), with the exception of updated age weights (see Helmus et al., 2010). Important differences between Static-99 and Static-2002 are that Static-2002 added and altered some items, organized items into meaningful subscales to aid interpretation, and has more

standardized coding rules. Static-2002 has a moderate relationship with sexual recidivism (weighted mean $d$ = 0.70, 95% CI = 0.59 to 0.81, $N$ = 3,330, $k$ = 8; which translate to an ROC of .69, 95% CI = .66 to .72; Hanson & Morton-Bourgon, 2009). Previous research found that Static-2002 was more predictive of sexual, violent, and any recidivism than Static-99 (Hanson, Helmus, & Thornton, 2010; Stalans, Hacker, & Talbot, 2010).

A list of the items in the RRASOR, Static-99R, and Static-2002R is provided in Table 1. For further information on Static-99R and Static-2002R, see http://www.static99.org.

## Samples

Multiple samples from diverse jurisdictions were used. Table 2 presents the main characteristics of each sample ($k$ = 20, $N$ = 7,491). All twenty samples had both RRASOR and Static-99R scores, but only 7 had Static-2002R scores. Most samples were drawn from Canada ($k$ = 10) or United States ($k$ = 4), followed by single samples from Austria, Denmark, Germany, New Zealand, Sweden, and United Kingdom. The current study examined three types of recidivism: sexual, violent (including sexual recidivism), and any recidivism. Of the 20 samples, 4 samples only reported sexual recidivism, 2 samples reported both sexual and violent recidivism, and 14 samples reported all three types of recidivism.

Each dataset was verified for internal inconsistencies (e.g., miscalculation of total scores or item scores contradicted by other information in the dataset). Identified errors were corrected if possible; otherwise, the case was deleted. Cases were also deleted under the following circumstances: missing follow-up information, any missing Static-99R item other than Ever Lived with a Lover (Item 2), more than one missing Static-2002R item, the offender was less than 18 years old at time of release or less than 16 years old when they committed the index offence, or if the offender was female. The age and gender exclusionary criteria are specified in the coding rules for Static-99 (A. J. R. Harris, Phenix, Hanson, & Thornton, 2003) and Static-2002 (Phenix et al., 2009). The new age item of Static-99R and Static-2002R was calculated from the verified datasets for each sample.

The number of participants in these samples was smaller than previously reported (e.g., Helmus, 2009) because (1) the date of birth or age of the offender at release was required to code the new Static-99R and Static-2002R age weights, and (2) the total scores of at least two of the scales included in this study had to be available in the dataset (e.g., Static-99 item scores were needed to calculate RRASOR total scores). The samples are described in detail in Helmus (2009; available from http://www.static99.org).

## Overview of Analyses

All analyses were conducted separately by the first and third author to ensure accuracy.

### *Predictive accuracy*

The first set of analyses used fixed-effect and random-effects meta-analyses to compute the weighted areas under receiver operating characteristic curves (ROC AUC) and 95% confidence intervals for each risk instrument. The AUC is a measure of relative risk and can be interpreted as the probability that a randomly selected recidivist has a higher score on the risk instrument than a randomly selected non-recidivist. The AUC is useful for comparing results across samples because it is not influenced by recidivism base rates (Rice & Harris, 1995). It is, however, influenced by the variance in the distribution of scores used to predict recidivism (Hanson, 2008; Humphreys & Swets, 1991).

Fixed-effect estimates of the AUCs and standard errors were calculated using the formula and procedures presented in Hedges (1994). Fixed-effect analyses have the advantage of providing an estimate of between-study variability (i.e., Cochran's $Q$ statistic; Hedges & Olkin, 1985). A significant Cochran's $Q$ statistic indicates that there is more variability across studies than expected by chance (the $Q$ statistic is distributed as a chi-square, with $k - 1$ degrees of freedom). In random-effects meta-analysis, the between-study variability is included in the error term, resulting in wider (and often more realistic) confidence intervals (Schmidt, Oh, & Hayes, 2009). The results of the random-effects and fixed-effect models

Table 2. *Descriptive Information of Samples*

| Study | *N* | Release Period | Follow-up (*SD*)[a] | Recidivism Rates | | | Age (*SD*) | RRASOR | Static-99R | Static-2002R |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Sexual | Violent[b] | Any | | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| Allan et al. (2007) | 492 | 1990-2000 | 5.7 (2.9) | 9.6 | 16.5 | 25.2 | 42.3 (12.2) | 1.4 (1.4) | 1.8 (2.3) | - |
| Bengtson (2008) | 308 | 1978-1995 | 16.2 (4.2) | 34.1 | 52.3 | 64.6 | 32.5 (10.4) | 1.8 (1.2) | 3.8 (2.4) | 4.6 (2.4) |
| Bigras (2007) | 457 | 1995-2004 | 4.6 (1.9) | 5.7 | 14.7 | 23.4 | 42.8 (12.0) | 1.3 (1.3) | 2.1 (2.4) | 3.5 (2.5) |
| Boer (2003) | 296 | 1976-1994 | 13.3 (2.1) | 8.8 | 23.3 | 48.3 | 41.2 (12.5) | 1.4 (1.2) | 2.8 (2.8) | 3.9 (2.7) |
| Bonta & Yessine (2005) | 133 | 1992-2004 | 5.5 (2.4) | 15.8 | 33.8 | 48.9 | 39.8 (9.6) | 2.7 (1.3) | 5.0 (2.1) | - |
| Brouillette-Alarie & Proulx (2008) | 228 | 1979-2006 | 9.9 (4.5) | 20.2 | 30.7 | - | 36.0 (10.2) | 2.1 (1.4) | 3.9 (2.3) | - |
| Cortoni & Nunes (2007) | 73 | 2001-2004 | 4.6 (0.6) | 0.0 | 8.2 | 12.3 | 41.6 (12.3) | 1.2 (1.0) | 2.2 (2.1) | - |
| Eher et al. (2008) | 706 | 2000-2005 | 3.9 (1.1) | 4.0 | 14.7 | 26.2 | 40.7 (12.6) | 1.2 (1.0) | 2.3 (2.3) | - |
| Epperson (2003) | 177 | 1989-1998 | 7.9 (2.5) | 14.1 | - | - | 37.2 (13.2) | 1.5 (1.2) | 2.5 (2.6) | - |
| Haag (2005) | 190 | 1995 | 7.0 (0.0) | 24.7 | - | - | 36.7 (9.7) | 2.0 (1.4) | 4.1 (2.2) | 5.7 (2.3) |
| Hanson et al. (2007) | 702 | 2001-2005 | 3.4 (1.0) | 8.1 | 16.4 | 27.9 | 41.6 (13.2) | 1.5 (1.2) | 2.4 (2.4) | 3.5 (2.5) |
| Harkins & Beech (2007) | 190 | 1994-1998 | 10.4 (1.1) | 14.2 | 21.1 | 36.3 | 43.3 (12.5) | 1.5 (1.3) | 2.2 (2.6) | 3.7 (2.8) |
| Hill et al. (2008) | 86 | 1971-2002 | 12.6 (6.6) | 15.1 | 29.1 | 61.6 | 39.4 (11.1) | 1.9 (1.0) | 4.7 (2.0) | - |
| Johansen (2007) | 273 | 1994-2000 | 9.1 (1.1) | 7.7 | 20.5 | 53.5 | 37.8 (10.8) | 1.8 (1.2) | 2.9 (2.3) | - |
| Knight & Thornton (2007) | 466 | 1957-1986 | 8.6 (2.6) | 26.2 | 36.9 | 53.0 | 36.1 (11.4) | 2.4 (1.3) | 4.6 (2.4) | 6.1 (2.5) |
| Långström (2004) | 1,278 | 1993-1997 | 8.9 (1.4) | 7.5 | 21.4 | - | 41.5 (12.0) | 0.8 (0.9) | 2.0 (2.4) | - |
| Nicholaichuk (2001) | 281 | 1983-1998 | 6.4 (4.0) | 18.5 | - | - | 34.8 (9.4) | 2.4 (1.4) | 4.8 (2.4) | - |
| Swinburne Romine et al. (2008) | 680 | 1977-2007 | 16.8 (7.8) | 13.8 | - | - | 38.2 (12.3) | 1.2 (1.1) | 1.7 (2.2) | - |
| Ternowski (2004) | 247 | 1994-1998 | 7.5 (1.0) | 8.1 | 15.4 | 19.8 | 43.9 (13.0) | 1.2 (1.2) | 1.6 (2.5) | - |
| Wilson et al. (2007 a & b) | 228 | 1994-2007 | 5.2 (3.0) | 10.5 | 25.9 | 35.5 | 41.7 (11.4) | 2.8 (1.5) | 5.1 (2.3) | - |
| Total | 7,491 | 1957-2007 | 8.3 (5.2) | 12.0 | 22.4 | 35.9 | 39.8 (12.2) | 1.5 (1.3) | 2.7 (2.6) | 4.3 (2.7) |

