Even Highly Correlated Measures Can Add Incrementally to Predicting Recidivism Among Sex Offenders

Assessment
19(4) 442–461

DOI: 10.1177/1073191112458312
http://asm.sagepub.com
SAGE

Kelly M. Babchishin[1,2], R. Karl Hanson[1], and Leslie Helmus[2]

Abstract

Criterion-referenced measures, such as those used in the assessment of crime and violence, prioritize predictive accuracy (discrimination) at the expense of construct validity. In this article, we compared the discrimination and incremental validity of three commonly used criterion-referenced measures for sex offenders (Rapid Risk Assessment for Sex Offence Recidivism [RRASOR], Static-99R, and Static-2002R). In a meta-analysis of 20 samples (n = 7,491), Static-99R and Static-2002R provided similar discrimination but outperformed the RRASOR in the prediction of sexual, violent, and any recidivism. Remarkably, despite large correlations between them (rs ranging from .70 to .92), these risk scales consistently added incremental validity to one another. 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 any recidivism. We also examined different methods of combining risk scales and found that the averaging approach produced better discrimination than choosing the highest score and produced better calibration than either choosing the lowest or highest risk score. The findings reinforce the importance of understanding the psychological content of criterion-referenced measures, even when the sole purpose is to predict a particular outcome and provide some direction concerning the best methods for combining risk scales.

Keywords

risk assessment, incremental validity, predictive accuracy, calibration, sex offenders

Predictions about future behavior have scientific merit to the extent that they are supported by empirical evidence (e.g., Meehl, 1954). This evidence is often expressed in the form of statements such as "previous studies have found that 45% of individuals with similar features to Mr. X will commit a violent act within two years." To make such predictions, it is possible and often useful to construct criterion-referenced measures that are driven almost exclusively by the empirical relationships between the items and the outcome of interest. Such empirical actuarial measures are increasingly being used in certain fields of psychology, and particularly in the assessment of risk for crime and violence (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Hanson, 2005).

Sex offender risk assessments often involve high stakes for both the offender and the general public, which has motivated the development of specialized risk scales. There are at least nine risk assessment scales specifically designed to predict recidivism for adult male sex offenders that have demonstrated moderate or high predictive accuracy (d > .50; Hanson & Morton-Bourgon, 2009). Among sex offender treatment providers, the most commonly used measure in the United States and Canada is Static-99 (Hanson & Thornton, 2000), followed by Rapid Risk Assessment for Sex Offence Recidivism (RRASOR; Hanson, 1997) in the United States, and Static-2002 (Hanson & Thornton, 2003) in Canada (McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010).

These measures (Static-99, Static-2002, and RRASOR) were created by the same research team using similar approaches to scale development. The initial pool of items was restricted to simple, easily scored, static items (e.g., prior offences, victim type) that predicted sexual recidivism in previous studies. As with many criterion-referenced measures, items were retained in the final measure if the item, or a meaningful subset of items, incrementally improved the predictive accuracy of the total scores. The Static-99

[1]Public Safety Canada, Ottawa, Ontario, Canada
[2]Carleton University, Ottawa, Ontario, Canada

Corresponding Author:
Kelly M. Babchishin, Corrections Research, Public Safety Canada, 340 Laurier Avenue West, Ottawa, Ontario, Canada, K1A 0P8
Email: Kelly_Babchishin@carleton.ca

and Static-2002 items were intended to cover the broad domains of risk factors previously identified in meta-analyses: age, general criminality, and sexual criminality (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005). Certain items of these scales, however, were retained based exclusively on their ability to predict recidivism, even when the reasons for the association with recidivism were unknown to the test developers (e.g., unrelated victims; see Hanson & Thornton, 2003).

Given the existence of multiple risk assessment scales, those involved in assessing sex offenders must choose whether to use one or multiple measures. Clinical practice has favored the use of multiple risk scales (Jackson & Hess, 2007). Research, however, has yet to provide clear direction on (a) the extent to which multiple risk scales improve predictive accuracy or (b) how to combine several risk scales into one overall judgment of risk.

The average predictive accuracy of the various risk scales are similar (Hanson & Morton-Bourgon, 2009), which is not surprising given the substantial overlap in item content. What is surprising is that different actuarial risk scales produce different estimates of risk (e.g., Barbaree, Langton, & Peacock, 2006a; Mills & Kroner, 2006). Evaluators have attempted to reconcile the divergent estimates of different risk tools by suggesting that they are saturated in different degrees by the two broad domains associated with recidivism risk: general criminality and sexual criminality (Barbaree, Langton, & Peacock, 2006b; Doren, 2002, 2004). Although informative, this does not answer the fundamental question of how to combine multiple actuarial scales. Certain approaches have been proposed, including believing the highest risk assessment, believing the lowest risk assessment, and averaging (Barbaree et al., 2006a; Doren, 2002; Seto, 2005; Vrieze & Grove, 2010). Few have been empirically evaluated and none have gained widespread acceptance. In practice, evaluators using multiple risk instruments tend to report each result independently from the other (as opposed to integrating the risk decisions of multiple risk instruments into one succinct judgment) or tend to believe the highest risk instrument (Jackson & Hess, 2007). Importantly, the optimal method of combining risk scales first requires examining the extent to which the scales provide unique information (i.e., incremental validity) in the prediction of the outcome of interest.

That the addition of new items or domains of risk factors add incrementally to the prediction of violence or crime is uncontroversial (e.g., Walters, 2011). In contrast, anticipating the incremental contribution of different criterion-referenced measures is difficult. Actuarial risk scales are rarely homogeneous because predictive accuracy increases as the number of distinct, risk-relevant domains increases. Furthermore, different domains may have different relationships with the outcome of interest. As such, when the variance of one of the latent constructs (e.g., sexual criminality) is statistically controlled for by the introduction of a second scale measuring the same construct (sexual deviance), the unique variance of a second latent construct (e.g., general criminality) becomes more evident. In this example, the relationship between the original scale and the outcome of interest now represents the unique variance derived from the items assessing general criminality. Consequently, the relationship between the total score (controlling for the other scale) and the outcome of interest can be different from the original, bivariate relationship. When the scale is saturated with different constructs, the relationship between a total score and the outcome can disappear, improve, or even reverse directions (also known as "suppression" effects) when other scales are added (Horst, 1941; Paulhus, Robins, Trzesniewski, & Tracy, 2004; Tzelgov & Henik, 1991). Despite having a negligible relationship with the outcome of interest, a scale can add incrementally to another by suppressing the irrelevant variance (e.g., measurement error), therefore allowing for a better prediction of the outcome of interest (Horst, 1941). Consequently, clinical inferences concerning the potential incremental contribution of different criterion-referenced measures need to be empirically validated.

The few studies that have directly examined the incremental contribution of different unmodified (i.e., no items removed) risk scales have yielded mixed results. Seto (2005) found no increase in predictive accuracy for multiple scales compared with one scale; the sample size, however, was too small ($N = 215$) to make strong statements about null hypotheses. Other researchers have found statistically significant incremental contributions to the prediction of sexual recidivism using information from more than one scale (Lloyd, 2008; Welsh, Schmidt, McKinnon, Chattha, & Meyers, 2008). Further research with larger samples is required to better understand whether there is practical utility in using several actuarial scales in risk assessments.

