Absolute Recidivism Rates Predicted By Static-99R and Static-2002R Sex Offender Risk Assessment Tools Vary Across Samples : A Meta-Analysis

Leslie Helmus, R. Karl Hanson, David Thornton, Kelly M. Babchishin and Andrew J. R. Harris

Criminal Justice and Behavior 2012 39: 1148 originally published online 21 May 2012
DOI: 10.1177/0093854812443648

ABSOLUTE RECIDIVISM RATES PREDICTED BY STATIC-99R AND STATIC-2002R SEX OFFENDER RISK ASSESSMENT TOOLS VARY ACROSS SAMPLES

A Meta-Analysis

LESLIE HELMUS
Public Safety Canada
Carleton University
R. KARL HANSON
Public Safety Canada
DAVID THORNTON
Sand Ridge Secure Treatment Center
KELLY M. BABCHISHIN
Public Safety Canada
Carleton University
ANDREW J. R. HARRIS
Forensic Assessment Group

There has been considerable research on relative predictive accuracy (i.e., discrimination) in offender risk assessment (e.g., Are high-risk offenders more likely to reoffend than low-risk offenders?), but virtually no research on the accuracy or stability of absolute recidivism estimates (i.e., calibration). The current study aimed to fill this gap by examining absolute and relative risk estimates for certain Static sex offender assessment tools. Logistic regression coefficients for Static-99R and Static-2002R were combined through meta-analysis (8,106 sex offenders; 23 samples). The sexual recidivism rates for typical sex offenders are lower than the public generally believes. Static-99R and Static-2002R both demonstrated remarkably consistent relative predictive accuracy across studies. For both scales, however, the predicted recidivism rates within each

AUTHORS' NOTE: *The views expressed are those of the authors and not necessarily those of Public Safety Canada or the Wisconsin Department of Health Services. Funding for this project was provided in part by the Social Science and Humanities Research Council of Canada and the Ontario Graduate Scholarship Program. We would like to thank Amy Phenix and Dennis Doren for their contributions throughout this project. We would also 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, Jackie Craissati, Margretta Dwyer, Reinhard Eher, Doug Epperson, Tina Garby, Randolph Grace, Steve Gray, Andy Haag, Leigh Harkins, Grant Harris, Andreas Hill, Steve Johansen, Ray Knight, Niklas Långström, Terry Nicholaichuk, Kevin Nunes, Jean Proulx, Martin Rettenberger, Marnie Rice, Steve Saum, Rebecca Swinburne Romine, Daryl Ternowski, Robin Wilson, and Annie Yessine. Correspondence concerning this article should be addressed to Karl Hanson, Corrections and Criminal Justice Research, Public Safety Canada, 10th Floor, 340 Laurier Avenue West, Ottawa, ON, Canada, K1A 0P8; e-mail: karl.hanson@ps.gc.ca.*

CRIMINAL JUSTICE AND BEHAVIOR, Vol. 39 No. 9, September 2012 1148-1171
DOI: 10.1177/0093854812443648

risk score demonstrated large and significant variability across studies. The authors discuss how the variability in recidivism rates complicates the estimation of recidivism probability in applied assessments.

Keywords: risk assessment; recidivism; prediction; sex offenders; absolute risk

Sex offending invokes considerable public concern, particularly when it appears that it should have been predicted and could have been prevented. In the past 20 years, there have been considerable advances in risk assessment for sex offenders. The major risk factors have been identified (Hanson & Morton-Bourgon, 2005; Mann, Hanson, & Thornton, 2010) and combined into structured risk scales demonstrating at least moderate predictive accuracy (Hanson & Morton-Bourgon, 2009). Virtually all of this research, however, has focused on relative risk (also referred to as discrimination by Gail & Pfeiffer, 2005), which considers the extent to which recidivists can be differentiated from nonrecidivists (i.e., Are high-risk offenders more likely to reoffend than low-risk offenders?).

There is surprisingly little research examining calibration, which is the ability of risk scales to estimate absolute recidivism rates (i.e., What is the likelihood of recidivism associated with each risk score?). This is concerning given the importance of absolute recidivism estimates for certain contexts, such as civil commitment in the United States (e.g., see Doren, 2002). Numeric information such as absolute probability estimates is also important for risk communication given the inconsistency with which nominal labels (e.g., "moderate risk") are interpreted by decision makers and clinicians (Hilton, Carter, Harris, & Sharpe, 2008; Monahan & Silver, 2003). For example, different clinical and social policy decisions would likely be made if the expected long-term recidivism rate for a "moderate risk" sex offender was 65% instead of 15%.

Empirically based risk assessment typically starts by considering the base rate, or the expected recidivism rate, of a typical member of the class. This overall rate can then be adjusted up or down based on aggravating or protective factors, thereby creating new, smaller subgroups. In the actuarial approach to risk assessment (Dawes, Faust, & Meehl, 1989), membership in such subgroups (typically represented by a score on a risk scale) is then linked to a table of expected recidivism rates.

Previous research has found that the overall sexual recidivism base rate is lower than commonly expected—often in the 10% to 15% range (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005; A. J. R. Harris & Hanson, 2004). Nevertheless, the rates vary considerably across settings and samples. Figure 1 displays sexual recidivism rates (weighted by sample size) from 53 studies in the Hanson and Bussière (1998) meta-analysis and a random sample of 20 newer studies drawn from Hanson and Morton-Bourgon (2009). For example, large studies in the United States have found rates as low as 3% (Boccaccini, Murrie, Caperton, & Hawes, 2009) and as high as 35% (Knight & Thornton, 2007). Such differences are difficult to interpret, especially given the differences in follow-up time and that many samples were preselected on risk relevant variables (e.g., evaluated for civil commitment).

The purpose of this study was threefold. First, we examined the accuracy with which absolute recidivism rates can be estimated by two of the most commonly used actuarial risk tools for sex offenders, namely Static-99R (Hanson & Thornton, 2000; Helmus, Thornton, Hanson, & Babchishin, 2012) and Static-2002R (Hanson, Helmus, & Thornton, 2010; Helmus et al., 2012). Such analyses have obvious utility for users of these instruments.