*Note.*[a]Follow-up period in years. [b]Violent recidivism, including sexual.

therefore converge as the amount of between-study variability decreases (when *Q* is less than the degrees of freedom, the results are identical). Random-effects estimates were calculated using Formulae 10, 12, and 14 from Hedges and Vevea (1998).

The Hanley and McNeil (1983) test of correlated ROC areas was used to test whether the risk instruments differed in their level of predictive accuracy. The Hanley and McNeil test requires the following: (1) the average AUC for the two risk instruments that are being compared, and (2) the average correlation between the two instruments being compared, computed separately for the recidivists and non-recidivists. The AUCs and average correlations were computed for each of the three recidivism type (sexual, violent including sexual, and any recidivism). Hanley and McNeil (1983) proposed the use of the Kendal Tau ($\tau$) correlation rather than the Pearson correlation. The $\tau$ is a rank correlation that represents the relationship between the ordering of the data when ranked by the two separate measures (i.e., for ordinal data). The $\tau$ therefore provides a more conservative test compared to the Pearson correlation, which assumes interval data. Table 1 (Hanley & McNeil, 1983, p. 841) associates an overall correlation based on the average AUC (for the two measures being compared) and the average $\tau$ (between the measures for the recidivists and the non-recidivists). We will refer this new correlation derived from Table 1 (Hanley & McNeil, 1983, p. 841) as the overall average *r*. Standard errors for the differences between two AUCs (A1 – A2) were based on Hanley and McNeil's (1983) Formula 3:

$$SD_{A_1-A_2} = \sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}$$

where *r* is the overall average *r*, and SE is the respective standard errors for the AUC of each measure. If the 95% confidence interval of the difference between measures included zero, the difference between the two scales was not statistically significant.

Incremental validity

Incremental validity was examined using Cox regression (Allison, 1984). Cox regression estimates relative risk ratios (hazard rates) associated with one or more predictor variables from survival data with unequal follow-up times. Each sample was used as a stratum to allow separate baseline hazard functions (i.e., recidivism rates) for each value of the stratified variable, effectively removing from the analysis the base rate variability across samples.

## Results

Predictive Validity

The predictive validity of the three scales was measured using AUCs. Appendix A presents the AUCs for the RRASOR, Static-99R and Static-2002R by sample. Tables 3 to 5 present the weighted AUC for each risk instrument and the Hanley and McNeil test. Static-99R and Static-2002R predicted sexual, violent, and any recidivism similarly, with no one scale displaying greater predictive accuracy (Table 3). Given that $\tau$ provides a more conservative test than the Pearson correlation, the analyses were also computed using Pearson correlations. The results were similar, with one exception: Static-2002R was significantly better than Static-99R in predicting any recidivism (difference with fixed-effect = 0.0133, 95% CI = 0.00275 to 0.0238; difference with random-effects = 0.0138, 95% CI = 0.00115 to 0.0265) using the Pearson correlation but this difference was not found when using $\tau$ correlation coefficient.

Table 4 presents the meta-analyzed AUC for the RRASOR and Static-99R. The Hanley and McNeil test found that Static-99R had significantly greater accuracy in predicting sexual, violent, and any recidivism than the RRASOR, with larger differences found for violent (including sexual) and any recidivism. The same pattern of results was found for the RRASOR and Static-2002R, with Static-2002R predicting sexual, violent, and any recidivism more accurately than the RRASOR (see Table 5).

The differences in predictive accuracy between scales were similar for both the fixed-effect and random-

Table 3. *Meta-analysis of Prediction AUC Areas for Static-99R and Static-2002R*

| Outcome | Measure | Fixed | | | Random | | | *k* | *N* | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Weighted ROC | 95% CI | | Weighted ROC | 95% CI | | | | |
| | | | *LL* | *UL* | | *LL* | *UL* | | | |
| Sexual | | | | | | | | | | |
| | Static-99R | **0.684** | 0.655 | 0.713 | **0.699** | 0.641 | 0.757 | 7 | 2,609 | 19.40** |
| | Static-2002R | **0.686** | 0.657 | 0.714 | **0.696** | 0.644 | 0.749 | 7 | 2,609 | 14.53* |
| Violent | | | | | | | | | | |
| | Static-99R | **0.703** | 0.679 | 0.727 | **0.705** | 0.658 | 0.752 | 6 | 2,419 | 16.05** |
| | Static-2002R | **0.708** | 0.684 | 0.731 | **0.708** | 0.659 | 0.756 | 6 | 2,419 | 18.02** |
| Any | | | | | | | | | | |
| | Static-99R | **0.718** | 0.697 | 0.739 | **0.712** | 0.657 | 0.768 | 6 | 2,419 | 32.50*** |
| | Static-2002R | **0.732** | 0.711 | 0.753 | **0.727** | 0.674 | 0.780 | 6 | 2,419 | 31.01*** |
| Difference Between Static-99R and Static-2002R | | | | | | | | | | |
| Sexual | | **0.00183** | -0.0183 | 0.0220 | **0.000787** | -0.0226 | 0.0242 | 7 | 2,609 | 2.15 |
| Violence | | **0.00341** | -0.0126 | 0.0194 | **0.00325** | -0.0146 | 0.0211 | 6 | 2,419 | 1.29 |
| Any | | **0.0132** | -0.000933 | 0.0274 | **0.0135** | -0.00224 | 0.0292 | 6 | 2,419 | 1.21 |

*Note.* CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$, ** $p < .01$, ***$p < .001$, on a Chi-Square distribution with $k - 1$ degrees of freedom.

Table 4. *Meta-analysis of Prediction AUC Areas for RRASOR and Static-99R*

| Outcome | Measure | Fixed | | | Random | | | *k* | *N* | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Weighted ROC | 95% CI | | Weighted ROC | 95% CI | | | | |
| | | | *LL* | *UL* | | *LL* | *UL* | | | |
| Sexual | | | | | | | | | | |
| | RRASOR | **0.661** | 0.642 | 0.680 | **0.660** | 0.628 | 0.691 | 19 | 7,418 | 28.16 |
| | Static-99R | **0.694** | 0.675 | 0.713 | **0.697** | 0.664 | 0.730 | 19 | 7,418 | 35.71** |
| Violent | | | | | | | | | | |
| | RRASOR | **0.614** | 0.597 | 0.631 | **0.605** | 0.574 | 0.636 | 16 | 6,163 | 29.76** |
| | Static-99R | **0.725** | 0.710 | 0.740 | **0.707** | 0.675 | 0.739 | 16 | 6,163 | 48.85*** |
| Any | | | | | | | | | | |
| | RRASOR | **0.582** | 0.564 | 0.600 | **0.576** | 0.547 | 0.605 | 14 | 4,657 | 19.79 |
| | Static-99R | **0.709** | 0.693 | 0.724 | **0.700** | 0.665 | 0.735 | 14 | 4,657 | 48.71*** |
| Difference Between RRASOR and Static-99R | | | | | | | | | | |
| Sexual | | **0.0304** | 0.0120 | 0.0489 | **0.0349** | 0.00847 | 0.0613 | 19 | 7,418 | 16.37 |
| Violence | | **0.104** | 0.0877 | 0.120 | **0.101** | 0.0752 | 0.127 | 16 | 6,163 | 20.69 |
| Any | | **0.123** | 0.106 | 0.139 | **0.124** | 0.0939 | 0.154 | 14 | 4,657 | 27.19** |

*Note.* CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$, ** $p < .01$, ***$p < .001$, on a Chi-Square distribution with $k - 1$ degrees of freedom.