Current Study

In this article, we compared the predictive and incremental validity of the total scores of three commonly used risk scales (RRASOR, Static-99R, and Static-2002R) and examined methods of combining scales. The items of Static-99R and Static-2002R are identical to their original scales, with the exception of updated age weights, which were adjusted to improve the calibration (i.e., agreement between the observed and predicted recidivism rates) of these measures with aging populations of sex offenders (see Helmus, Thornton, Hanson, & Babchishin, 2012). Static-99R was selected because it is the current version of the most frequently used scale for the evaluation of risk among sex offenders (Jackson & Hess, 2007; McGrath et al., 2010). RRASOR and Static-2002R were selected because (a) they

Table 1. Items Contained in the RRASOR (Rapid Risk Assessment for Sex Offence Recidivism), Static-99R, and Static-2002R

RRASOR	Static-99/Static-99R	Static-2002/Static-2002R
Offender's age at release[a]	Offender's age at release[a]	Offender's age at release[a]
Number of prior sexual offence charges and convictions[b]	Number of prior sexual offence charges and convictions[b]	Prior sentencing occasions for sexual offences[b]
Any unrelated victims of sexual assaults[c]	Any unrelated victims of sexual assaults[c]	Any unrelated victims of sexual assaults[c]
Any male victims of sexual assaults[c]	Any male victims of sexual assaults[c]	Any male victims of sexual assaults[c]
	Convictions for noncontact sexual offences[d]	Convictions for noncontact sexual offences[d]
	Any stranger victims of sexual assaults[d]	Any stranger victims of sexual assaults[d]
	Number of prior sentencing dates[a]	Prior sentencing occasions for anything[a]
	Conviction for nonsexual violence prior to the Index Offence[e]	Prior violent nonsexual sentencing occasion[e]
	Conviction for nonsexual violence at the time of the Index Offence[f]	Any prior involvement with the criminal justice system[f]
	Ever lived with an intimate partner for two consecutive years[f]	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[f]

Note. Adapted from 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).

are the next most commonly used scales and (b) they closely resemble Static-99R in purpose and content. Additionally, a large number of validation studies for these scales allowed us to obtain numerous data sets for this study.

Previous research has suggested that these measures assign different relative weight to the different domains associated with sexual recidivism risk: namely, general criminality, sexual criminality, and age. Whereas these three factors have relatively equal weight in Static-99R and Static-200R (Thornton, 2010), RRASOR predominantly measures sexual criminality and age (e.g., Barbaree et al., 2006b). The specific items of these scales are listed in Table 1.

Based on a reanalysis of 20 samples (7,491 sex offenders), we examined (a) whether the RRASOR, Static-99R, or Static-2002R total scores predicted sexual, violent, and any recidivism more accurately than the others and (b) whether the total scores of the three scales added incremental validity to one another in the prediction of the three types of recidivism. If one of the scales 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" risk scale (as per Seto's [2005] recommendation). The choice of risk scales would be less clear, however, if none of the measures had superior discrimination or if they were found to add incrementally to one another. As such, a secondary purpose of the current study was to examine the merits of three commonly proposed methods of combining risk scales: (a) choosing the lowest score, (b) choosing the highest score, and, (c) taking an average.

Method

Measures

Rapid Risk Assessment for Sex Offence Recidivism. The RRASOR (Hanson, 1997) is an actuarial scale designed to measure risk of sexual recidivism among adult male sex offenders. Scores range from 0 to 6, with a higher score indicating greater risk of sexual recidivism. The RRASOR

is a subset of four items of Static-99 and, for the purpose of 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 additional offences for prior sexual offences. Pseudo-recidivism occurs when offenders have additional charges laid against them for crimes they committed before they were apprehended for the current offence. In RRASOR, these new charges are coded as recidivism (or in the case of priors, as separate offences) whereas in Static-99 they are included in the current offence (or clustered with other 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 rules of Static-99 rather than RRASOR is expected to be minimal.

In the sample used to create the RRASOR, it was found to differentiate sexual recidivists from nonrecidivists, AUC (area under the curve) = 0.71 (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 $d = .60$, 95% confidence interval [CI] = [0.54, 0.65], $N = 11{,}031$, $k = 34$; this translates to an AUC of 0.66, 95% CI = [0.65, 0.68]).

Static-99R. Static-99R (Hanson & Thornton, 2000; Helmus, Thornton, et al., 2012) is a 10-item actuarial scale that assesses recidivism risk of adult male sex offenders. The items and scoring rules are identical to Static-99 (Hanson & Thornton, 2000), with the exception of updated age weights (Helmus, Thornton, et al., 2012). Offenders can be placed in one of four risk categories based on their total score (ranging from −3 to 12): low (−3 to 1), moderate–low (2 to 3), moderate–high (4 to 5), and high (6+). Static-99R contains all the RRASOR items as well as additional items concerned with relationship history (ever lived with a lover), sexual offence history (stranger victims, noncontact sexual offences), and general criminal history (number of prior sentencing occasions, index nonsexual violence, prior nonsexual violence; see Table 1). A recent meta-analysis found a moderate relationship between Static-99R and sexual recidivism (AUC = 0.69, 95% CI = [0.66, 0.72], $k = 22$, $n = 8{,}033$; Helmus, Hanson, Thornton, Babchishin, & Harris, 2012).

Static-2002R. Static-2002R (Hanson & Thornton, 2003; Helmus, Thornton, et al., 2012) is a 14-item actuarial measure that assesses recidivism risk of adult male sex offenders. The items and scoring rules are identical to Static-2002 (Hanson & Thornton, 2003), with the exception of updated age weights (Helmus, Thornton, et al., 2012). Static-2002 (Hanson & Thornton, 2003) was created with the aim of improving Static-99. Important differences between Static-99R and Static-2002R are that Static-2002R has some additional or altered items, organized items into meaningful subscales to aid interpretation, and has more standardized coding rules. Static-2002R has five subscales: age (one item), persistence of sex offending (three items), deviant sexual interests (three items), relationship to victims (two items), and general criminality (five items). Offenders can be placed in one of five risk categories based on their total score (ranging from −2 to 13): low (−2 to 2), low–moderate (3 to 4), moderate (5 to 6), moderate–high (7 to 8), and high (9+). Previous research found that Static-2002 was significantly more predictive of sexual, violent, and any recidivism than Static-99 (Hanson, Helmus, & Thornton, 2010).

Samples

Tables 2 and 3 present the characteristics of each sample ($k = 20$, $N = 7{,}491$). All 20 samples had both RRASOR and Static-99R scores and 7 samples 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 sexual recidivism and violent recidivism, and 14 samples reported all three types of recidivism. Of the 11 studies that could be classified in terms of their treatment status, six samples were mostly treated (defined as more than 75% of the offenders), whereas four were mixed in their treatment exposure, and only one sample was mostly untreated (defined as less than 25% of the offenders). All samples used official criminal records to measure recidivism, but 10 samples used charges as the recidivism criteria whereas 10 used convictions as the recidivism criteria.

Each data set was verified for internal inconsistencies (e.g., miscalculation of total scores or item scores contradicted by other information in the data set). 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 younger than 18 years old at time of release or younger than 16 years old when they committed the index offence, or if the offender was female. These exclusionary criteria are specified in the coding rules for Static-99 (Harris, Phenix, Hanson, & Thornton, 2003) and Static-2002 (Phenix et al., 2009). The new age items of Static-99R and Static-2002R were calculated from the verified data sets for each sample.

The number of participants in these samples and the total number of samples were smaller than previously reported

Table 2. Descriptive Information of the Samples

Study	*n*	Age in years (*SD*)	Country	Recidivism criteria	Type of sample	Mostly treated	Release period	*Mdn* Year Release
Allan, Grace, Rutherford, and Hudson (2007)	492	42.3 (12.2)	New Zealand	Charges	Prison treatment	Yes	1990-2000	1994
Bengtson (2008)	308	32.5 (10.4)	Denmark	Charges	Forensic psychiatric evaluations		1978-1995	1986
Bigras (2007)	457	42.8 (12.0)	Canada	Charges	Routine CSC	Mixed	1995-2004	1999
Boer (2003)	296	41.2 (12.5)	Canada	Convictions	Routine CSC		1976-1994	1990
Bonta and Yessine (2005)	133	39.8 (9.6)	Canada	Convictions	Preselected high risk	Mixed	1992-2004	1999
Brouillette-Alarie and Proulx (2008)	228	36.0 (10.2)	Canada	Convictions	Prison and community treatment		1979-2006	1996
Cortoni and Nunes (2007)	73	41.6 (12.3)	Canada	Charges	Prison treatment	Yes	2001-2004	2003
Eher, Rettenberger, Schilling, and Pfafflin (2008)	706	40.7 (12.6)	Austria	Convictions	Routine European prison		2000-2005	2003
Epperson (2003)	177	37.2 (13.2)	United States	Charges	Routine correctional		1989-1998	1995
Haag (2005)	190	36.7 (9.7)	Canada	Convictions	Detained until end of sentence	Mixed	1995	1995
Hanson, Harris, Scott, and Helmus (2007)	702	41.6 (13.2)	Canada	Charges	Routine community supervision		2001-2005	2002
Harkins and Beech (2007)	190	43.3 (12.5)	United Kingdom	Convictions	Prison and community treatment	Yes	1994-1998	1995
Hill, Habermann, Klusmann, Berner, and Briken (2008)	86	39.4 (11.1)	Germany	Convictions	Sexual homicide perpetrators		1971-2002	1989
Johansen (2007)	273	37.8 (10.8)	United States	Charges	Prison treatment	Yes	1994-2000	1996
Knight and Thornton (2007)	466	36.1 (11.4)	United States	Charges	Civil commitment evaluation		1957-1986	1970
Långström (2004)	1,278	41.5 (12.0)	Sweden	Convictions	Routine European prison	No	1993-1997	1995
Nicholaichuk (2001)	281	34.8 (9.4)	Canada	Convictions	High intensity treatment	Yes	1983-1998	1992
Swinburne Romine, Dwyer, Mathiowetz, and Thomas (2008)	680	38.2 (12.3)	United States	Convictions	Community treatment	Mixed	1977-2007	1988
Ternowski (2004)	247	43.9 (13.0)	Canada	Charges	Prison treatment	Yes	1994-1998	1996
Wilson, Cortoni, and Vermani (2007); Wilson, Picheca, and Prinzo (2007)	228	41.7 (11.4)	Canada	Charges	Preselected high risk		1994-2007	2002
Total	7,491	39.8 (12.2)					1957-2007	1995