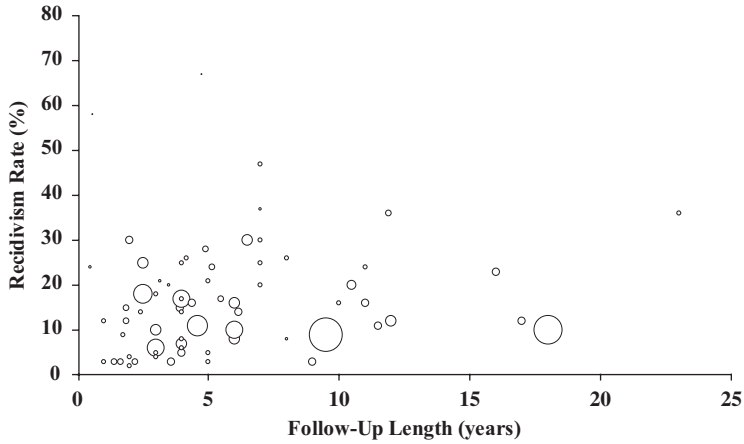

Figure 1: Sexual Recidivism Rates, as a Function of Follow-Up, From 73 Studies (*N* = 35,522)
Note. The size of the bubble is proportional to the sample size of the study.

These analyses also have a second, more general contribution: They provide refined estimates of absolute sexual recidivism rates based on a consistent follow-up period and an explicit definition of risk level. These estimates should be of particular value to decision makers concerned with the extent to which identifiable subgroups of sex offenders exceed certain risk thresholds (e.g., "substantial probability," "more likely than not"). Last, this study contributes to offender risk assessment by advancing our understanding of the absolute risk properties of actuarial scales and proposing a method for examining their stability across diverse samples.

Static-99R and Static-2002R are both actuarial risk assessment scales designed to predict sexual recidivism among adult male sex offenders. The original Static-99 (Hanson & Thornton, 2000) is the most commonly used actuarial scale for sex offenders. It is widely used in Canada and the United States for treatment planning (Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010), community supervision (Interstate Commission for Adult Offender Supervision, 2007), and civil commitment evaluations (Jackson & Hess, 2007). Static-2002 was developed to increase coherence and conceptual clarity compared to Static-99 (see Hanson & Thornton, 2003) and has also become one of the most widely used scales, particularly in Canada (McGrath et al., 2010). Recently, the age item in both scales was revised to better reflect the relationship between age and sexual recidivism (Helmus et al., 2011). The widespread use and validation of Static-99 and Static-2002 affords a unique opportunity to examine the stability of various properties of these actuarial risk scales across samples.

Although there are at least 63 replications of the relative predictive accuracy of Static-99 (Hanson & Morton-Bourgon, 2009), typically using statistics such as the area under the receiver operating characteristic curve or Cohen's *d*, we were able to identify only two studies that explicitly examined the stability of the Static-99 recidivism estimates, with mixed results (Doren, 2004; G. T. Harris et al., 2003). Combining data across seven samples, Doren (2004) found that Static-99 significantly overestimated recidivism for scores of

"4." G. T. Harris and colleagues (2003), however, found that Static-99 significantly underestimated recidivism; for low-risk scores, the observed recidivism rates were approximately twice the predicted rates.

The paucity of research on the stability of absolute recidivism rates is not limited to these particular scales. Most of the recidivism risk scales used in corrections and forensic psychology include recidivism rate tables (e.g., the Violence Risk Appraisal Guide [VRAG; Quinsey, Harris, Rice, & Cormier, 2006], the Sex Offender Risk Appraisal Guide [Quinsey et al., 2006], the Level of Service/Case Management Inventory [Andrews, Bonta, & Wormith, 2004], the Statistical Information on Recidivism - Revised 1 [Nafekh & Motiuk, 2002], the Ontario Domestic Assault Risk Assessment [Hilton et al., 2004], the Minnesota Sex Offender Screening Tool - Revised [Epperson, Kaul, Huot, Goldman, & Alexander, 2003], and STABLE-2007 [Hanson, Harris, Scott, & Helmus, 2007]). However, in contrast to the large number of studies documenting their ability to discriminate between recidivists and nonrecidivists (see meta-analyses by Campbell, French, & Gendreau, 2009; Hanson, Helmus, & Bourgon, 2007; Hanson & Morton-Bourgon, 2009), there have been only a handful of studies examining the stability of the recidivism rate estimates (G. T. Harris, Rice, & Cormier, 2002; G. T. Harris et al., 2003; Mills, Jones, & Kroner, 2005; Snowden, Gray, Taylor, & MacCulloch, 2007).

HOW SHOULD WE EVALUATE ABSOLUTE AND RELATIVE ACCURACY?

Given the limited research attention to absolute recidivism rates, it is not surprising that psychology has yet to develop conventions for testing absolute predictive accuracy. The most commonly recommended statistic for reporting predictive accuracy is the area under the receiver operating characteristic curve (AUC for ROC; Mossman, 1994; Rice & Harris, 2005; Swets, Dawes, & Monahan, 2000). The AUC, however, is a measure of relative accuracy (i.e., the extent to which recidivists differ from nonrecidivists) and is independent of absolute recidivism base rates. Even as an effect size for relative accuracy, the AUC (as well as Cohen's *d*) is influenced by the variance in the scores used to predict recidivism (Hanson, 2008; Humphreys & Swets, 1991), suggesting that other statistics should also be considered.

For absolute accuracy, useful statistics for comparing predicted and observed values include the global chi-square goodness-of-fit significance test (Pearson, 1900) and the E/O effect size index (Gail & Pfeiffer, 2005; Rockhill, Byrne, Rosner, Louie, & Colditz, 2003). When these are nonsignificant (indicating good fit), researchers can have increased confidence in the generalizability of the predicted values. When they are significant, these statistics provide little information concerning the reasons for the poor fit.

In agreement with Mossman (2006), we believe that evaluating the accuracy of absolute recidivism rate estimates for risk tools should consider separately (a) the ability of the test to discriminate between recidivists and nonrecidivists and (b) the stability of base rates across studies and samples. Given that individuals can change from nonrecidivists to recidivists without changing their initial risk scores, the meaning of risk scores must include a sense of enduring relative risk. Inferences concerning absolute recidivism rates must, therefore, combine inferences concerning relative risk with inferences concerning the expected

base rate. Consequently, it is beneficial to use statistical methods that provide separate estimates of the discrimination ability of the scores and the stability of estimated base rates.