Table 5. *Meta-analysis of Prediction AUC Areas for RRASOR and Static-2002R*

| Outcome | Measure | Fixed | | | Random | | | *k* | *N* | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Weighted ROC | 95% CI | | Weighted ROC | 95% CI | | | | |
| | | | *LL* | *UL* | | *LL* | *UL* | | | |
| Sexual | | | | | | | | | | |
| | RRASOR | **0.650** | 0.621 | 0.680 | **0.655** | 0.609 | 0.702 | 7 | 2,609 | 8.76 |
| | Static-2002R | **0.686** | 0.657 | 0.714 | **0.696** | 0.644 | 0.749 | 7 | 2,609 | 14.53* |
| Violent | | | | | | | | | | |
| | RRASOR | **0.603** | 0.577 | 0.630 | **0.604** | 0.566 | 0.642 | 6 | 2,419 | 5.37 |
| | Static-2002R | **0.708** | 0.684 | 0.731 | **0.708** | 0.659 | 0.756 | 6 | 2,419 | 18.02** |
| Any | | | | | | | | | | |
| | RRASOR | **0.586** | 0.562 | 0.610 | **0.585** | 0.553 | 0.618 | 6 | 2,419 | 4.52 |
| | Static-2002R | **0.732** | 0.711 | 0.753 | **0.727** | 0.674 | 0.780 | 6 | 2,419 | 31.01*** |
| Difference Between RRASOR and Static-2002R | | | | | | | | | | |
| Sexual | | **0.0349** | 0.00650 | 0.0633 | **0.0370** | 0.00311 | 0.0710 | 7 | 2,609 | 2.40 |
| Violence | | **0.0985** | 0.0739 | 0.123 | **0.102** | 0.0613 | 0.142 | 6 | 2,419 | 9.41 |
| Any | | **0.139** | 0.117 | 0.161 | **0.139** | 0.0952 | 0.184 | 6 | 2,419 | 17.34** |

*Note.* CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$, ** $p < .01$, ***$p < .001$, on a Chi-Square distribution with $k - 1$ degrees of freedom.

effects analyses. In addition, the differences in predictive accuracy between the scales were remarkably consistent across samples for the prediction of sexual and violent recidivism, as indicated by a non-significant *Q*. For any recidivism, the comparison between the RRASOR and Static-99R as well as the comparison between the RRASOR and Static-2002R had significant variability, indicating that the difference in predictive accuracy for these comparisons were inconsistent across the samples.

Incremental Validity

Tables 6 to 8 present the Cox regression analyses used to examine the incremental validity of the risk instruments for each recidivism type. For the prediction of sexual recidivism, risk instruments were found to add incrementally to one another despite large correlations between instruments, ranging between .70 and .92 (Table 6). The RRASOR and Static-99R each added incrementally to one another; Static-99R and Static-2002R each added incrementally to one another; and, finally, Static-2002R added incremental validity to the RRASOR but the RRASOR did not add incrementally to Static-2002R. In addition, entering all three risk instruments into a model found that Static-99R and Static-2002R added incrementally to the model, but not the RRASOR. Namely, adding the RRASOR after accounting for both Static-99R and Static-2002R did not significantly improve the predictive accuracy of the model ($\chi^2$ change = 0.48, $df$ = 1, $p$ = .49).

Table 6. *Incremental Validity of the Risk Instrument for Predicting Sexual Recidivism*

| | Sexual Recidivism | | | | | |
|---|---|---|---|---|---|---|
| | *N* | *r* | *Exp(B)* | 95% CI | | Wald |
| | | | | *LL* | *UL* | |
| Comparison 1 | | | | | | |
| RRASOR | 7,410 | .702 | 1.11 | 1.04 | 1.19 | 9.75** |
| Static-99R | | | 1.26 | 1.21 | 1.31 | 143.15*** |
| Comparison 2 | | | | | | |
| RRASOR | 2,606 | .703 | 1.06 | 0.96 | 1.17 | 1.27 |
| Static-2002R | | | 1.23 | 1.17 | 1.30 | 55.17*** |
| Comparison 3 | | | | | | |
| Static-99R | 2,606 | .925 | 1.14 | 1.04 | 1.25 | 8.22** |
| Static-2002R | | | 1.12 | 1.03 | 1.23 | 6.62* |
| Comparison 4 | | | | | | |
| RRASOR | 2,606 | - | 1.04 | 0.94 | 1.15 | 0.48 |
| Static-99R | | | 1.14 | 1.04 | 1.25 | 7.41** |
| Static-2002R | | | 1.11 | 1.02 | 1.22 | 5.14* |

*Note.* Analyses conducted separately for each comparison, with each sample entered as strata and both risk instruments entered in Step 1. Sample sizes fluctuate due to the amount of cases censored before earliest event. *r* = correlation between measures; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
*$p$ < .05, **$p$ < .01, ***$p$ < .001.

For the prediction of violent (including sexual) recidivism, all three instruments added incremental information for all analyses. Of note, however, was that the incremental effect for the RRASOR was reversed – namely, low scores on the RRASOR were associated with higher rates of violent recidivism once the other scales were controlled for (see Table 7). In addition, a model that included all three risk instruments found significant incremental validity for each instrument (with low scores on the RRASOR predicting violent recidivism).

Table 7. *Incremental Validity of the Risk Instrument for Predicting Violent Recidivism*

| | Violent (including Sexual) Recidivism | | | | | |
|---|---|---|---|---|---|---|
| | *N* | *r* | *Exp(B)* | 95% CI | | Wald |
| | | | | *LL* | *UL* | |
| Comparison 1 | | | | | | |
| RRASOR | 6,161 | .691 | 0.83 | 0.79 | 0.88 | 40.30*** |
| Static-99R | | | 1.42 | 1.37 | 1.46 | 499.54*** |
| Comparison 2 | | | | | | |
| RRASOR | 2,417 | .708 | 0.83 | 0.76 | 0.91 | 17.09*** |
| Static-2002R | | | 1.34 | 1.28 | 1.40 | 165.81*** |
| Comparison 3 | | | | | | |
| Static-99R | 2,417 | .927 | 1.16 | 1.08 | 1.26 | 15.38*** |
| Static-2002R | | | 1.10 | 1.02 | 1.18 | 6.33* |
| Comparison 4 | | | | | | |
| RRASOR | 2,417 | - | 0.80 | 0.74 | 0.88 | 23.53*** |
| Static-99R | | | 1.20 | 1.11 | 1.30 | 22.04*** |
| Static-2002R | | | 1.16 | 1.07 | 1.25 | 13.88*** |

*Note.* Analyses conducted separately for each comparison, with each sample entered as strata and both risk instruments entered in Step 1. Sample sizes fluctuate due to the amount of cases censored before earliest event. *r* = correlation between measures; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
*$p < .05$, **$p < .01$, ***$p < .001$.

For the prediction of any recidivism, all comparisons found that the risk instruments added incremental validity to one another (see Table 8). Specifically, the RRASOR and Static-99R added incrementally to one another, Static-99R and Static-2002R added incrementally to one another, and, finally, the RRASOR and Static-2002R added incrementally to one another. Similarly to the prediction of violent recidivism, higher scores on the RRASOR were associated with lower probability of any recidivism. Lastly, a model that included all three risk instruments found significant incremental validity for each instrument (with low RRASOR score predicting high rates of general recidivism).