Note. CSC = Correctional Service Canada (administers all sentences of at least 2 years).

(e.g., Helmus, 2009) because (a) the age of the offender at release was required to compute the revised age item, and (b) the total scores of at least two of the scales included in this study had to be available in the data set.

Plan of Analyses

The first and third author conducted all analyses separately to ensure accuracy.

Discrimination. Discrimination describes the extent to which recidivists are different from nonrecidivists (also referred to as relative predictive accuracy; Gail & Pfeiffer, 2005). The Area Under the Receiver Operating Characteristic Curves (AUC ROC) is the most common method of assessing discrimination (Pintea & Moldovan, 2009; Rice & Harris, 2005; Swets, Dawes, & Monahan, 2000). In this context, the AUC can be interpreted as the probability that a randomly selected recidivist has a higher score on the risk scale than a randomly selected nonrecidivist. AUCs are robust statistics based on rank orders and are unlikely to be distorted by outliers or the arbitrary effects of scaling (Blanton & Jaccard, 2006). In addition, AUCs are useful for comparing results across samples because they are not influenced by recidivism base rates (Rice & Harris, 1995). Readers should note, however, that AUCs are influenced by the variance in the predictor variable (Humphreys & Swets, 1991).

The first set of analyses used fixed-effect and random-effects meta-analyses to compute the weighted AUCs and

Table 3. Scores and Recidivism Information

Study	*N*	*M* (*SD*) RRASOR	Static-99R	Static-2002R	Follow-up (*SD*)[a]	Recidivism rates Sexual	Violent[b]	Any
Allan et al. (2007)	492	1.4 (1.4)	1.8 (2.3)	—	5.7 (2.9)	9.6	16.5	25.2
Bengtson (2008)	308	1.8 (1.2)	3.8 (2.4)	4.6 (2.4)	16.2 (4.2)	34.1	52.3	64.6
Bigras (2007)	457	1.3 (1.3)	2.1 (2.4)	3.5 (2.5)	4.6 (1.9)	5.7	14.7	23.4
Boer (2003)	296	1.4 (1.2)	2.8 (2.8)	3.9 (2.7)	13.3 (2.1)	8.8	23.3	48.3
Bonta and Yessine (2005)	133	2.7 (1.3)	5.0 (2.1)	—	5.5 (2.4)	15.8	33.8	48.9
Brouillette-Alarie and Proulx (2008)	228	2.1 (1.4)	3.9 (2.3)	—	9.9 (4.5)	20.2	30.7	—
Cortoni and Nunes (2007)	73	1.2 (1.0)	2.2 (2.1)	—	4.6 (0.6)	0.0	8.2	12.3
Eher et al. (2008)	706	1.2 (1.0)	2.3 (2.3)	—	3.9 (1.1)	4.0	14.7	26.2
Epperson (2003)	177	1.5 (1.2)	2.5 (2.6)	—	7.9 (2.5)	14.1	—	—
Haag (2005)	190	2.0 (1.4)	4.1 (2.2)	5.7 (2.3)	7.0 (0.0)	24.7	—	—
Hanson et al. (2007)	702	1.5 (1.2)	2.4 (2.4)	3.5 (2.5)	3.4 (1.0)	8.1	16.4	27.9
Harkins and Beech (2007)	190	1.5 (1.3)	2.2 (2.6)	3.7 (2.8)	10.4 (1.1)	14.2	21.1	36.3
Hill et al. (2008)	86	1.9 (1.0)	4.7 (2.0)	—	12.6 (6.6)	15.1	29.1	61.6
Johansen (2007)	273	1.8 (1.2)	2.9 (2.3)	—	9.1 (1.1)	7.7	20.5	53.5
Knight and Thornton (2007)	466	2.4 (1.3)	4.6 (2.4)	6.1 (2.5)	8.6 (2.6)	26.2	36.9	53.0
Långström (2004)	1,278	0.8 (0.9)	2.0 (2.4)	—	8.9 (1.4)	7.5	21.4	—
Nicholaichuk (2001)	281	2.4 (1.4)	4.8 (2.4)	—	6.4 (4.0)	18.5	—	—
Swinburne Romine et al. (2008)	680	1.2 (1.1)	1.7 (2.2)	—	16.8 (7.8)	13.8	—	—
Ternowski (2004)	247	1.2 (1.2)	1.6 (2.5)	—	7.5 (1.0)	8.1	15.4	19.8
Wilson, Cortoni, et al. (2007); Wilson, Picheca, et al. (2007)	228	2.8 (1.5)	5.1 (2.3)	—	5.2 (3.0)	10.5	25.9	35.5
Total	7,491	1.5 (1.3)	2.7 (2.6)	4.3 (2.7)	8.3 (5.2)	12.0	22.4	35.9

a. Follow-up period in years.
b. Violent recidivism, including sexual.

their 95% CIs for each risk scale. 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. In random-effects meta-analysis, the between-study variability is included in the error term, resulting in wider confidence intervals (Schmidt, Oh, & Hayes, 2009). The results of the random-effects and fixed-effect models converge as the amount of between-study variability decreases. Although there are several methods available to compute random-effects and fixed-effect estimates of the AUCs and standard errors, we used the formulas and procedures recommended by Hedges and Vevea (1998) and Hedges (1994), respectively.

To test the extent to which the risk scales differed in their level of discrimination, differences between AUCs were meta-analyzed using the Delong method for computing the standard error of the difference (Delong, Delong, & Clarke-Pearson, 1988). Of note, there are several different methods to compare differences in AUCs (Bandos, Rockette, & Gur, 2005; Braun & Alonzo, 2008; Delong et al., 1988; Hanley & Hajian-Tilaki, 1997), and all continue to be underused in the literature (Robin et al., 2011). The two most popular methods are those described by Delong et al. (1988) and Hanley and McNeil (1983; Vickers, Cronin, & Begg, 2011). Use of the Hanley and McNeil method is expected to decline given that Hanley recommends that it be replaced by the Delong et al. method (J. A. Hanley, personal communication, May 20, 2011), as do other commentators (Begg, 1991; Hanley & Hajian-Tilaki, 1997; Le & Lindgren, 1995). The Delong et al. method uses a more precise method of computing variances by assigning positions to scores based on whether each recidivist has a higher, lower, or equal score on the measure than a nonrecidivist. A detailed explanation of the Delong et al. method is provided in the appendix.

To compute the standard errors derived from the Delong method, the pROC package for R program (Robin et al., 2011) was updated to include formulae from Hanley and Hajian-Tilaki (1997). If the 95% confidence interval of the difference between measures included zero, the difference in discrimination between the two scales was not statistically significant.

Incremental validity. Incremental validity was examined using Cox regression (Allison, 1984). The dependent variable was time-at-risk or "survival time," and each case was identified as either a failure (e.g., recidivist) or censored

(i.e., still at risk at the end of the follow-up time). Cox regression calculates hazard rates or the extent to which the probability of failure varies as a function of predictor variables. 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. The analyses provide the Wald statistic that, if significant, indicates that the scale adds incremental validity to the other scale(s) included in the model. The analyses also provide Exp(*B*), which is a hazard ratio and indicates the scale's relationship with recidivism. For example, a hazard ratio of 1.10 indicates that each one-score increase on the scale increases the hazard by a factor of 1.10, or 10%.