Mossman (2006) recommended likelihood ratios as a base-rate independent measure of the detection properties of risk assessment instruments. A likelihood ratio is defined as P(S+|D+)/P(S+|D−), or the probability of a symptom being present among diseased cases divided by the probability of that symptom among nondiseased cases. Likelihood ratios are invariant to disease base rates when the probabilities of symptoms being present are consistent within diseased and within nondiseased cases. In medicine, for example, a patient who becomes diseased would be expected to increase his or her likelihood of displaying a particular symptom. In contrast, offenders become a recidivist without changing their initial risk scores. Consequently, likelihood ratios associated with specific risk scores vary with changes in the recidivism base rate within the same study (e.g., when extending the follow-up period; see similar criticism by G. T. Harris & Rice, 2007). Likelihood ratios for specific risk scores would also be expected to change depending on the distribution of risk scale scores in different samples, even if the probability of recidivism per risk score remains identical. Consequently, likelihood ratios are not a useful indicator for evaluating the stability of relative or absolute estimates.

Logistic regression is a promising approach to estimating base rates while controlling for risk levels (Hosmer & Lemeshow, 2000). In logistic regression, the dichotomous dependent variable (recidivism) is transformed into log odds (also called *logits*). With one predictor variable (Static-99R or Static-2002R), logistic regression estimates two regression coefficients (B_0 and B_1). B_1 (the slope) is an estimate of relative predictive accuracy, or the average change in recidivism rates for each one-unit increase in risk scores, expressed as a log odds ratio. B_0 is an estimate of the recidivism base rate for offenders with a score of 0 (also expressed in logit units). Unlike AUCs, logistic regression B_1 coefficients are not expected to be influenced by restriction of range in the predictor variable (Hanson, 2008).

PURPOSE OF THE CURRENT STUDY

The current study is a meta-analysis examining the stability of the absolute recidivism rates and relative predictive accuracy of Static-99R and Static-2002R across diverse samples and settings. The primary metric used to index the absolute recidivism rates was the logistic regression estimates of the base rates while controlling for risk levels. The logistic regression estimate of the slope was used to examine relative predictive accuracy. Given that the necessary information for this meta-analysis was rarely reported in the original studies, the analyses were conducted using data sets graciously provided by the authors of previous validation studies.

METHOD

MEASURES

Static-99R. Static-99R (Hanson & Thornton, 2000; Helmus et al., 2012) is an empirically derived actuarial risk assessment tool designed to predict sexual recidivism in adult male sex offenders (also see www.static99.org). It has 10 items, and the total score (ranging from

–3 to 12) can be used to place offenders in one of four risk categories: low (–3 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6+). The Static-99R items are identical to those of Static-99 with the exception of updated age weights.

Static-2002R. Similar to Static-99R, Static-2002R (Hanson & Thornton, 2003; Helmus et al., 2012) is an empirical actuarial risk assessment tool for adult male sex offenders (also see www.static99.org). It has 14 items grouped into 5 main subscales: Age at Release, Persistence of Sex Offending, Sexual Deviance, Relationship to Victims, and General Criminality. The total score (ranging from –2 to 14) can be used to place offenders in one of five risk categories: low (–2 to 2), low-moderate (3 to 4), moderate (5 to 6), moderate-high (7 to 8), and high (9+). The items are identical to those of Static-2002 with the exception of updated age weights. Previous research found that Static-2002 was significantly more predictive of sexual, violent, and any recidivism than Static-99 (Hanson et al., 2010). In contrast, Static-99R and Static-2002R have similar predictive accuracy (Babchishin, Hanson, & Helmus, 2011).

SAMPLES

Raw data (typically in the form of SPSS data sets) were obtained by contacting the authors of all Static-99 and Static-2002 replications known to us in 2009. From the original list of 63 studies (Hanson & Morton-Bourgon, 2009), approximately half the data sets were excluded because they did not have information to score all items, age at release, or the dates required to compute fixed follow-up times. Several samples were also excluded because they examined samples outside of the intended target population (e.g., juveniles). Approximately 90% of authors we contacted graciously agreed to share their data. In total, 23 Static-99 samples with sexual recidivism data were obtained, 7 of which also included Static-2002 scores. These samples were used in a related study to develop the new age weights for Static-99R and Static-2002R (Helmus et al., 2011). Although it was not possible to evaluate the quality of the Static ratings in the samples, minimal standards for quality control were applied. Cases were deleted if there were unresolved coding inconsistencies. In addition, as per the coding rules (A. J. R. Harris, Phenix, Hanson, & Thornton, 2003; Phenix, Doren, Helmus, Hanson, & Thornton, 2009), cases were deleted if more than one Static-2002 item was missing, any Static-99 item was missing other than "ever lived with a lover" (Item 2), the offender was younger than 18 years old at the time of release or younger than 16 years old when he or she committed the index offense, or if the offender was female. In addition, cases were deleted if the index sex offense was more than 2 years prior to the current offense because Static-99R and Static-2002R were developed on, and intended for, sex offenders with a current (or recent) sex offense.

Compared to the 63 Static-99 replications in the Hanson and Morton-Bourgon (2009) meta-analysis, the final, cleaned cases included in the current study displayed greater discrimination for the original Static-99 ($d = 0.73$, $k = 23$, $n = 8{,}106$ vs. $d = 0.67$, $k = 63$, $n = 20{,}010$). This was expected given that the current cases were drawn from appropriate populations and had complete information. In contrast, Hanson and Morton-Bourgon (2009) included studies that only approximated Static-99 scores and those that applied the scale to offenders outside of its intended sampling frame (e.g., juveniles).