To examine the practical importance of the incremental finding, participants were also sorted into risk categories (low, moderate, and high) based on a scale-independent definition of nominal risk categories suggested by Babchishin and Hanson (2009) for Static-99R and Static-2002R. Specifically, offenders with a score associated with less than half the rate of sexual re-offending than the typical offender (risk ratio < 0.50) were classified as "low-risk." Offenders with a score associated with more than half the rate of re-offending than the typical offender, but less than twice the rate of re-offending of a typical offender

Table 8. *Incremental Validity of the Risk Instrument for Predicting Any Recidivism*

| | Any Recidivism | | | | | |
|---|---|---|---|---|---|---|
| | *N* | *r* | *Exp(B)* | 95% CI | | Wald |
| | | | | *LL* | *UL* | |
| Comparison 1 | | | | | | |
| RRASOR | 4,655 | .697 | 0.77 | 0.73 | 0.81 | 98.04*** |
| Static-99R | | | 1.40 | 1.36 | 1.44 | 538.38*** |
| Comparison 2 | | | | | | |
| RRASOR | 2,418 | .708 | 0.74 | 0.68 | 0.79 | 71.66*** |
| Static-2002R | | | 1.40 | 1.35 | 1.46 | 337.57*** |
| Comparison 3 | | | | | | |
| Static-99R | 2,418 | .927 | 1.10 | 1.03 | 1.17 | 9.09** |
| Static-2002R | | | 1.15 | 1.09 | 1.22 | 21.76*** |
| Comparison 4 | | | | | | |
| RRASOR | 2,418 | - | 0.72 | 0.67 | 0.77 | 81.16*** |
| Static-99R | | | 1.15 | 1.08 | 1.23 | 19.32*** |
| Static-2002R | | | 1.25 | 1.17 | 1.33 | 48.36*** |

*Note*. Analyses conducted separately for each comparison, with each sample entered as strata and the risk instruments entered in Step 1. Sample sizes fluctuate due to the amount of cases censored before earliest event. *r* = correlation between measures; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
*$p < .05$, **$p < .01$, ***$p < .001$.

(risk ratio 0.50–1.99) were classified as "moderate-risk." Lastly, offenders with a score associated with twice the re-offending rate of a typical offender (risk ratio > 2.00) were classified as "high-risk" (see Hanson, Lloyd, Helmus, & Thornton, 2010 for more details about relative risk ratios). The relative risk ratios, centered on the median scores in routine (non-selected) samples of sex offenders, were calculated in a previous study (Hanson et al., 2010) using Cox regression coefficients, yielding the following categories for Static-99R (low: -3 to -1; moderate: 0 to 4; high: 5+) and Static-2002R (low: -2 to 1; moderate: 2 to 6; high: 7+; see Appendix B).

A simple crosstab of the sexual recidivism rates by Static-99R and Static-2002R risk categories is presented in Table 9 to allow for a visual representation of the recidivism rates of offenders for whom the scales provide discordant results (when both instruments sort offenders into different risk categories). Recidivism rates for discordant groups were intermediate between the two adjacent risk categories. For example, when both instruments classified offenders as moderate risk, the observed recidivism rate was 10.7% (146/1,360), and when both instruments rated offenders as high risk, the observed rate was 34.4% (174/506). When one instrument classified the offender moderate and the other instrument classified the offender high, the observed sexual recidivism rate was 21.9% (73/334).

Table 9. *Distribution of Static-99R/2002R Risk Category and Observed Sexual Recidivism Rates*

| | Static-2002R | | | |
|---|---|---|---|---|
| | Low | Moderate | High | Total |
| | % ($n_{recidivist}/n$) | % ($n_{recidivist}/n$) | % ($n_{recidivist}/n$) | % ($N_{recidivist}/N$) |
| Static-99R | | | | |
| Low | 2.9% (7/244) | 0.0% (0/5) | - | 2.8% (7/249) |
| Moderate | 6.2% (10/160) | 10.7% (146/1,360) | 20.4% (10/49) | 10.6% (166/1,569) |
| High | - | 22.1% (63/285) | 34.4% (174/506) | 30.0% (237/791) |
| Total | 4.2% (17/404) | 12.7% (209/1,650) | 33.2% (184/555) | $N$ = 2,609 |

*Note.* Sexual recidivism rates from all cases, not controlling for length of follow-up. Average follow-up = 8.0 years ($SD$ = 4.9).

## Discussion

The purpose of the present study was to examine the relative and incremental validity of three scales designed to predict recidivism among sexual offenders. The current study found that Static-99R and Static-2002R outperformed the RRASOR in the prediction of sexual, violent, and any recidivism. No differences in predictive accuracy were found between Static-99R and Static-2002R. Despite large correlations between the scales, they all added incremental validity to one another for predicting sexual, violent, and any recidivism, with one exception: the RRASOR not adding incremental validity to the prediction of sexual recidivism after controlling for Static-2002R. Interestingly, the RRASOR added incrementally in a negative direction for violent and any recidivism, with higher scores indicating lower risk.

The finding of incremental validity in the current study is truly remarkable given the substantial overlap in the items of these scales, and is in stark contrast with Seto (2005) who did not find incremental validity of similar risk scales (albeit using a much smaller sample). It would be easy to assume that the high correlations between risk scales would preclude incremental validity. Given substantial overlap in content, Vrieze and Grove (2010) assumed that discordant results between the measures would form "…a prima facie reason to disbelieve" either scale and would "…undercut each others' statuses as knowledge claims" (Vrieze & Grove, 2010, p. 388). The current findings suggest that Vrieze and Grove (2010) are only partially correct. Equally valid measures can give divergent results. Even when the items "look" similar, they can be related to recidivism through different causal mechanisms, a point we will return to later in the discussion.

A previous meta-analysis with seven of the datasets used in the current study found that Static-2002 outperformed Static-99 in predicting sexual, violent, and any recidivism (Hanson et al., 2010). Stalans and colleagues (2010) also found that Static-2002 outperformed Static-99 in predicting sexual recidivism. The reason for the lack of differences in predictive accuracy between the revised versions of Static-99 and Static-2002 in the current study (despite using the same samples as Hanson et al., 2010) can most likely be attributed to the updated age weights in the revised scales. The revised age weights notably increased the predictive accuracy of Static-99R, whereas a smaller improvement was found in Static-2002R (Helmus et al., 2010). As such, Static-99R and Static-2002R are more similar in predictive accuracy than the original scales. There were also differences in statistical analyses between Hanson and colleagues (2010) and the current study. Specifically, Hanson and colleagues (2010) used Pearson correlation coefficients to compute the Hanley and McNeil (1983) test (a less conservative test than the Kendall's Tau correlation coefficients) whereas in the current study we used the Kendall's Tau. Re-analyzing Hanson and colleagues' (2010) data using Kendall's Tau, however, did not alter the findings (i.e., Static-

2002 still significantly outperformed Static-99). As such, the similarity in predictive accuracy for the revised version of Static-99 and Static-2002 is likely due to the revised age weights of the scales rather than the method used to examine the difference in predictive accuracy.

Item Weighting

The finding of incremental validity in the current study demonstrates that the original weighting of the items in the RRASOR, Static-99R, and Static-2002R was not optimal. Remarkably, the RRASOR was found to add incremental validity to Static-99R in the prediction of sexual recidivism, despite the fact that all the items of the RRASOR are included in Static-99R. (In fact, we used the items of Static-99R to calculate the RRASOR.) The incremental validity findings therefore cannot be attributed to new constructs being captured by the RRASOR, but to the different weighting of the items.