Comparing combinations of risk scales. To compare different rules for combining risk scales (lowest, highest, average), the scales needed to be standardized to a common metric. Based on previous research (Hanson, Babchishin, Helmus, & Thornton, 2012), the metric used was the log of the hazard ratios associated with each score, centered on the median value of the scale in routine samples. For Static-99R and Static-2002R, the necessary parameters (i.e., median, log hazard ratios) have previously been reported based on samples that overlap with those reported in the current study (Babchishin, 2011; Hanson, Babchishin, et al., 2012; Hanson, Lloyd, Helmus, & Thornton, 2012). The parameters for the RRASOR had not been previously reported and were calculated for this analysis using the same data sets and procedures used for Static-99R and Static-2002R (using routine samples, $k = 6$, $n = 3{,}642$). Specifically, the RRASOR was considered to have a median value of 1 (based on Hanson, Lloyd, et al.'s [2012] sample and method) and a log hazard ratio of 0.578 (based on Hanson, Babchishin, et al.'s [2012] sample and method). The corresponding values for Static-99R were a median of 2 and a log hazard ratio of 0.332; for Static-2002R the median was 3 and log hazard ratio was 0.322.

To calculate the predicted recidivism rate, the hazard ratios were multiplied by the sexual recidivism base rate (see Hanson, Babchishin, et al., 2012, for the strengths and limitations of this approach to estimating absolute recidivism rates). The base rate of sexual recidivism in this sample (8.8%) was estimated from logistic regression with the three risk scales (centered on their respective medians) entered as predictors and the ragged recidivism information (i.e., with varying lengths of follow-up) entered as the outcome. These analyses were, therefore, restricted to cases with scores on all three measures ($k = 7$, $n = 2{,}609$). Estimated recidivism rates over 100% were trimmed to 100%.[1]

Calibration

The *E*/*O* index was used to examine the fit between the expected and observed recidivism rates derived from the three combination methods and can be considered a measure of effect size. *E* refers to expected number of recidivists and *O* refers to observed number of recidivists. Perfect fit is indicated by an *E*/*O* index of 1.0. Following Rockhill, Byrne, Rosner, Louie, and Colditz (2003), the 95% confidence intervals for the *E*/*O* index were calculated using the Poisson variance for the logarithm of the observed number of cases (*O*):

$$95\%\ \mathrm{CI}(E/O) = (E/O)\exp(\pm 1.96\sqrt{1/O})$$

Calibration was also examined using the Hosmer–Lemeshow test (Hosmer & Lemeshow, 2000) and provides an overall significance test for sets of predicted values. The Hosmer–Lemeshow test is a chi-square goodness-of-fit statistic with sparse data clumped into a limited number of roughly equal-sized groupings, or "bins" ($g \leq 10$). The fit between the observed and expected recidivism rates is calculated for each bin, thus highlighting potential areas of misfit. In our study, the expected recidivism rates for groupings were estimated as the weighted average of the expected values for each score within groupings. A nonsignificant Hosmer–Lemeshow test suggests good fit between the observed and expected recidivism rates (equivalent to a Pearson chi-square test with $df = g - 2$).

Results

Discrimination

Tables 4 to 6 present the weighted AUC for each risk scale and differences between scales tested using the Delong method. The total scores of Static-99R and Static-2002R predicted sexual and violent recidivism similarly. However, the total scores of Static-2002R predicted any recidivism with significantly greater discrimination than Static-99R (Table 4). The fixed-effect and random-effects meta-analyses of differences between AUCs produced identical results because the between-study variability was less than would be expected by chance ($Q < df$).

Table 5 presents the meta-analyzed AUCs for the RRASOR and Static-99R. The Delong test found that Static-99R total scores had significantly greater discrimination in predicting sexual, violent, and any recidivism than the RRASOR total scores, with the largest differences found for violent (including sexual) and any recidivism. The same pattern of results was found for the total scores of the RRASOR and Static-2002R, with Static-2002R predicting sexual, violent, and any recidivism more accurately than the RRASOR (Table 6).

The differences in discrimination between the scales were remarkably consistent across samples for the prediction of sexual and violent recidivism, as indicated by nonsignificant *Q* statistics (with the exception of comparing Static-99R and RRASOR for violent recidivism). For any recidivism, the comparison between the RRASOR and

Table 4. Meta-Analysis of the Discrimination of Static-99R and Static-2002R

		Fixed			Random					
			95% CI			95% CI				
Outcome	Measure	Weighted AUC	LL	UL	Weighted AUC	LL	UL	*k*	*N*	*Q*
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		**−0.00268**	−0.0154	0.0100	**−0.00268**	−0.0154	0.0100	7	2,609	4.26
Violence		**−0.00377**	−0.0145	0.0070	**−0.00377**	−0.0145	0.0070	6	2,419	3.32
Any*		**−0.0127**	−0.0222	−0.00327	**−0.0127**	−0.0222	−0.00327	6	2,419	2.61

Note. AUC = area under the curve; CI = confidence interval; LL = lower limit; UL = upper limit. A negative difference score indicates greater discrimination by the Static-2002R compared with the Static-99R.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5. Meta-Analysis of the Discrimination of RRASOR (Rapid Risk Assessment for Sex Offence Recidivism) and Static-99R

		Fixed			Random					
			95% CI			95% CI				
Outcome	Measure	Weighted AUC	LL	UL	Weighted AUC	LL	UL	*k*	*N*	*Q*
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 Static-99R and RRASOR										
Sexual*		**0.0295**	0.0147	0.0444	**0.0344**	0.110	0.0578	19	7,418	23.59
Violence*		**0.102**	0.0877	0.116	**0.100**	0.0757	0.124	16	6,163	28.53*
Any*		**0.122**	0.108	0.136	**0.124**	0.0951	0.152	14	4,657	38.37***

Note. AUC = area under the curve; CI = confidence interval; LL = lower limit; UL = upper limit. A positive difference score indicates greater discrimination by the Static-99R compared with the RRASOR.
*$p < .05$. **$p < .01$. ***$p < .001$.

Static-99R and the comparison between the RRASOR and Static-2002R had significant variability, indicating that the difference in discrimination for these comparisons were inconsistent across the samples.

Incremental Validity

Tables 7 to 9 present the Cox regression analyses used to examine the incremental validity of the total scores of risk

Table 6. Meta-Analysis of the Discrimination of RRASOR (Rapid Risk Assessment for Sex Offence Recidivism) and Static-2002R

		Fixed			Random					
			95% CI			95% CI				
Outcome	Measure	Weighted AUC	LL	UL	Weighted AUC	LL	UL	*k*	*N*	*Q*
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 Static-2002R and RRASOR										
Sexual*		**0.0365**	0.0142	0.0588	**0.0365**	0.0142	0.0588	7	2,609	3.80
Violence*		**0.0968**	0.0767	0.117	**0.102**	0.0647	0.138	6	2,419	13.38
Any*		**0.138**	0.119	0.156	**0.139**	0.0965	0.182	6	2,419	25.94***

Note. AUC = area under the curve; CI = confidence interval; LL = lower limit; UL = upper limit. A positive difference score indicates greater discrimination by the Static-2002R compared with the RRASOR.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 7. Incremental Validity of the Risk Instruments for Predicting Sexual Recidivism

	Sexual recidivism					
				95% CI		
	N	*r*	Exp(*B*)	LL	UL	Wald
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 simultaneously in the model. Sample sizes fluctuate because of the amount of cases censored before earliest event. RRASOR = Rapid Risk Assessment for Sex Offence Recidivism; *r* = Pearson's correlation between measures; CI = confidence interval; LL = lower limit; UL = upper limit.
*$p < .05$. **$p < .01$. ***$p < .001$.

scales for each recidivism type. For the prediction of sexual recidivism, all risk scales were generally found to add incrementally to one another despite large correlations between scales, ranging from .70 and .92 (Table 7). Namely, the RRASOR (Wald = 9.75, $p < .01$) and Static-99R (Wald = 143.15, $p < .001$) each added incrementally