Tables 1 and 2 provide descriptive information for the studies included. For additional information, readers are encouraged to obtain a more detailed report of this project

TABLE 1: Descriptive Information

Study	n	Age M	Age SD	Country	Recidivism Criteria	Type of Sample	Mostly Treated	Release Period	Mdn Year Release
Allan et al. (2007)	492	42	12	New Zealand	Charges	Prison treatment	Yes	1990–2000	1994
Bartosh et al. (2003)	186	38	12	United States	Charges	Routine correctional	—	1996	1996
Bengtson (2008)	311	33	10	Denmark	Charges	Forensic psychiatric evaluations	—	1978–1995	1986
Bigras (2007)	483	43	12	Canada	Charges	Routine CSC	Mixed	1995–2004	1999
Boer (2003)	299	41	12	Canada	Conviction	Routine CSC	—	1976–1994	1990
Bonta & Yessine (2005)	133	40	10	Canada	Conviction	Preselected high risk	Mixed	1992–2004	1999
Brouillette-Alarie & Proulx (2008)	228	36	10	Canada	Conviction	Prison & community treatment	—	1979–2006	1996
Cortoni & Nunes (2007)	73	42	12	Canada	Charges	Prison treatment	Yes	2001–2004	2003
Craissati et al. (2011)	209	38	12	United Kingdom	Conviction	Routine community supervision	Mixed	1992–2005	1998
Eher et al. (2008, 2009)	706	41	12	Austria	Conviction	Routine European prison	—	2000–2005	2003
Epperson (2003)	177	37	13	United States	Charges	Routine correctional	—	1989–1998	1995
Haag (2005)	198	37	10	Canada	Conviction	Detained until end of sentence	Mixed	1995	1995
Hanson, Harris, et al. (2007)	702	42	13	Canada	Charges	Routine community supervision	—	2001–2005	2002
Harkins & Beech (2007)	197	43	12	United Kingdom	Convictions	Prison & community treatment	Yes	1994–1998	1995
Hill et al. (2008)	86	39	11	Germany	Conviction	Sexual homicide perpetrators	—	1971–2002	1989
Johansen (2007)	273	38	11	United States	Charges	Prison treatment	Yes	1994–2000	1996
Knight & Thornton (2007)	466	36	11	United States	Charges	Civil commitment evaluation	—	1957–1986	1970
Långström (2004)	1,278	41	12	Sweden	Conviction	Routine European prison	No	1993–1997	1995
Nicholaichuk (2001)	281	35	9	Canada	Conviction	High intensity treatment	Yes	1983–1998	1992
Saum (2007)	169	46	12	United States	Charges	Community supervision	Yes	1988–1998	—
Swinburne Romine et al. (2008)	680	38	12	United States	Conviction	Community treatment	Mixed	1977–2007	1988
Ternowski (2004)	247	44	13	Canada	Charges	Prison treatment	Yes	1994–1998	1996
Wilson, Cortoni, et al. (2007) and Wilson, Picheca, et al. (2007)	232	42	11	Canada	Charges	Preselected high risk	—	1994–2007	2002
Total	8,106	40	12	—	—	—	—	1957–2007	1996

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

TABLE 2: Recidivism Information

	Static-99R		*Static-2002R*		*Years Follow-Up*		*Overall*		*5 Years*		*10 Years*	
	M	SD	M	SD	M	SD	n	*Recid (%)*	n	*Recid (%)*	n	*Recid (%)*
Allan et al. (2007)	1.8	2.3	—		5.7	2.9	492	9.6	298	11.7	25	20.0
Bartosh et al. (2003)	3.3	2.9	—		5.0	0.2	186	11.8	90	13.3	—	—
Bengtson (2008)	3.8	2.4	4.6	2.4	16.2	4.2	311	33.8	310	19.7	291	28.5
Bigras (2007)	2.1	2.4	3.5	2.5	4.6	1.9	483	6.2	207	9.2	—	—
Boer (2003)	2.8	2.8	3.9	2.7	13.3	2.1	299	8.7	299	3.7	295	7.8
Bonta & Yessine (2005)	5.0	2.1	—		5.5	2.4	133	15.8	81	17.3	3	0.0
Brouillette-Alarie & Proulx (2008)	3.9	2.3	—		9.9	4.5	228	20.2	199	14.6	110	20.9
Cortoni & Nunes (2007)	2.2	2.1	—		4.6	0.6	73	0.0	17	0.0	—	—
Craissati et al. (2011)	2.2	2.3	—		9.1	2.7	209	11.5	200	7.5	66	9.1
Eher et al. (2008, 2009)	2.3	2.3	—		3.9	1.1	706	4.0	151	2.0	—	—
Epperson (2003)	2.5	2.6	—		7.9	2.5	177	14.1	150	10.7	36	22.2
Haag (2005)	3.9	2.3	5.7	2.3	7.0	0.0	198	25.3	198	19.7	—	—
Hanson, Harris, et al. (2007)	2.4	2.4	3.5	2.5	3.4	1.0	702	8.1	31	0.0	—	—
Harkins & Beech (2007)	2.2	2.6	3.7	2.8	10.4	1.1	197	14.2	197	9.6	127	16.5
Hill et al. (2008)	4.7	2.0	—		12.6	6.6	86	15.1	73	11.0	54	18.5
Johansen (2007)	2.9	2.3	—		9.1	1.1	273	7.7	272	5.9	62	12.9
Knight & Thornton (2007)	4.6	2.4	6.1	2.5	8.6	2.6	466	26.2	433	24.7	353	30.0
Långström (2004)	2.0	2.4	—		8.9	1.4	1,278	7.5	1,278	5.4	353	7.4
Nicholaichuk (2001)	4.8	2.4	—		6.4	4.0	281	18.5	168	22.6	59	25.4
Saum (2007)	1.2	2.3	—		5.0	0.0	169	33.7	169	29.6	—	—
Swinburne Romine et al. (2008)	1.7	2.2	—		16.8	7.8	680	13.8	569	8.4	542	11.3
Ternowski (2004)	1.6	2.5	—		7.5	1.0	247	8.1	247	6.5	—	—
Wilson, Cortoni, et al. (2007) and Wilson, Picheca, et al. (2007)	5.1	2.3	—		5.2	3.0	232	10.3	103	11.7	16	6.3
Overall	2.7	2.6	4.3	2.7	8.2	5.1	8,106	12.4	5,740	11.1	2,392	16.6

Note. Recidivism information is from fixed follow-up periods, not controlling for Static scores.