Our findings provide clear evidence that the weightings for actuarial scales are unlikely to ever be optimal. Given large enough samples sizes, the null hypothesis (finding no incremental validity) can almost always be rejected (Cohen, 1994). The refinement of weights, however, is a never-ending task requiring larger sample sizes for decreasingly small gains in precision. Test developers also need to be vigilant about over-fitting the data, as small adjustments rarely generalize to other datasets (Cureton, 1950). As well, complex weights reduce practical ease of the scoring and increase the risk of error; integers are relatively simple.

Although some progress in risk assessment can be made by improving item weights, we do not believe this will solve the most pressing problems of applied risk assessment. Instead, we believe the way forward involves increasing attention to the construct validity of prediction tools.

Construct Validity and Combining Multiple Risk Scales

Most psychological tests are designed to assess latent constructs, such as mental health and intelligence. As such, concordance among alternate measures of the same construct (e.g., different intelligence tests) is expected, and evaluators routinely average findings from multiple measures (Weiner, 2003). Such an averaging approach is based on the assumption of classical test theory that increasing the item pool should reduce sampling error and produce more reliable results (Nunally & Bernstein, 1994). Evaluators who find concordance between measures have increased confidence in the results.

The scores used for violence risk prediction, however, have often been selected on a purely empirical basis, with little attention to construct validity. Without knowing what is being assessed, it is difficult for evaluators to know how to combine the results of different risk tools. The preferred method of combination will depend on whether or not the scales are measuring similar or different constructs.

When scales sample items from the same domains and have similar relationships with the outcome (i.e., recidivism), then it is plausible to base conclusions on the average of the measures. For example, in the current study Static-99R and Static-2002R had similar contributions to the prediction of sexual recidivism and can be assumed to sample from, and give similar weights to, the same latent constructs. Despite a relatively small incremental effect between Static-99R and Static-2002R, there was a noticeable difference in the recidivism rates of discordant cases. Namely, when Static-99R and Static-2002R were discordant, there was an approximately 10% difference in observed recidivism rates, with the recidivism rates of the discordant cases being intermediate between the two respective risk categories. A 10% difference is similar in size to the effects found for most of the well established risk factors (e.g., any male victims, single, any unrelated victims; Hanson & Bussière, 1998).

When scales sample items from different domains, it is less clear how to combine their findings into one coherent judgment. When scales measure different constructs, it should not be a surprise that the scales rank offenders differently. The average of the two distinct scales may not be advisable as it may result in decreased predictive accuracy compared to other methods of combining the results. For example, the RRASOR attributes more weight to sexual deviancy than Static-99 (Doren, 2004; Roberts, Doren, & Thornton, 2002), which includes items from the domains of sexual deviancy as well as general

antisociality. The method of combining results from scales sampling different domains must therefore also consider (1) what are the domain(s) being assessed by the scales and (2) how each of the domains are related to the outcome of interest (i.e., recidivism). In the current study, the RRASOR added incrementally to Static-99R, but in different directions depending on the recidivism type (i.e., positive incremental validity to Static-99R for sexual recidivism, but negative incremental validity for violent and any recidivism). The negative relationship of the RRASOR to violent and any recidivism suggests that subtracting the RRASOR from Static-99R would be a better method of combination than averaging. For sexual recidivism, however, where both scales add incrementally with positive weights, it is possible that an approach that adds or averages the scales together would be more accurate.

In summary, the method used to combine findings from risk scales assessing different domains necessitates the identification of what the scales are actually measuring. This, however, is not an obvious task. Despite all the items of the RRASOR being included in Static-99R, the two scales had opposite relationships with violent recidivism, once the other scales were controlled for. Consequently, it can be assumed (albeit post-hoc) that the two scales are sampling different domains. Identifying the constructs being measured requires both theory and empirical evidence; without such evidence, reliability between assessors concerning the latent constructs would be expected to be low.

Implications for Researchers

We believe the results of the current study should motivate further consideration of construct validity in the development of empirical risk prediction tools. Although it is possible to address the problem of combining multiple measures without understanding what they are measuring, a pure prediction approach to this problem has considerable limitations. Vrieze and Grove (2010), for example, have proposed creating a superscale, with existing scales treated as items in the superscale. Although such an approach is logically consistent, it is inefficient and impractical. Specifically, such a superscale would require all the same steps required when creating any new scale, such as generating a scoring manual and completing cross-validation. Given that many of the individual scales have identical or nearly identical items, evaluators would soon tire of the repetition and quickly look for ways of combining items rather than the total scores of diverse measures.

We believe that future research on risk assessment should focus on identifying and assessing the psychologically meaningful characteristics associated with recidivism (Mann, Hanson, & Thornton, 2010). For example, a single dimension or propensity (e.g., antisociality) would be composed of and influenced by several markers (e.g., unemployment, substance abuse, history of criminal behaviour, procriminal attitudes). Once valid measures of the core constructs have been assessed, researchers can examine the independent contribution of these dimensions. Following dimensional theory (Loftus, Oberg, & Dillon, 2004) risk factors could be weighted at the construct level (e.g., antisociality, sexual deviancy) and the weight allocated to each construct can depend on the type of recidivism being predicted (e.g., violence vs. sexual).

One advantage of such a conceptual actuarial measure would be that the subcomponents are defined and, consequently, evaluators could identify the reasons for an offender's score. Understanding what items are measuring would allow evaluators to explain inconsistencies in risk rating across measures, thereby helping inform the method of combining multiple risk scales. This task would, however, be difficult as it requires not only an understanding of the underlying constructs, but knowledge of how the specific items measure these constructs. Nevertheless, this type of task is essential given that the incremental addition of scales is most likely not limited to the three actuarial scales examined in this study. If the scales are created using a purely predictive approach, risk evaluators will continue to be faced with the knowledge that other variables (and scales) add incremental validity without being able to explain why. The direction forward for risk assessment combines empirical prediction with the construct validity tradition.

Implications for Current Practice

The current study did not find a clear superiority for either Static-99R or Static-2002R for the prediction of sexual, violent, or general recidivism (both scales were, however, superior to the RRASOR). Consequently, evaluators choosing between them would need to consider other criteria. For example, evaluators interested in estimating absolute recidivism rates may prefer Static-99R over Static-2002R because of the relatively large normative samples available (Helmus, 2009). For other assessments, Static-2002R may be preferable to Static-99R because the items are grouped into subscales (i.e., age, sexual deviancy, general criminality) that suggest the source of the risk. In high stakes situations, evaluators may want to use both measures: both scales add incremental validity to one another, with recidivism rates of discordant cases being intermediate between the rates suggested by the individual scales.

For violent recidivism, both Static-99R and Static-2002R can be used. Risk evaluators should be aware, however, that the item weighting of these scales is not optimal for the prediction of violent recidivism (i.e., too much weight allocated to items assessing sexual deviancy). As such, if the evaluation is primarily concerned with violent recidivism, we recommend scales designed for that purpose (e.g., VRAG, SORAG – Quinsey et al., 2006; Risk Matrix-2000v and Risk Matrix-2000c – Thornton et al., 2003). These measures have stronger weights for general criminality than the RRASOR, Static-99R, and Static-2002R.

In the current study, we presented the prediction weights (standardized regression coefficients from the Cox regression analyses) of the RRASOR, Static-99R, and Static-2002R for illustrative purposes only. We do not advocate the use of these weights in applied practice because they would likely be affected by overfitting (Cureton, 1950). Without further replication studies (with large sample sizes), the extent to which the weights found are accurate and generalizable is unknown.

In summary, for evaluators who select scales that measure similar domains of risk factors (e.g., Static-99R and Static-2002R), then it is likely that an averaging approach would be the optimal method of combining the findings of the multiple scales. Such an approach follows classical test theory, in that a greater number of items measuring the same construct pool (and having similar predictive accuracy) reduces measurement error and increases predictive accuracy. Consequently, concordance among scales would increase evaluators' confidence in the accuracy of the risk assessment. In contrast, if the selected scales are not sampling the same latent constructs, then the evaluators would require a defensible model concerning (1) the latent constructs measured by the scales, (2) how the domains relate to the outcome of interest, and (3) empirical evidence concerning how the constructs should be weighted and combined. In the absence of such an empirically supported model, it would be prudent for evaluators to privilege the scale for which the evaluator holds the most confidence.