Table 8. Incremental Validity of the Risk Instruments for Predicting Violent Recidivism

	Violent (including sexual) recidivism					
				95% CI		
	N	*r*	Exp(*B*)	LL	UL	Wald
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 simultaneously in the model. Sample sizes fluctuate because of the amount of cases censored before earliest event. RRASOR = Rapid Risk Assessment for Sex Offence Recidivism; *r* = Pearson's correlation between measures; CI = confidence interval; LL = lower limit; UL = upper limit.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 9. Incremental Validity of the Risk Instruments for Predicting Any Recidivism

	Any recidivism					
				95% CI		
	N	*r*	Exp(*B*)	LL	UL	Wald
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 simultaneously in the model. Sample sizes fluctuate because of the amount of cases censored before earliest event. RRASOR = Rapid Risk Assessment for Sex Offence Recidivism; *r* = Pearson's correlation between measures; CI = confidence interval; LL = lower limit; UL = upper limit.
**$p < .01$. ***$p < .001$.

to one another; Static-99R (Wald = 8.22, $p < .01$) and Static-2002R (Wald = 6.62, $p < .05$) each added incrementally to one another; and, finally, Static-2002R (Wald = 55.17, $p < .001$) added incremental validity to the RRASOR but the RRASOR did not add incrementally to Static-2002R (Wald = 1.27, $p = .26$). The incremental

Table 10. Incremental Validity of Constructs

			95% CI		
	N	Exp(*B*)	LL	UL	Wald
Sexual recidivism					
Sexual criminality	2,594	1.189	1.135	1.246	53.24***
General criminality		1.185	1.121	1.253	35.85***
Age		0.979	0.970	0.988	19.60***
Violent recidivism					
Sexual criminality	2,405	1.061	1.018	1.105	8.00**
General criminality		1.335	1.278	1.396	164.45***
Age		0.964	0.957	0.972	79.17***
Any recidivism					
Sexual criminality	2,406	1.009[a]	0.975	1.044	0.26
General criminality		1.399	1.350	1.450	342.48***
Age		0.961	0.955	0.967	141.07***

Note. Analyses conducted separately for each comparison, with each sample entered as strata and components entered simultaneously. Sexual criminality = Prior sentencing occasion, any juvenile arrest for sexual offences, rate of sexual offending, noncontact offence, male victim, young and unrelated victim, unrelated victim, and stranger victim; α = .684, *n* = 2,594. Criminality = Any prior involvement, prior sentencing occasions, community violation, years free prior to index, any prior nonsexual violent occasion, and nonsexual violent convictions at time of index; α = .786, *n* = 2,594. Cases required all items (13 Static-2002R items and 1 Static-99R item) on subscales to be included in the analyses. RRASOR = Rapid Risk Assessment for Sex Offence Recidivism; CI = confidence interval; LL = lower limit; UL = upper limit.

a.For the prediction of any recidivism, sexual criminality did not improve the model after accounting for general criminality and age, χ^2 change = 0.26, *df* = 1, *p* = .61.

**p < .01. ***p < .001.

contributions of the scales were consistently positive when predicting sexual recidivism. For example, the Exp(*B*) of the RRASOR for the model that included both the RRASOR and Static-99R was 1.26 (95% CI = [1.21, 1.31]), indicating that each one-score increase on the RRASOR increases the hazard by a factor of 26%, after controlling for Static-99R. In addition, entering all three risk scales into a model (Comparison 4) found that Static-99R and Static-2002R added incrementally to the prediction of sexual recidivism, but the RRASOR did not add incrementally (Wald = 0.48, *p* = .49), after accounting for both Static-99R and Static-2002R.

For the prediction of violent (including sexual) recidivism, all three scales added incremental information for all analyses. Of note, the incremental effect for the RRASOR was reversed—namely, *lower* scores on the RRASOR were associated with *higher* rates of violent recidivism after controlling for the other scales (Table 8). In addition, the model that included the three risk scales found significant incremental validity for each scale (with lower scores on the RRASOR predicting higher rates of violent recidivism).

For the prediction of any recidivism, each risk scale once again added incremental validity to one another (Table 9). Specifically, the RRASOR and Static-99R added incrementally to each other, Static-99R and Static-2002R added incrementally to each other, and, finally, the RRASOR and Static-2002R added incrementally to each other. Similar to the prediction of violent recidivism, lower scores on the RRASOR were associated with higher rates of any recidivism, after controlling for the other scales. Last, the model that included all three risk scales found significant incremental validity for each scale (with lower RRASOR scores predicting higher rates of any recidivism).

The incremental validity of the constructs measured by the risk scales were also examined (see Table 10). Specifically, based on factor analyses (Barbaree et al., 2006b; Thornton, 2010), the items of the risk scales were separated into the domain of sexual criminality, general criminality, and age. For the prediction of sexual recidivism, each domain was found to add incrementally to one another. As expected, the incremental effect for age was negative—younger age was associated with higher rates of sexual, violent, and any recidivism. In addition, all three domains were found to add incremental information in the prediction of violent (including sexual) recidivism. For any recidivism, sexual criminality did not add incremental validity to general criminality and age (Wald = 0.26, *p* = .61).

To examine the practical importance of the incremental finding, offenders were also sorted into risk categories (for Static-99R: low, −3 to −1; moderate, 0 to 4; high, 5+ and for Static-2002R: low, −2 to 0; moderate, 1 to 5; high, 6+). These categories were based on a scale-independent definition of nominal risk categories suggested by Babchishin

Table 11. 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	3.2 (6/189)	1.7 (1/60)	—	2.8 (7/249)
Moderate	5.9 (1/17)	9.8 (134/1,374)	17.4 (31/178)	10.6 (166/1,569)
High	—	26.4 (29/110)	30.5% (208/681)	30.0 (237/791)
Total	3.4 (7/206)	10.6 (164/1,544)	27.8 (239/859)	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).

and Hanson (2009). Specifically, offenders with a score associated with less than half the rate of sexual reoffending than the typical offender (hazard ratio < 0.50) were classified as "low risk." Offenders with a score associated with more than half the rate of reoffending than the typical offender, but less than twice the rate of reoffending of the typical offender (hazard ratio 0.50-1.99) were classified as "moderate risk." Last, offenders with a score associated with twice the rate of a typical offender (hazard ratio > 2.00) were classified as "high risk."

A simple crosstab of the sexual recidivism rates by Static-99R and Static-2002R risk categories is presented in Table 11 to allow for a visual representation of the recidivism rates of offenders for whom the scales provide discordant results (when both scales sort offenders into different risk categories) and concordant results. Recidivism rates for discordant groups were intermediate between the two adjacent risk categories. For example, when both scales classified offenders as moderate risk, the observed recidivism rate was 9.8% (134/1,374), and when both scales rated offenders as high risk, the observed rate was 30.5% (237/791). When one scale classified the offender as moderate and the other scale classified the offender as high, the observed sexual recidivism rate was 20.8% (60/288). This 10% difference in recidivism is similar in size to the effects found for most of the well-established risk factors (e.g., any male victim, any unrelated victims; Hanson & Bussière, 1998).

Combining Across Risk Scales

Three methods for combining risk scales were examined: (a) choosing the lowest, (b) choosing the highest, and (c) taking an average. Analyses were restricted to cases with scores on all three measures (k = 7, n = 2,609). Each method produced scores that were highly correlated (Kendall's tau correlation coefficients ranged from .77 to .88) and provided similar discrimination (i.e., sorting offenders from lowest risk to highest risk to reoffend). The only significant difference in discrimination was that choosing the highest score did worse than averaging (see Table 12). The estimated recidivism rates were derived from an exponential (ever increasing) model. The exponential is an acceptable fit to Static-99R scores when the expected recidivism rates are less than 50% (Hanson, Babchishin, et al., 2012). With high scores, however, the exponential model can produce impossible estimates (i.e., >100%), as was the case in the current study. Consequently, we removed the highest risk bin defined by the Hosmer–Lemeshow test (i.e., the bin with expected recidivism rates over 100%, remaining n = 2,362). With this trimmed data set, the average produced an E/O index of 1.03 (95% CI = [0.92, 1.15]), indicating good agreement between the observed recidivism rates and the recidivism rates predicted by this approach. As could be expected, choosing the lowest systematically and significantly underestimated the recidivism rates (E/O = 0.75, 95% CI = [0.67, 0.83]), and choosing the highest method systematically and significantly overestimated the recidivism rates (E/O = 1.49, 95% CI = [1.33, 1.66]). The Hosmer–Lemeshow test found that the averaging approach produced good calibration (χ^2 = 12.20, df =7, p = .094; n = 2,362), whereas the other two methods deviated from perfect calibration (χ^2_{min} = 55.10, df = 7, p < .001; n = 2,362; χ^2_{max} = 78.00, df = 7, p < .001; n = 2,362).