(Helmus, 2009) or to refer to the original studies. The total sample included 8,106 sex offenders with Static-99R scores and 2,609 with Static-2002R scores. In all, 10 samples were from Canada, 6 were from the United States, and 2 were from the United Kingdom, and there was one each from Denmark, Austria, Sweden, Germany, and New Zealand. Of the 13 studies that could be classified in terms of their treatment status, 7 samples were mostly treated (defined as more than 75% of the offenders), whereas 5 were mixed in their treatment exposure, and only 1 sample was mostly untreated (less than 25%). The average age at release was 40 years old ($SD = 12$). Offenders were released between 1957 and 2007, although 83% were released in 1990 or later. All samples used official criminal records to measure recidivism, but 13 samples used charges as the recidivism criteria and 11 used convictions. Note that either definition underestimates the true rate of recidivism because of underreporting (e.g., Dobash & Dobash, 1995).

Table 2 presents average Static-99R/2002R scores per sample (for Static-99R, $M = 2.7$, $SD = 2.6$; for Static-2002R, $M = 4.3$, $SD = 2.7$). Offenders were followed up for an average of 8.2 years ($SD = 5.1$). Table 2 also includes sexual recidivism rates overall (not controlling for follow-up time) and rates from fixed 5- and 10-year follow-up periods. Note that these data do not control for Static scores. The observed sexual recidivism rate for all cases was 12.4%, with a 5-year rate of 11.1% and a 10-year rate of 16.6%. From the full samples, approximately 70% of cases had at least a 5-year follow-up, whereas only 30% had 10-year follow-up data. Consequently, more fluctuation across samples was expected (and observed) at 10 years than at 5 years.

OVERVIEW OF ANALYSES

All data analyses were conducted independently by the first and fourth authors to ensure accuracy. Analyses examined both relative and absolute risk. To evaluate differences between recidivists and nonrecidivists (discrimination, or relative risk), we used slope coefficients from logistic regression (Hosmer & Lemeshow, 2000). Slope coefficients (i.e., B_1) are log odds ratios but can also be transformed to odds ratios for easier interpretation.

To evaluate absolute predictive accuracy, we used the intercept (i.e., B_0) from logistic regression, which estimates the recidivism rate (as a logit) for offenders with a score of 0. Recentering the scales can produce B_0s that examine predicted base rates for any score. Examining base rates at multiple scores provides a more accurate picture of variability because base rate differences may exist at all scores or at restricted ranges (e.g., only among the highest scores). We used B_0 coefficients centered on Static-99R scores of 0, 2, and 5, and Static-2002R scores of 1, 3, and 6. For Static-99R, the score of 2 was chosen because it is the median (Hanson, Lloyd, Helmus, & Thornton, in press) and, therefore, is considered to describe the "typical" sex offender. Static-99R scores of 0 and 5 were chosen because they are approximately one standard deviation away from the median and, therefore, represent higher and lower ranges of scores while still being a somewhat common (i.e., high frequency) value. Static-2002R scores of 1, 3, and 6 were chosen because they were the closest equivalent (in terms of percentiles) to the Static-99R scores (Hanson et al., in press). For clarity in terminology, $B_{0(0)}$, $B_{0(1)}$, $B_{0(2)}$, $B_{0(3)}$, $B_{0(5)}$, and $B_{0(6)}$ were used to denote intercept coefficients centered on Static-99R/2002R scores of 0, 1, 2, 3, 5, and 6, respectively. For ease of interpretation, the B_0 logits and their confidence intervals were transformed into probabilities (p), where $p = e^{\mathrm{LOGIT}}/(1 + e^{\mathrm{LOGIT}})$.

Logistic regression is appropriate for dichotomous outcome variables where the relationship between the predictor(s) and the outcome follows a logistic distribution. In the sample-level logistic regression analyses, this assumption (tested with the Hosmer and Lemeshow test in SPSS) was met in 43 of 45 analyses. Given that we would expect 2 of 40 tests to be significant merely by chance, the assumption was considered upheld.

For samples in which there were no recidivists, logistic regression coefficients could not be computed. Rather than deleting these low base rate samples, the recidivism base rate (p) was estimated as $1/4n$ (i.e., Bartlett's adjustment; see Cohen, 1988, p. 183; Eisenhart, 1947, §4.3), with an estimated variance of $1/(np(1 - p))$ (see Fleiss, Levin, & Paik, 2003, §2.6). In this formula, n refers to the total number with follow-up information, not just those with a particular score, because if there were any recidivists in the sample, B_0 would not equal zero. For meta-analysis, the proportions were transformed into log odds. B_1 coefficients (discrimination) could not be computed with zero variance in the outcome variable.

Findings across studies were aggregated using fixed-effect and random-effects meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009). Whereas the results of fixed-effect meta-analysis are conceptually restricted to the particular set of studies included in the meta-analysis, random-effects meta-analysis estimates effects for the population of which the current sample of studies is a part (Hedges & Vevea, 1998). More specifically, random-effects meta-analysis incorporates variability across samples into the error term, whereas fixed-effect meta-analysis separates that variability. When variability across studies is low ($Q <$ degrees of freedom), random-effects and fixed-effect meta-analysis produce identical results. As the variability across studies increases, the confidence intervals for random-effects meta-analysis get wider than the fixed-effect results, and the random-effects

method gives more weight to smaller studies. Conceptually, as variability across studies approaches infinity, the random-effects mean approaches the unweighted average.

To test the variability of findings across studies, we used Cochran's Q statistic and the I^2 statistic (Borenstein et al., 2009). A significant Q statistic indicates that there is more variability across studies than would be expected by chance. The Q statistic is distributed as a chi-square with $k - 1$ degrees of freedom (k = the number of studies). Following Hanson and Morton-Bourgon (2009), a finding was considered an outlier if it was the single extreme value and accounted for more than 50% of the total variance (Q).