## References

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


*Allan, M., Grace, R. C., Rutherford, B., & Hudson, S. M. (2007). Psychometric assessment of dynamic risk factors for child molesters. *Sexual Abuse: A Journal of Research and Treatment, 19,* 347-367. doi: 10.1007/s11194-007-9052-5

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data.* Beverly Hills, CA: Sage.

Anderson, D. & Hanson, R. K. (2010). Static-99: An actuarial tool to assess risk of sexual and violent recidivism among sexual offenders. In R. K. Otto & K. S. Douglas *Handbook of Violence Risk Assessment* (pp. 251-267). New York: Taylor & Francis Group.

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

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84-94. doi: 10.1207/s15327752jpa8701_07

Avis, J. M., Kudisch, J. D., & Fortunato, V. J. (2002). Examining the incremental validity and adverse impact of cognitive ability and conscientiousness on job performance. *Journal of Business and Psychology, 17*, 87-105. doi: 0889-3268/02/0900-0087/0

Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Moving beyond "low", "moderate", and "high" in risk communication. *Crime Scene, 16*(1), 11-14. Available from http://www.cjsw.ac.uk/cjsw/files/Hanson%202009.pdf

Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006). Different actuarial risk measures produce different risk rankings for sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 18*, 423–440. doi: 10.1007/s11194-006-9029-9

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior, 28*, 490-521. doi: 10.1177/009385480102800406

*Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime & Law, 14,* 85-106. doi: 10.1080/10683160701483104

*Bigras, J. (2007). La prédiction de la récidive chez les délinquants sexuels [Prediction of recidivism among sex offenders]. *Dissertation Abstracts International : Section B, 68* (09). (UMI No. NR30941).

*Boer, A. (2003). *Evaluating the Static-99 and Static-2002 risk scales using Canadian sexual offenders* (Unpublished master's thesis). University of Leicester, United Kingdom.

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

*Bonta, J., & Yessine, A. K. (2005). [Recidivism data for 124 released sexual offenders from the offenders identified in *The National Flagging System: Identifying and responding to high-risk, violent offenders* (User Report 2005-04). Ottawa, ON: Public Safety and Emergency Preparedness Canada]. Unpublished raw data.

Borum, R., Bartel, P., Forth, A. (2002). *Manual for the Structured Assessment of Violence Risk in Youth (SAVRY)*. Tampa, FL: University of South Florida.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessments: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, 50, 196-209. doi: 10.1037/a0016066

*Brouillette-Alarie, S., & Proulx, J. (2008, October). *Predictive and convergent validity of phallometric assessment in relation to sexual recidivism risk.* Poster presented at the annual conference for the Association for the Treatment of Sexual Abusers, Atlanta, GA.

Cella, D. F., & Bonomi, A. E., Lloyd, S. R., Tulsky, D. S., Kaplan, E., & Bonomi, P. (1995). Reliability and validity of the Functional of Cancer Therapy- Lung (FACT_L) quality of life instrument. *Lung Cancer, 12*, 199-220.

Cohen, J. (1994). The earth is round ($p < .05$). *The American Psychologist, 49*, 997-1003.

*Cortoni, F., & Nunes, K. L. (2007). *Assessing the effectiveness of the National Sexual Offender Program* (Research Report No. R-183). Ottawa, ON: Correctional Service of Canada. Retrieved from http://www.csc-scc.gc.ca/text/rsrch/reports/r183/r183-eng.shtml

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10*, 94-96.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

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

Doren, D. (2004). Toward a multidimensional model for sexual recidivism risk. *Journal of Interpersonal Violence, 19*, 835-856. doi: 10.1177/0886260504266882

Douglas, K., & Kropp, P. R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior, 29*, 617-658. doi: 10.1177/009385402236735

*Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2009). [*Data from sex offenders released from prison in Austria*]. Unpublished raw data.

*Epperson, D. L. (2003). *Validation of the MnSOST-R, Static-99, and RRASOR with North Dakota prison and probation samples.* Unpublished Technical Assistance Report, North Dakota Division of Parole and Probation.

Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Goldman, R., & Alexander, W. (1998). *Minnesota Sex Offender Screening Tool-Revised (MnSOST-R).* St. Paul, MN: Minnesota Department of Corrections.

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Hare Psychopathy Checklist: Youth Version.* North Tonawanda, NY: Multi-Health Systems.

Garb, H. N. (2003). Clinical judgment and mechanical prediction. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Volume 10, Assessment psychology* (pp. 27-42). Hoboken, NJ: John Wiley & Sons, Inc.

Gendreau, P., Goggin, C., & Law, M. (1997). Predicting prison misconduct. *Criminal Justice & Behavior, 24*, 414-431. doi: 10.1177/0093854897024004002

Grove, W. M. Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical

prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30. doi: 10.1037/1040-3590.12.1.19

*Haag, A. M. (2005). [recidivism data from 198 offenders detained until their warrant expiry date. From: Do psychological interventions impact on actuarial measures: An analysis of the predictive validity of the Static-99 and Static-2002 on a re-conviction measure of sexual recidivism. *Dissertation Abstracts International, Section B, 66* (08), 4531]. Unpublished raw data.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the Areas under ROC curves derived from same cases. *Radiology, 148*, 839-843.

Hanson, R. K. (1997). *The development of a brief actuarial scale for sex offender recidivism* (User Report No. 1997-04). Ottawa, ON: Department of the Solicitor General of Canada. Available from http://www.defenseforsvp.com/Resources/Hanson_Static-99/RRASOR.pdf

Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence, 20*, 212-217. doi: 10.1177/0886260504267740

Hanson, R. K. (2008). What statistics should we use to report predictive accuracy. *Crime Scene, 15*(1), 15-17. Available from http://www.cpa.ca/cpasite/userfiles/Documents/Criminal%20Justice/Crime%20Scene%202008-04.pdf

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology, 20*, 172-182. doi: 10.1037/a0015726

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.

*Hanson, R. K., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (Corrections Research User Report No. 2007-05). Ottawa, ON: Public Safety Canada. Available from http://www.publicsafety.gc.ca/res/cor/rep/_fl/crp2007-05-en.pdf

Hanson, R. K., Helmus, L., & Thornton, D. (2010). Predicting recidivism among sexual offenders: A multi-site study of Static-2002. *Law and Human Behavior, 34*, 198-211. doi: 10.1007/s10979-009-9180-1

Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2010). *Using multiple samples to estimate relative risk for actuarial risk tools: A Canadian example using Static-99 and Static-2002*. Unpublished manuscript.

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

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:1005482921 333.

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of Static-2002*. (Corrections Research User Report No. 2003-01). Ottawa, ON: Department of the Solicitor General of Canada. Available from http://www.publicsafety.gc.ca/res/cor/rep/2003-01-not-sttc-eng.aspx

*Harkins, L., & Beech, A.R. (2007). *Examining the effectiveness of sexual offender treatment using risk band analysis*. Unpublished manuscript.

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

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

Ottawa, ON: Department of the Solicitor General of Canada. Available from http://www.publicsafety.gc.ca/res/cor/rep/_fl/2003-03-stc-cde-eng.pdf

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumière, M. L., Boer, D., & Lang, C. (2003). A multi-site comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15*, 413-425. doi: 10.1037/1040-3590.15.3.413

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *Manual for the Hare Psychopathy Checklist-Revised: Screening Version (PCL:SV)*. Toronto, ON: Multi-Health Systems.