A similar pattern occurred if we included all risk scores (n = 2,609); however, each approach produced expected recidivism rates that were statistically different from the observed rates; $E/O_{average}$ = 1.13, 95% CI = [1.02, 1.24], E/O_{min} = 0.81, 95% CI = [0.74, 0.90], and E/O_{max} = 1.60, 95% CI = [1.45, 1.76], respectively. Notably, however, the averaging method produced the least deviation of the three. Figure 1 presents a calibration plot of the three methods including all bins (n = 2,609), with the dashed line indicating perfect calibration. Misfit between the expected and observed values was evident in the highest scores of the respective combination methods.

Table 12. Comparing the Accuracy of Methods for Combining Risk Scales in Predicting Sexual Recidivism

	Fixed-effect			Random-effects					
		95% CI			95% CI				
	Weighted AUC	LL	UL	Weighted AUC	LL	UL	*k*	*n*	*Q*
Believe the lowest score (Min)	**0.686**	0.658	0.715	**0.694**	0.647	0.742	7	2,609	14.81*
Believe the highest score (Max)	**0.678**	0.648	0.706	**0.686**	0.640	0.730	7	2,609	13.22*
Average	**0.690**	0.661	0.719	**0.698**	0.651	0.746	7	2,609	14.79*
Differences between combination methods									
Min vs. Max[a]	**0.00841**	−0.00575	0.0226	**0.00841**	−0.00575	0.0226	7	2,609	3.94
Min vs. Average[b]	**−0.00381**	−0.0125	0.00485	**−0.00381**	−0.0125	0.00485	7	2,609	4.34
Max vs. Average[c]*	**−0.0133**	−0.0204	−0.00621	**−0.0124**	−0.0226	−0.00228	7	2,609	6.42

Note. AUC = area under the curve; CI = confidence interval; LL = lower limit; UL = upper limit.
a. A positive difference score indicates greater discrimination by believing the lowest score compared with believing the highest score combination method.
b. A positive difference score indicates greater discrimination by believing the lowest score compared with the averaging method.
c. A positive difference score indicates greater discrimination by believing the highest score compared with the averaging method.
*$p < .05$.

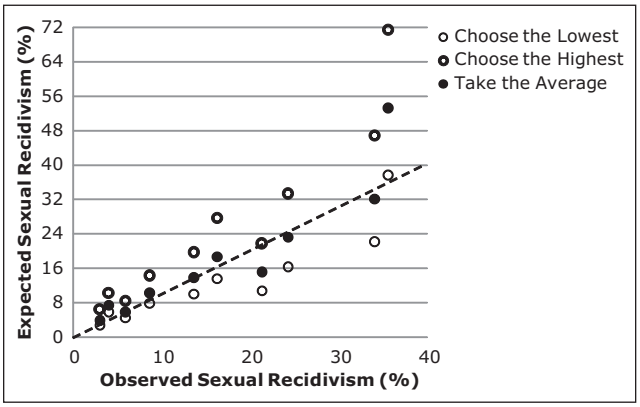

Figure 1. Calibration plot for three combination methods for predicting sexual recidivism rates
Dashed line indicates perfect calibration. Each set of markers represents a bin grouping used in the Hosmer–Lemeshow test.

Discussion

The purpose of the study was to compare the predictive accuracy and incremental validity of three actuarial risk assessment scales for sex offenders and by doing so, demonstrate methods appropriate more generally for the validation of criterion-referenced measures. We found that the total scores of Static-99R and Static-2002R provided better discrimination than the total scores of the RRASOR for sexual, violent, and any recidivism. No differences in discrimination were found between Static-99R and Static-2002R for sexual and violent recidivism, but Static-2002R was found to outperform Static-99R in predicting any recidivism. Remarkably, despite large correlations and substantial item overlap between the risk scales, the scales consistently added incremental validity to one another, with one exception: the RRASOR did not add incremental validity to the prediction of sexual recidivism after controlling for Static-2002R.

The study also examined three methods of combining risk tools: (a) choose the lowest, (b) choose the highest, and (c) take an average. Each of these methods was highly correlated with the others; however, the averaging approach produced better discrimination than choosing the highest risk scale. For estimating absolute recidivism rates, the averaging showed the best calibration. The other methods systematically underestimated (choose the lowest) or overestimated (choose the highest) the observed recidivism rates.

Incremental Contribution of Risk Scales

Our findings are in stark contrast with Seto (2005) who did not find incremental validity of similar risk scales (albeit using a much smaller sample size). Both studies found that various decision rules for combining different risk tools had little effect on discrimination. Unlike Seto (2005), however, the current findings suggest that there are clear benefits to considering multiple scales. Large correlations between risk scales are routinely assumed to preclude incremental validity. This is not the case. The presence of incremental validity even between these highly correlated scales suggests that incremental validity should be expected for most risk scales (provided the samples sizes are large enough).

It is not obvious, however, how different findings across risk scales should be interpreted. Vrieze and Grove (2010)

assumed that discordant results between two measures with substantial overlap in content would form ". . . a prima facie reason to disbelieve" both scales and would ". . . undercut each others' statuses as knowledge claims" (Vrieze & Grove, 2010, p. 388). We believe 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 or different weighting of domains predictive of recidivism.

Based on previous research (Barbaree et al., 2006b; Thornton, 2010), there are at least three broad dimensions represented by the items in sex offender risk assessment tools: age, sexual criminality, and general criminality. RRASOR has only items related to age and sexual criminality, whereas Static-99R and Static-2002R have items from all three dimensions. We found that each of these dimensions had different relative contributions to the different recidivism outcomes. In particular, a history of sexual criminality was related to sexual recidivism and (to a lesser extent) violent recidivism. Sexual criminality was not predictive of general (i.e., any) recidivism. Consequently, it is possible to increase the measures' overall association with any and violent recidivism by rebalancing the weights of these dimensions, according more weight to general criminality and age and less weight to sexual criminality. Although initially unexpected, the negative weights for the RRASOR (after controlling for the Static scales) for violent and any recidivism make sense: negative weights for RRASOR scores decreased the predictor's saturation with sexual criminality for outcomes that had little or no relationship to sexual criminality (a classic suppression effect).

Suppression effects, however, cannot provide a complete explanation for our findings. The RRASOR added incremental validity (positively) to Static-99R in the prediction of sexual recidivism, despite all the RRASOR items being already included in Static-99R. We even used the items of Static-99R to calculate the RRASOR. These incremental validity findings, therefore, cannot be attributed to new constructs being captured by the RRASOR, but to different weighting of the items in each scale. 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.

Combining Risk Scales

Given the presence of incremental validity between risk scales, evaluators interested in systematic risk assessments may wish to use multiple risk scales in their evaluations. Rules for combining risk scales first require an understanding of the domains being measured by the scales and how the domains relate to the outcome of interest. Understanding what the scale items are measuring would allow evaluators to explain inconsistencies in risk rating across measures and inform methods of combining 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, such research is essential given that the incremental addition of risk scales is most likely not limited to the three actuarial scales examined in this study.

Although the current results (and logic) favor the averaging approach, other rules may be better for other sets of measures. Specifically, psychological measures typically assess one construct (e.g., depression) and are created by sampling from a pool of possible items. When these measures provide discordant results, psychometric theory supports an averaging or additive approach. Such approaches are based on the assumption of classical test theory that increasing the item pool should reduce sampling error and produce more reliable results (Nunnally & Bernstein, 1994). Of course, if the additional items are substantially worse (less predictive; or less related to the latent construct) than the items already considered, the accuracy of the overall prediction would deteriorate. Unlike traditional psychological measures, however, combining multiple risk scales requires a construct-level approach because these measures tend to assess multiple constructs (e.g., general criminality, sexual deviancy). Specifically, the preferred method of combination will depend on whether or not the scales are measuring similar or different domains, as well as the domains' relationships with the outcome of interest.