Although the Q statistic is commonly used to measure variability across studies, it provides a significance test and is dependent on the number of studies included in the analysis, which makes it difficult to compare to other analyses. To describe the magnitude of variability using a k-independent statistic, the I^2 was used. The I^2 statistic describes the proportion of the overall variability (the Q) that is beyond what you would expect by chance from sampling error (i.e., the proportion of variability that can be considered "true" differences; Borenstein et al., 2009). Specifically, I^2 is $(Q - df)/Q$. For easier interpretation, I^2 was reported as a percentage. As a rough heuristic, I^2 values of 25%, 50%, and 75% can be considered low, moderate, and high variability, respectively (Higgins, Thompson, Deeks, & Altman, 2003).

RESULTS

AUCs are currently the most commonly reported and recommended effect size statistic for examining the relative accuracy of risk scales and, therefore, allow comparison with much of the published research on risk scales. Analyzing the combined data set with all cases (k = 23, n = 8,106), the Static-99R significantly predicted sexual recidivism with a moderate to large effect size (AUC = 0.705, 95% CI of 0.688 to 0.722). When the AUCs from each study were meta-analyzed, the average weighted AUC was slightly lower (fixed-effect AUC = 0.693, 95% CI of 0.675 to 0.711; random-effects AUC = 0.695, 95% CI of 0.664 to 0.725, k = 22, n = 8,055; note that Cortoni & Nunes, 2007, was excluded because there were zero sexual recidivists). This difference is likely because AUCs tend to decrease when the variance (s^2) in risk scores decreases (for Static-99R, overall s^2 = 6.9 and median s^2 within studies = 5.5). In addition, there was significant variability in the AUCs across studies (Q = 36.64, p = .018). Given the effect of restriction of range on the AUC values, it is unclear whether this variability represents true variation in the accuracy of Static-99R or a statistical artifact. Consequently, logistic regression was used to summarize relative risk properties of Static-99R and Static-2002R for each sample, as well as absolute properties.

For Static-99R, Tables 3 and 4 present the logistic regression results per sample at 5- and 10-year fixed follow-up periods, with B_0s centered on Static-99R scores of 0, 2, and 5. Figure 2 graphs the variability across samples in the Static-99R $B_{0(2)}$ at 5 years. Table 5 presents logistic regression results at both 5 and 10 years for the samples with Static-2002R scores, with B_0s centered on scores of 1, 3, and 6. The meta-analysis results are displayed in Table 6 (for Static-99R) and Table 7 (for Static-2002R).

STATIC-99R

The weighted average odds ratio for Static-99R at 5 years was 1.34 (95% CI of 1.29 to 1.40 for both fixed-effect and random-effects, k = 21, n = 5,692; see Table 6), meaning that

TABLE 3: Static-99R Logistic Regression Analyses for Sexual Recidivism at 5 Years With B_0s Converted to Percentages

	Static-99R														
	M	SD	N	*% Recid*	B_1	B_1SE	$B_{0(0)}$	$B_{0(0)}$	*95% CI*	$B_{0(2)}$	$B_{0(2)}$	*95% CI*	$B_{0(5)}$	$B_{0(5)}$	*95% CI*
Allan et al. (2007)	1.9	2.3	298	11.7	0.439	0.087	3.9	2.0	7.4	8.8	5.8	13.2	26.6	18.6	36.5
Bartosh et al. (2003)	3.0	2.7	90	13.3	0.137	0.114	8.9	3.4	21.4	11.4	5.9	20.8	16.2	8.9	27.7
Bengtson (2008)	3.8	2.4	310	19.7	0.208	0.065	9.4	5.2	16.3	13.6	9.3	19.3	22.7	17.9	28.2
Bigras (2007)	2.1	2.3	207	9.2	0.354	0.111	3.6	1.5	8.4	7.1	4.1	12.0	18.0	10.9	28.4
Boer (2003)	2.8	2.8	299	3.7	0.478	0.139	0.5	0.1	2.4	1.2	0.4	3.8	4.8	2.6	8.9
Bonta & Yessine (2005)	5.1	2.3	81	17.3	0.358	0.164	2.7	0.4	17.7	5.3	1.3	19.4	14.2	7.6	24.8
Brouillette-Alarie & Proulx (2008)	3.9	2.4	199	14.6	0.332	0.092	3.7	1.4	9.4	6.9	3.6	12.9	16.7	11.8	23.2
Cortoni & Nunes (2007)	2.7	1.8	17	0.0	—	—	1.5	0.03	43.6	1.5	0.03	43.6	1.5	0.03	43.6
Craissati et al. (2011)	2.1	2.3	200	7.5	0.341	0.112	2.9	1.1	7.3	5.6	3.0	10.2	14.2	8.2	23.5
Eher et al. (2008, 2009)	1.9	2.1	151	2.0	1.013	0.401	0.04	<0.01	3.3	0.3	0.02	5.6	5.9	1.7	18.8
Epperson (2003)	2.6	2.6	150	10.7	0.347	0.107	3.4	1.2	9.2	6.6	3.3	12.8	16.6	10.2	26.0
Haag (2005)	3.9	2.3	198	19.7	0.299	0.090	6.2	2.6	14.1	10.7	6.1	18.1	22.7	17.0	29.8
Hanson, Harris, et al. (2007)	2.2	2.6	31	0.0	—	—	0.8	0.02	29.4	0.8	0.02	29.4	0.8	0.02	29.4
Harkins & Beech (2007)	2.2	2.6	197	9.6	0.350	0.097	3.4	1.4	8.1	6.6	3.6	11.8	16.8	10.7	25.5
Hill et al. (2008)	4.8	1.9	73	11.0	0.377	0.231	1.6	0.1	20.1	3.4	0.6	18.5	9.9	4.6	20.0
Johansen (2007)	2.9	2.3	272	5.9	0.208	0.111	3.0	1.2	7.6	4.5	2.4	8.3	8.1	4.8	13.4
Knight & Thornton (2007)	4.6	2.4	433	24.7	0.240	0.052	9.2	5.4	15.3	14.1	10.0	19.6	25.2	21.2	29.7
Långström (2004)	2.0	2.4	1278	5.4	0.308	0.050	2.3	1.5	3.6	4.2	3.2	5.6	10.0	7.8	12.9
Nicholaichuk (2001)	5.0	2.3	168	22.6	0.362	0.099	3.9	1.2	12.0	7.8	3.5	16.4	20.0	14.1	27.4
Saum (2007)	1.2	2.3	169	29.6	0.388	0.085	18.8	12.7	27.0	33.5	26.1	41.9	61.8	46.0	75.4
Swinburne Romine et al. (2008)	1.7	2.2	569	8.4	0.249	0.066	5.0	3.2	7.8	8.0	6.0	10.6	15.5	10.8	21.8
Ternowski (2004)	1.6	2.5	247	6.5	0.296	0.100	3.3	1.5	7.1	5.8	3.4	9.8	13.0	7.4	21.9
Wilson, Cortoni, et al. (2007) and Wilson, Picheca, et al. (2007)	5.2	2.3	103	11.6	0.037	0.136	9.8	2.3	33.7	10.5	3.8	25.4	11.6	6.6	19.4

Note. In samples with no recidivists, B_0 was estimated but B_1 could not be computed.