Hedges, L. V. (1994). Fixed effect models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-299). New York: Russell Sage.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504. doi: 10.1037/1082-989X.3.4.486

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* (master's thesis). Available from ProQuest Dissertation and Theses database. (UMI No. MR58443). Also available from www.static99.org

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2010). *Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights*. Unpublished manuscript.

*Hill, A., Habermann, N., Klusmann, D., Berner, W., & Briken, P. (2008). Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy and Comparative Criminology, 52*, 5-20. doi: 10.1177/0306624X07307450

Hoge, R. D., & Andrews, D. A. (2002). *The Youth Level of Service/Case Management Inventory manual and scoring key*. Toronto, ON: Multi-Health Systems.

Humphreys, L. G., & Swets, J. A. (1991). Comparison of predictive validities measured with biserial correlations and ROCs of signal detection theory. *Journal of Applied Psychology, 76*, 316-321.doi: 10.1037/0021-9010.76.2.316

Interstate Commission for Adult Offender Supervision. (2007). *Sex offender assessment information survey* (ICAOS Documents No. 4-2007). Lexington, KY: Author.

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts**.** *Sexual Abuse: A Journal of Research and Treatment, 19*, 409-448. doi: 10.1007/s11194-007-9062-3

*Johansen, S. H. (2007). Accuracy of predictions of sexual offense recidivism: A comparison of actuarial and clinical methods. *Dissertation Abstracts International, Section B, 68* (03). (UMI No. 3255527).

*Knight, R. A., & Thornton, D. (2007). *Evaluating and improving risk assessment schemes for sexual recidivism: A long-term follow-up of convicted sexual offenders* (Document No. 217618). Submitted to the U.S. Department of Justice.

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

*Langström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: A Journal of Research and Treatment, 16*, 107-120. doi: 10.1177/107906320401600202

Lloyd, M. D. (2008). Incremental validity of commonly-used risk assessment measures in predicting serious sexual recidivism. *Dissertation Abstracts International: Section B. The Sciences and*

*Engineering, 69*(9), 5784.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review, 111,* 835-863. doi: 10.1037/0033-295X.111.4.835

Malloy, P. F., Cummings, J. L., Coffey, C. E., Duffy, J., Fink, M., Lauterbach, E. C., Lovell, M., Royall, D., & Salloway, S. (1997). Cognitive screening instruments in neuropsychiatry: A report of the Committee on Research of the American Neuropsychiatric Association. *The Journal of Neuropsychiatry and Clinical Neurosciences, 9,* 189-197.

Mann, R. E., Hanson, R. K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. *Sexual Abuse: A Journal of Research and Treatment, 22,* 191-217. doi: 10.1177/1079063210366039

McGrath, R. J., Cumming, G. F., & Burchard, B. L., Zeoli, S., & Ellerby, L. (2010). *Current practices and emerging trends in sexual abuser management: The Safer Society 2009 North American Survey* (ISBN: 978-1-884444-85-2)**.** Brandon, VT: Safer Society Press.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1956). Wanted-A good cook-book. *American Psychologist, 11,* 263-272. doi: 10.1037/h0044164

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

Monahan, J., & Walker, L. (2010). *Social science in law: Cases and materials*. New York: Foundation Press.

*Nicholaichuk, T. (2001, November). *The comparison of two standardized risk assessment instruments in a sample of Canadian Aboriginal sexual offenders.* Paper presented at the annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, San Antonio, TX.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

Phenix, A., Doren, D., Helmus, L., Hanson, R. K., & Thornton, D. (2009) *Coding rules for Static-2002.* Ottawa, ON: Public Safety Canada. Available from http://www.publicsafety.gc.ca/res/cor/rep/sttc-2002-eng.aspx

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd Ed.). Washington, DC: American Psychological Association.

Rettenberger, M., Matthes, A., Boer, D. P., Eher, R. (2010). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology, 54,* 169-186. doi: 10.1177/0306624X08328755

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

Roberts, C. F., Doren, D. M., & Thornton, D. (2002). Dimensions associated with assessments of sex offender recidivism risk. *Criminal Justice and Behavior, 29,* 569-589. *doi:* 10.1177/009385402236733

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Taylor and Francis.

Schmidt, F. L., Oh, I., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97-128. doi: 10.1348/000711007X255327

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17,* 156-167. doi: 10.1037/1040-3590.17.2.156

Sledjeski, E. M., Dierker, L. C., Costello, D., Shiffman, S., Donny, E., & Flay, B. R. (2007). Predictive validity of four nicotine dependence measures in a college sample. *Drug and Alcohol Dependence, 87,* 10-19. doi: 10.1016/j.drugalcdep.2006.07.005

Stalans, L. J., Hacker, R., & Talbot, M. E. (2010). Comparing non-violent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behaviour, 37,* 613-628. doi: 10.1177/0093854810363794

*Swinburne Romine, R., Dwyer, S. M., Mathiowetz, C., & Thomas, M. (2008, October). *Thirty years of sex offender specific treatment: A follow-up study*. Poster presented at the conference for the Association for the Treatment of Sexual Abusers, Atlanta, GA.

*Ternowski, D. R. (2004). Sex offender treatment: An evaluation of the Stave Lake Correctional Centre Program. *Dissertation Abstracts International: Section B, 66* (06), 3428.

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment, 14,* 139-153. doi: 10.1177/107906320201400205

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

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40,* 525-531. doi: 10.1037/a0014693

Vrieze, S. I., & Grove, W. M. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological Assessment, 22,* 382-395. doi: 10.1037/a0019228

Weiner. I. B. (2003). The assessment process. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Volume 10, Assessment psychology* (pp. 3-26). Hoboken, NJ: John Wiley & Sons, Inc.

Welsh J. L., Schmidt F., McKinnon L., Chattha H. K., & Meyers J. R. (2008). A comparative study of adolescent risk assessment instruments: Predictive and incremental validity. *Assessment, 15,* 104-115. doi: 10.1177/1073191107307966

*Wilson, R. J., Cortoni, F., & Vermani, M. (2007a). *Circles of support and accountability: A national replication of outcome findings* (Report No. R-185). Ottawa, ON: Correctional Service of Canada. Available from http://www.csc-scc.gc.ca/text/rsrch/reports/r185/r185-eng.shtml

*Wilson, R. J., Picheca, J. E., & Prinzo, M. (2007b). Evaluating the effectiveness of professionally-facilitated volunteerism in the community-based management of high-risk sexual offenders: Part two – A comparison of recidivism rates. *The Howard Journal, 46,* 327-337. doi: 10.1111/j.1468-2311.2007.00480.x