If the two risk scales are sampling from the same domain(s), then similar to traditional psychological measures, an averaging approach can decrease measurement error and increase discrimination. In contrast, if the selected scales are not sampling the same domains, then the evaluators would require a defensible model concerning (a) the domains measured by the scales, (b) how the domains relate to the outcome of interest, and (c) 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. The direction forward for risk assessment combines empirical prediction with the construct validity tradition. Until then, risk evaluators will continue to be faced with the knowledge that other variables (and scales) add incremental validity without being able to explain why.

An additional concern is the metric used to combine scales. In the current study, we used relative risk (from the median value) as the scale-independent metric. As we have argued elsewhere, relative risk is an intrinsic and surprisingly stable characteristic of this type of risk scale (Hanson, Babchishin, et al., 2012; Helmus, Hanson, et al., 2012). Absolute recidivism rates estimates could also be used

(either in raw or logit metrics); however, absolute recidivism rates vary across samples for reasons that are not fully understood (Helmus, Thornton, et al., 2012). Seto (2005) and Barbaree et al. (2006a) used percentile ranks to compare scales; although plausible, percentiles measure the rarity of particular scores and provide no direct information concerning the extent to which the score is associated with the outcome of interest. Further work is required to develop both the evidence base and practical methods of combining risk scales.

Future Directions

Although some progress in risk assessment can be made by refining item weights, we do not believe this will solve the most pressing problems of applied risk assessment (Hanson, 2009). Specifically, the current study provides clear evidence that the item weights in actuarial scales are unlikely to ever be optimal. Given large-enough sample sizes, the null hypothesis (finding no incremental validity) can almost always be rejected (Cohen, 1994). In addition, the refinement of weights is a never-ending task requiring larger sample sizes for decreasingly small gains in precision. Test developers would also need to be vigilant about overfitting the data, as small adjustments rarely generalize to other data sets (Cureton, 1950). As well, complex weights will reduce the practical ease of scoring and increase the risk of error; integers are simple.

We believe the way forward involves increasing attention to the construct validity of risk scales. Determining the construct validity of a test requires identification of the latent constructs of interest and the integration of multifaceted evidence regarding the extent to which the test reflects these latent constructs. Construct validity can be assessed, for example, by examining content validity, factor structure, known-groups validity, predictive validity, temporal stability, and the results of experiments (Cronbach & Meehl, 1955; Embretson, 2010). The degree of relationship between the risk scale and other relevant measures is but one step required to establish construct validity (see Embretson, 2007, for a review). Construct validity research would be an essential step in identifying (a) the latent constructs measured by the risk scales, (b) how these constructs relate to the outcome of interest, and (c) how these constructs should be weighted and combined.

The development, validation, and interpretation of empirical actuarial measures produce unique challenges. Large samples (thousands) are required to compare the predictive accuracy of actuarial measures. An additional challenge is that the statistics used to evaluate the predictive accuracy of risk scales have not been commonly included in psychometric textbooks or statistical courses in psychology (e.g., the Delong test of differences between AUC ROC; Delong et al., 1988); consequently, many researchers are unaware or inexperienced with the appropriate analyses. The analyses presented in the current study are not novel and are commonly used to assess diagnosis accuracy and prognostic indicators in other fields such as medical research (AUC ROC—Begg, 1991; Grzybowski & Younger, 1997; Swets, 1988; Zweig & Campbell, 1993; survival analysis—Horton & Switzer, 2005; Sasieni, 2005). To support cumulative knowledge development, studies comparing the discrimination of risk scales should routinely report a standard set of indicators. Specifically, we recommend that authors present correlations between the measures, the AUC of each scale, and the standard error of the difference. Authors interested in examining the calibration of risk scales should report other indicators as well, such as the *E*/*O* index and calibration plots (Altman, Vergouwe, Royston, & Moons, 2008; Harrell, Lee, & Mark, 1996).

Recommendations on Which Risk Scales to Use

Both Static-99R and Static-2002R showed equivalent predictive accuracy for sexual and violent recidivism. This contrasts with a previous meta-analysis of seven of the current data sets in which Static-2002 outperformed Static-99 in predicting sexual, violent, and any recidivism (Hanson et al., 2010). A subsequent study has similarly found Static-2002 to predict sexual recidivism better than Static-99 (Stalans, Hacker, & Talbot, 2010). The revised age weights in the *R* versions notably increased the discrimination of Static-99R, whereas a smaller improvement was found in Static-2002R (the original Static-2002 age weights were quite similar; Helmus, Thornton, et al., 2012). Consequently, evaluators wishing to choose between Static-99R and Static-2002R must make that decision based on criteria other than discrimination ability (e.g., calibration, ease of coding, conceptual clarity, depth of replication).

Finally, we recommend that evaluators do not rely primarily on the RRASOR in their sex offender risk assessments. Despite the RRASOR being the second most used scale in the United States (McGrath et al., 2010), it had meaningfully lower predictive accuracy than Static-99R and Static-2002R in all analyses. Nevertheless, the RRASOR was found to add incremental validity to Static-99R, which will motivate some evaluators to consider both the RRASOR and Static-99R for high-stakes cases (e.g., civil commitment). Use of both RRASOR and Static-99, however, invites the problems associated with interpreting a subset of items already included in a longer, better validated measure of sexual recidivism. In cases where the outcome of primary interest is violent or any recidivism, we suggest evaluators use scales specifically designed for these outcomes (for reviews, see Hanson & Morton-Bourgon, 2009; Yang, Wong, & Coid, 2010).

Appendix

Delong, Delong, and Clarke-Pearson (1988) Method for Comparing Area Under the Receiver Operating Characteristic Curves

The Delong et al. (1988) approach starts by calculating for each observation a quantity called the "placement." Specifically, the scores of the recidivists are arranged as table columns and the scores of the nonrecidivists are arranged as rows (see example below). Next, a score of 1 is allocated if a recidivist has a higher score than a nonrecidivist ("correct" ordering), a score of 0 if a nonrecidivist has a higher score than a recidivist ("incorrect" ordering), and a score of 0.5 if the scores are equal. The placement for an observation is the average of these 0/.5/1 (incorrect/tie/correct) orderings. For example, the placement for the 6th nonrecidivist in the example (NR_6) is 3.5/6 = 0.58. There are two sets of placements: a set for the recidivists, which in the example below ranges from 0.31 to 0.94 (n_R = 6; M = 0.677, S_R^2 = 0.04284), and a set for the nonrecidivists, which in the example below ranges from 0.08 to 1.00 (n_{NR} = 8; M = 0.677, S_{NR}^2 = 0.1338). The means of these sets are identical and are equal to the AUC (0.677). In the Delong method, the variance of the AUC is a weighted sum of the variances of the placements:

$$S_{AUC}^2 = (S_R^2 / n_R) + (S_{NR}^2 / n_{NR})$$

In our example, S_{AUC}^2 = (0.04284/6) + (0.1338/8) = 0.02386. (Note that these calculations were completed retaining more significant figures than shown in Table 1, e.g., 0.31250 rather than 0.31.)

To compare two measures (Measure A and Measure B), the AUCs and the variances for each measure are calculated using the method described above. The covariance is similarly estimated from the placements on each measure. The covariance between the placements for Measure A and Measure B are calculated separately for the recidivists and the nonrecidivists, using the standard formula (covariance = correlation × SD_1 × SD_2). The overall covariance (for both Measure A and Measure B) is a weighted sum of the covariance for the recidivists and the covariance for the nonrecidivists:

$$\text{Covariance}_{\text{AB}} = (\text{Covariance}_{\text{AB-R}/nR}) + (\text{Covariance}_{\text{AB-NR}/nNR})$$

The standard error of the difference between the AUCs is then defined in the usual way:

$$SE_{A-B} = \sqrt{SE_A^2 + SE_B^2 - 2\,\text{cov}_{AB}}$$

A worked example is provided in Hanley and Hajian-Tilaki (1997). With small data sets, it is relatively easy to compute the values in Delong et al.'s method using a hand calculator or generic spreadsheets. Purpose-built software programs are available and are convenient for the analysis of larger data sets (e.g., Robin et al., 2011).