TABLE 4: Static-99R Logistic Regression Analyses for Sexual Recidivism at 10 Years With B_0s Converted to Percentages

	Static-99R														
	M	SD	N	*% Recid*	B_1	B_1SE	$B_{0(0)}$	$B_{0(0)}$	*95% CI*	$B_{0(2)}$	$B_{0(2)}$	*95% CI*	$B_{0(5)}$	$B_{0(5)}$	*95% CI*
Allan et al. (2007)	1.6	2.1	25	20.0	0.900	0.419	3.1	0.3	27.4	16.4	5.0	42.2	74.4	22.1	96.8
Bengtson (2008)	3.8	2.3	291	28.5	0.230	0.061	13.5	8.0	21.8	19.8	14.4	26.6	33.0	27.2	39.3
Boer (2003)	2.8	2.8	295	7.8	0.381	0.094	1.8	0.7	4.9	3.8	1.9	7.4	11.1	7.4	16.3
Brouillette-Alarie & Proulx (2008)	3.9	2.4	110	20.9	0.261	0.106	7.9	2.8	20.2	12.7	6.5	23.1	24.1	16.4	33.9
Craissati et al. (2011)	1.4	2.1	66	9.1	0.200	0.182	6.5	2.2	18.1	9.4	4.3	19.6	15.9	4.8	41.2
Epperson (2003)	3.7	2.9	36	22.2	0.794	0.290	0.5	0.02	14.8	2.5	0.2	22.7	22.0	8.2	47.0
Harkins & Beech (2007)	2.3	2.6	127	16.5	0.362	0.102	6.1	2.6	13.8	11.9	6.8	19.8	28.5	18.5	41.1
Hill et al. (2008)	5.0	1.8	54	18.5	0.230	0.207	6.3	0.6	41.0	9.7	2.2	34.1	17.7	9.4	30.6
Johansen (2007)	3.6	2.3	62	12.9	0.067	0.165	10.4	2.6	32.9	11.7	4.9	25.5	13.9	6.6	27.1
Knight & Thornton (2007)	4.4	2.5	353	30.0	0.198	0.051	14.6	9.0	22.7	20.2	14.9	26.9	31.5	26.7	36.8
Långström (2004)	2.0	2.5	353	7.4	0.407	0.086	2.2	1.0	4.8	4.8	2.8	8.0	14.6	9.9	21.1
Nicholaichuk (2001)	5.2	2.4	59	25.4	0.234	0.143	8.5	1.6	34.8	12.9	4.1	33.8	23.0	13.6	36.4
Swinburne Romine et al. (2008)	1.7	2.2	542	11.2	0.213	0.061	7.5	5.1	10.8	11.0	8.6	14.0	19.0	13.6	26.0

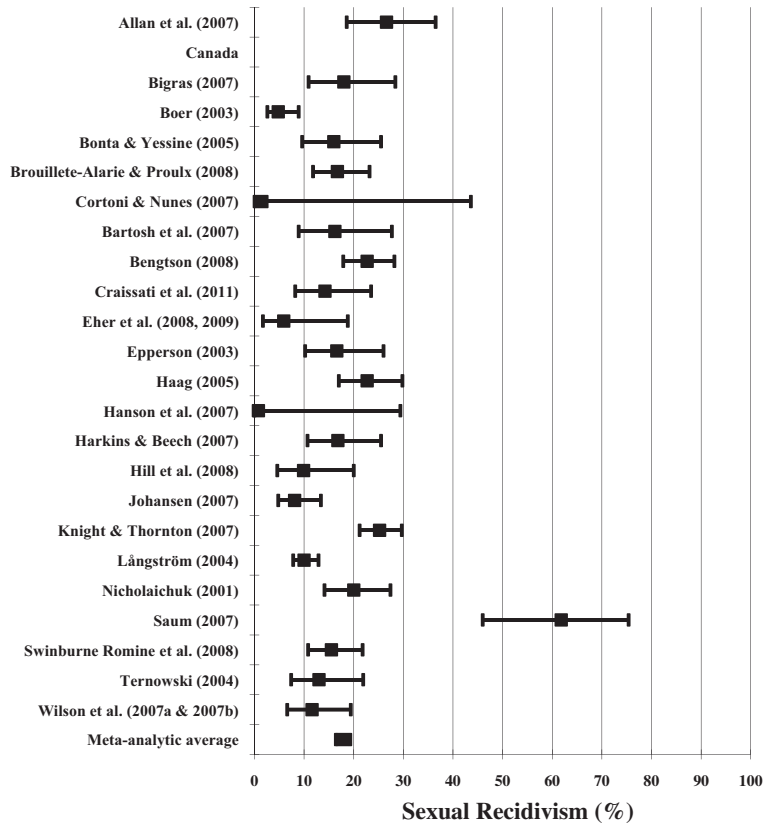

Figure 2: Estimated 5-Year Recidivism Rate for a Static-99R Score of 2 ($B0_{(2)}$) for Each Sample and Combined Weighted Average

TABLE 5: Static-2002R Logistic Regression Analyses for Sexual Recidivism With B_0s Converted to Percentages