## Appendix A

Table 1A. *ROC Areas for the RRASOR, Static-99R, and Static-2002R by Sample*

| Sample | *N* | Sexual Recidivism | | | Violent Recidivism | | | Any Recidivism | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ROC | 95% CI | | ROC | 95% CI | | ROC | 95% CI | |
| Allan et al. (2007) | 492 | | | | | | | | | |
| RRASOR | | **.70** | .62 | .78 | **.60** | .53 | .67 | **.57** | .51 | .63 |
| Static-99R | | **.72** | .64 | .80 | **.69** | .63 | .75 | **.70** | .65 | .75 |
| Bengtson (2008) | 308 | | | | | | | | | |
| RRASOR | | **.61** | .54 | .68 | **.60** | .54 | .67 | **.59** | .52 | .66 |
| Static-99R | | **.62** | .56 | .68 | **.66** | .60 | .72 | **.64** | .57 | .70 |
| Static-2002R | | **.64** | .57 | .70 | **.66** | .60 | .72 | **.67** | .60 | .73 |
| Bigras (2007) | 457 | | | | | | | | | |
| RRASOR | | **.60** | .48 | .72 | **.54** | .47 | .62 | **.55** | .48 | .61 |
| Static-99R | | **.71** | .60 | .82 | **.69** | .62 | .75 | **.69** | .64 | .75 |
| Static-2002R | | **.70** | .59 | .81 | **.69** | .63 | .75 | **.71** | .65 | .76 |
| Boer (2003) | 296 | | | | | | | | | |
| RRASOR | | **.69** | .57 | .81 | **.62** | .55 | .70 | **.60** | .53 | .66 |
| Static-99R | | **.75** | .65 | .85 | **.75** | .69 | .81 | **.79** | .74 | .84 |
| Static-2002R | | **.73** | .63 | .83 | **.74** | .67 | .80 | **.81** | .76 | .86 |
| Bonta & Yessine (2005) | 133 | | | | | | | | | |
| RRASOR | | **.50** | .36 | .64 | **.47** | .36 | .57 | **.48** | .38 | .58 |
| Static-99R | | **.64** | .52 | .77 | **.66** | .57 | .76 | **.65** | .56 | .74 |
| Brouillette-Alarie & Proulx (2008) | 228 | | | | | | | | | |
| RRASOR | | **.67** | .59 | .75 | **.60** | .53 | .68 | - | - | - |
| Static-99R | | **.68** | .60 | .76 | **.69** | .62 | .76 | - | - | - |
| Cortoni & Nunes (2007)[a] | 73 | | | | | | | | | |
| RRASOR | | - | - | - | **.73** | .56 | .90 | **.64** | .49 | .80 |
| Static-99R | | - | - | - | **.71** | .53 | .89 | **.70** | .54 | .87 |
| Eher et al. (2008) | 706 | | | | | | | | | |
| RRASOR | | **.75** | .65 | .85 | **.68** | .63 | .74 | **.63** | .58 | .68 |
| Static-99R | | **.71** | .61 | .82 | **.76** | .71 | .80 | **.71** | .67 | .75 |
| Epperson (2003) | 177 | | | | | | | | | |
| RRASOR | | **.73** | .62 | .84 | - | - | - | - | - | - |
| Static-99R | | **.78** | .67 | .88 | - | - | - | - | - | - |
| Haag (2005) | 190 | | | | | | | | | |
| RRASOR | | **.64** | .55 | .74 | - | - | - | - | - | - |
| Static-99R | | **.70** | .61 | .78 | - | - | - | - | - | - |
| Static-2002R | | **.67** | .58 | .76 | - | - | - | - | - | - |

(Table continues)

Table 1A. Continued.

| Sample | *N* | Sexual Recidivism ROC | 95% CI | | Violent Recidivism ROC | 95% CI | | Any Recidivism ROC | 95% CI | |
|---|---|---|---|---|---|---|---|---|---|---|
| Hanson et al. (2007) | 702 | | | | | | | | | |
| RRASOR | | **.70** | .64 | .77 | **.63** | .57 | .68 | **.61** | .56 | .66 |
| Static-99R | | **.76** | .69 | .82 | **.75** | .70 | .80 | **.76** | .72 | .80 |
| Static-2002R | | **.75** | .69 | .82 | **.76** | .72 | .81 | **.76** | .73 | .80 |
| Harkins & Beech (2007) | 190 | | | | | | | | | |
| RRASOR | | **.73** | .64 | .83 | **.66** | .56 | .75 | **.62** | .54 | .71 |
| Static-99R | | **.78** | .68 | .88 | **.75** | .67 | .84 | **.77** | .69 | .84 |
| Static-2002R | | **.79** | .68 | .89 | **.77** | .69 | .85 | **.77** | .70 | .84 |
| Hill et al. (2008) | 86 | | | | | | | | | |
| RRASOR | | **.61** | .46 | .76 | **.58** | .45 | .71 | **.53** | .40 | .65 |
| Static-99R | | **.67** | .52 | .82 | **.62** | .50 | .75 | **.61** | .48 | .74 |
| Johansen (2007) | 273 | | | | | | | | | |
| RRASOR | | **.63** | .52 | .75 | **.61** | .53 | .69 | **.57** | .50 | .64 |
| Static-99R | | **.65** | .53 | .77 | **.71** | .63 | .79 | **.72** | .66 | .78 |
| Knight & Thornton (2007) | 466 | | | | | | | | | |
| RRASOR | | **.62** | .56 | .68 | **.59** | .53 | .64 | **.56** | .51 | .61 |
| Static-99R | | **.62** | .56 | .67 | **.64** | .59 | .69 | **.63** | .58 | .68 |
| Static-2002R | | **.63** | .58 | .69 | **.64** | .59 | .69 | **.64** | .59 | .69 |
| Långström (2004) | 1,278 | | | | | | | | | |
| RRASOR | | **.72** | .66 | .78 | **.65** | .61 | .68 | - | - | - |
| Static-99R | | **.73** | .68 | .79 | **.78** | .75 | .81 | - | - | - |
| Nicholaichuk (2001) | 281 | | | | | | | | | |
| RRASOR | | **.67** | .60 | .75 | - | - | - | - | - | - |
| Static-99R | | **.74** | .67 | .81 | - | - | - | - | - | - |
| Swinburne Romine et al. (2008) | 680 | | | | | | | | | |
| RRASOR | | **.62** | .55 | .68 | - | - | - | - | - | - |
| Static-99R | | **.63** | .57 | .70 | - | - | - | - | - | - |
| Ternowski (2004) | 247 | | | | | | | | | |
| RRASOR | | **.61** | .48 | .74 | **.60** | .50 | .70 | **.61** | .52 | .69 |
| Static-99R | | **.75** | .66 | .85 | **.76** | .69 | .84 | **.77** | .70 | .84 |
| Wilson et al. (2007 a & b) | 228 | | | | | | | | | |
| RRASOR | | **.63** | .51 | .75 | **.52** | .43 | .60 | **.50** | .42 | .58 |
| Static-99R | | **.57** | .43 | .71 | **.62** | .54 | .70 | **.60** | .52 | .68 |

*Note.* Sexual recidivism rates from all cases, not controlling for length of follow-up used to compute AUC values.
[a] No AUC value for sexual recidivism could be computed for Cortoni and Nunes (2007) because they were no recidivists

## Appendix B

Table 1B. *Relative Risk Ratios of Static-99R and Static-2002R*

| | | Static-99R | | Static-2002R | |
|---|---|---|---|---|---|
| | | **Score** | **RR** | **Score** | **RR** |
| **Low** | **< 0.50** | | | -2 | 0.11 |
| | | -3 | 0.26 | -1 | 0.17 |
| | | -2 | 0.34 | 0 | 0.26 |
| | | -1 | 0.45 | 1 | 0.38 |
| **Moderate** | **0.50 – 2.00** | 0 | 0.59 | 2 | 0.54 |
| | | 1 | 0.77 | 3 | 0.74 |
| | | 2 | 1.00 | 4 | 1.00 |
| | | 3 | 1.31 | 5 | 1.31 |
| | | 4 | 1.71 | 6 | 1.68 |
| **High** | **≥ 2.00** | 5 | 2.23 | 7 | 2.08 |
| | | 6 | 2.91 | 8 | 2.52 |
| | | 7 | 3.80 | 9 | 2.97 |
| | | 8 | 4.96 | 10 | 3.40 |
| | | 9 | 6.48 | 11+ | 3.79 |
| | | 10+ | 8.47 | | |

*Note*. RR = Relative risk. Relative risk for Static-99R calculated in a previous study (Hanson et al., 2010) based on Cox regression coefficients derived from entering Static-99R scores ($B$ = 0.267; $SE$ = 0.013; Wald = 413.11; $p$ <.001), with sample as strata ($k$ = 22, $n$ = 8,047). Relative risk for Static-2002R calculated in a previous study (Hanson et al., 2010) based on Cox regression coefficients derived from entering Static-2002R scores ($\beta$ = 0.285; $SE$ = 0.033; Wald = 74.24; $p$ < .001) and squared Static-2002R scores ($\beta$ = -.013; $SE$ = 0.006; Wald = 4.66; $p$ = .031), with sample as strata ($k$ = 7, $n$ = 2,610).