Example of Calculating Placements for the Delong et al. Method

Observations for nonrecidivists	Observations for recidivists						Placement of nonrecidivists
	Rec_1 = 1	Rec_2 = 3	Rec_3 = 4	Rec_4 = 5	Rec_5 = 5	Rec_6 = 8	
NR_1 = −2	1	1	1	1	1	1	1.00
NR_2 = −1	1	1	1	1	1	1	1.00
NR_3 = 1	0.5	1	1	1	1	1	0.92
NR_4 = 2	0	1	1	1	1	1	0.83
NR_5 = 2	0	1	1	1	1	1	0.83
NR_6 = 4	0	0	0.5	1	1	1	0.58
NR_7 = 6	0	0	0	0	0	1	0.17
NR_8 = 8	0	0	0	0	0	0.5	0.08
Placement of recidivists	0.31	0.62	0.69	0.75	0.75	0.94	

Acknowledgments

We would like to thank Michael Seto for his helpful comments on an earlier version of this manuscript. Thank you to 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. We would also like to thank Ian Broom and Xavier Robin for their help with the pROC program for R.

Authors' Note

The views expressed are those of the authors and not necessarily those of Public Safety Canada. A related version of this work has been published as a government report (Babchishin, Hanson, & Helmus, 2011).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this project was provided in part by the Social Science and Humanities Research Council of Canada.

Note

1. Rates of more than 100% occurred in the highest scores, and specifically, for five scores when using the "choose the highest" method, once when using the "choose the lowest" method, and for four scores when averaging.

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.

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2008). Prognosis and prognostic research: Validating a prognostic model. *British Medical Journal, 338*, 1432-1435. doi:10.1136/bmj.b605

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

Babchishin, K. M. (2011, November). Risk ratios for Static-99/R and Static-2002/R. In L. Helmus (Chair), *Improving risk communication: Non-arbitrary methods for quantifying risk.* Paper presented at 30th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Toronto, Ontario, Canada.

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

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2011). *The RRASOR, Static-99R, and Static-2002R all add incrementally to the prediction of recidivism among sex offenders* (Corrections Research User Rep. No. 2011-02). Ottawa, Ontario: Public Safety Canada.

Bandos, A. I., Rockette, H. E., & Gur, D. (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine, 24*, 2873-2893. doi:10.1002/sim.2149

Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006a). 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., Langton, C. M., & Peacock, E. J. (2006b). The factor structure of static actuarial items: Its relation to prediction. *Sexual Abuse: A Journal of Research and Treatment, 18*, 207-226. doi:10.1007/s11194-006-9011-6

Begg, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine, 10*, 1887-1895. doi:10.1002/sim.4780101205

*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]. *Dissertations Abstracts International: Section B, 68*(09). (UMI No. NR30941)

Blanton, H., & Jaccard J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41. doi:10.1037/0003-066X.61.1.27

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

*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 Rep. 2005-04). Ottawa, Ontario: Public Safety and Emergency Preparedness Canada. Unpublished raw data.

Braun, T. M., & Alonzo, T. A. (2008). A modified sign test for comparing paired ROC curves. *Biostatistics, 9*, 364-372. doi:10.1093/biostatistics/kxm036

*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.

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

*Cortoni, F., & Nunes, K. L. (2007). *Assessing the effectiveness of the National Sexual Offender Program* (Research Rep. No. R-183). Ottawa, Ontario: 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.

Delong, E. R., Delong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*, 837-845.

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

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

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

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*, 449-455. doi:10.3102/0013189X07311600

Embretson, S. E. (Ed.). (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Association.

*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.

Gail, M. H., & Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics, 6*, 227-239. doi:10.1093/biostatistics/kxi005

Grzybowski, M., & Younger, J. G. (1997). Statistical methodology: III. Receiver Operating Characteristics (ROC) curves. *Academic Emergency Medicine, 4*, 818-826.

*Haag, A. M. (2005). Recidivism data from 198 offenders detained until their warrant expiry date. In *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* (UMI No. NR05662). *Dissertations Abstracts International, Section B, 66*(08), 4531B. Unpublished raw data.

Hanley, J. A., & Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology, 4*, 49-58.

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 Rep. No. 1997-04). Ottawa, Ontario: Department of the Solicitor General of Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/cprmindex-eng.aspx#a97

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. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology, 50*, 172-182. doi:10.1037/a0015726

Hanson, R. K., Babchishin, K. M., Helmus, L., & Thornton, D. (2012). *Quantifying the relative risk of sex offenders: Risk ratios for Static-99R*. Manuscript submitted for publication.

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 Rep. No. 2007-05). Ottawa, Ontario: Public Safety Canada. Retrieved 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. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. *International Journal of Forensic Mental Health, 11*, 9-23. doi:10.1080/14999013.2012.667511

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*, 1154-1163. doi:10.1037/0022-006X.73.6.1154

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:1005482921333

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of Static-2002.* (Corrections Research User Rep. No. 2003-01). Ottawa, Ontario: Department of the Solicitor General of Canada. Retrieved 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.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*, 361-387. doi:10.1002/(SICI)1097-0258

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. doi:10.1080/13552600.2010.494772

Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, Ontario: Department of the Solicitor General of Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/2003-03-stc-cde-eng.pdf

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

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York, NY: 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 Dissertations and Theses database. (UMI No. MR58443). Retrieved from http://www.static99.org/pdfdocs/helmus2009-09static-99normsmathesis.pdf

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior, 39*, 1148-1171. doi:10.1177/0093854812443648

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment, 24*, 64-101. doi:10.1177/1079063211409951

*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

Horst, P. (1941). The role of the predictor variables which are independent of the criterion. *Social Science Research Council, 48*, 431-436.

Horton, N. J., & Switzer, S. S. (2005) Statistical methods in the journal. *New England Journal of Medicine, 353*, 1977-1979.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

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

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. *Dissertations 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* (Doc. No. 217618). Submitted to the U.S. Department of Justice.

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

Le, C. T., & Lindgren, B. R. (1995). Construction and comparison of two receiving operating characteristic curves derived from the same sample. *Biometrical Journal, 37*, 869-877. doi:10.1002/bimj.4710370709

Lloyd, M. D. (2008). Incremental validity of commonly-used risk assessment measures in predicting serious sexual recidivism. *Dissertation Abstracts International: Section B. Sciences and Engineering, 69*(9), 5784.

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

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

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

*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.

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

Paulhus, D. L., Robins, R. W., Trzesniewski, K. H., & Tracy, J. L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research, 39*, 303-328. doi:10.1207/s15327906mbr3902_7

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

Pintea, S., & Moldovan, R. (2009). The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9*, 49-66.

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

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. doi:10.1186/1471-2105-12-77

Rockhill, B., Byrne, C., Rosner, B., Louie, M. M., & Colditz, G. (2003). Breast cancer risk prediction with a log-incidence model: Evaluation of accuracy. *Journal of Clinical Epidemiology, 56*, 856-861. doi:10.1016/S0895-4356(03)00124-0

Sasieni, P. D. (2005). Survival analysis. In W. Athrens & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 693-728). Heidelberg, Germany: Springer.

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

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 Behavior, 37*, 613-628. doi:10.1177/0093854810363794

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293. doi:10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

*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. *Dissertations Abstracts International: Section B, 66*(06), 3428.

Thornton, D. (October, 2010). *Interpreting Static-99R and Static-2002R in light of recent research.* Paper presented at the annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Phoenix, AZ.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109*, 524-536. doi:10.1037/0033-2909.109.3.524

Vickers, A. J., Cronin, A. M., & Begg, C. B. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology, 11*(13), 1-7. doi:10.1186/1471-2288-11-13

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

Walters, G. D. (2011). Taking the next step: Combining incrementally valid indicators to improve recidivism prediction. *Assessment, 18*, 227-233. doi:10.1177/1073191110397484

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. (2007). *Circles of support and accountability: A national replication of outcome findings* (Rep. No. R-185). Ottawa, Ontario: Correctional Service of Canada. Retrieved from http://www.csc-scc.gc.ca/text/rsrch/reports/r185/r185-eng.shtml

*Wilson, R. J., Picheca, J. E., & Prinzo, M. (2007). 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

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*, 740-767. doi:10.1037/a0020473

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561-577.