	Static-2002R														
	M	SD	N	*% Recid*	B_1	B_1SE	$B_{0(1)}$	$B_{0(1)}$	*95% CI*	$B_{0(3)}$	$B_{0(3)}$	*95% CI*	$B_{0(6)}$	$B_{0(6)}$	*95% CI*
5 years															
Bengtson (2008)	4.6	2.4	307	19.9	0.178	0.060	11.0	6.6	18.0	15.0	10.8	20.6	23.2	18.3	29.0
Bigras (2007)	3.5	2.4	197	8.6	0.386	0.118	2.5	0.9	7.1	5.4	2.7	10.3	15.3	9.4	24.0
Boer (2003)	3.9	2.7	296	3.7	0.492	0.151	0.4	0.08	2.5	1.2	0.4	3.9	5.0	2.7	9.0
Haag (2005)	5.7	2.3	190	18.9	0.308	0.091	4.6	1.6	12.1	8.1	4.1	15.6	18.2	13.2	24.8
Hanson, Harris, et al. (2007)	3.2	2.7	31	0	—	—	0.8	0.02	29.4	0.8	0.02	29.4	0.8	0.02	29.4
Harkins & Beech (2007)	3.7	2.8	190	9.5	0.362	0.097	2.6	0.9	7.0	5.2	2.5	10.2	13.9	8.8	21.2
Knight & Thornton (2007)	6.2	2.5	433	24.7	0.219	0.048	8.9	5.2	15.1	13.2	9.1	18.9	22.7	18.8	27.1
10 years															
Bengtson (2008)	4.6	2.4	288	28.8	0.231	0.058	14.2	8.9	22.0	20.8	15.6	27.2	34.5	28.4	41.1
Boer (2003)	3.9	2.7	293	7.8	0.371	0.099	1.9	0.7	5.2	3.9	2.0	7.6	11.0	7.4	16.2
Harkins & Beech (2007)	3.8	2.7	123	16.3	0.381	0.103	4.6	1.7	11.9	9.4	4.9	17.4	24.7	16.1	35.8
Knight & Thornton (2007)	6.1	2.6	353	30.0	0.183	0.047	13.9	8.4	22.1	18.8	13.4	25.8	28.7	24.0	33.8

Note. In samples with no recidivists, B_0 was estimated but B_1 could not be computed.

TABLE 6: Fixed-Effect Meta-Analysis of Logistic Regression Coefficients for Static-99R for Sexual Recidivism

	Fixed Effect		*Random Effects*						
	M	*95% CI*	M	*95% CI*	Q	I^2*(%)*	*95% CI*	k	n
5 years									
B_1 (*OR*)	1.34	1.29–1.40	1.34	1.29–1.40	20.35	1.7	0.0–48.0	21	5,692
B_0 (%)									
Centered 0	5.2	4.5–6.2	4.3	3.0–6.2	84.35***	73.9	60.7–82.7	23	5,740
Centered 2	8.9	8.0–9.9	7.3	5.3–10.1	155.19***	85.8	80.0–90.0	23	5,740
Centered 5	17.8	16.4–19.3	16.0	12.8–19.8	131.45***	83.3	75.9–88.4	23	5,740
Without Saum									
B_1 (*OR*)	1.34	1.28–1.39	1.34	1.28–1.39	19.09	0.5	0.0–48.2	20	5,523
B_0 (%)									
Centered 0	4.2	3.6–5.1	4.1	3.1–5.4	42.70**	50.8	19.7–69.9	22	5,571
Centered 2	7.4	6.6–8.3	7.0	5.5–8.8	64.59***	67.5	49.3–79.1	22	5,571
Centered 5	17.2	15.8–18.6	15.0	12.3–18.1	92.77***	77.4	66.1–84.9	22	5,571
10 years									
B_1 (*OR*)	1.30	1.23–1.36	1.31	1.23–1.39	15.10	20.5	0.0–58.2	13	2,373
B_0 (%)									
Centered 0	7.7	6.2–9.5	6.5	4.2–9.8	33.38***	64.0	34.9–80.2	13	2,373
Centered 2	12.3	10.7–14.1	10.8	7.8–14.8	47.46***	74.7	56.4–85.3	13	2,373
Centered 5	24.6	22.4–27.1	21.7	16.9–27.4	54.41***	77.9	62.7–87.0	13	2,373

Note. B_{0S} converted to percentage, B_{1s} converted to odds ratios. Total sample size and number of studies differ between B_0 and B_1 because B_1 cannot be computed without recidivists. *OR* = odds ratio.
**$p < .01$. ***$p < .001$.

for each one-point increase in Static-99R score, the odds of recidivism increase by 1.34. The Q indicated nonsignificant variability ($Q = 20.34$, $p = .437$), and the I^2 indicated that only 2% of the variability across studies was more than would be expected by chance (i.e., true variability).

The B_0s, however, showed significant and large variability across samples regardless of where they were centered on Static-99R (see Table 6), with I^2 statistics ranging between 74% and 86%. Among the individual studies (Table 3), the predicted recidivism rate for a Static-99R score of 0 varied between 0.04% and 18.8%, with a weighted average of 5.2% (95% CI of 4.5 to 6.2) from fixed-effect analyses and 4.3% from random-effects analyses (95% CI of 3.0 to 6.2). The predicted recidivism rate for a Static-99R score of 2 varied between 0.3% and 33.5%, with a weighted average of 8.9% (95% CI of 8.0 to 9.9) from fixed-effect analyses and 7.3% from random-effects analyses (95% CI of 5.3 to 10.1). The predicted recidivism rate for a Static-99R score of 5 varied between 0.8% and 61.8%, with a weighted average of 17.8% (95% CI of 16.4 to 19.3) from fixed-effect analyses and 16.0% from random-effects analyses (95% CI of 12.8 to 19.8). For $B_{0(0)}$, $B_{0(2)}$, and $B_{0(5)}$, the highest base rate was from Saum (2007). For $B_{0(0)}$ and $B_{0(2)}$, the lowest base rate was from Eher, Rettenberger, Schilling, and Pfafflin (2008, 2009), whereas for $B_{0(5)}$, the lowest base rate was from Hanson, Harris, and colleagues (2007). In these base rate analyses, more than half of the observed values fell outside the 95% confidence interval for the random-effects meta-analysis.

One study (Saum, 2007) appeared to be a statistical outlier in the B_0 analyses but not the B_1 analyses (see Figure 2). This study accounted for approximately 50% of the Q statistic in the analyses of $B_{0(0)}$ and $B_{0(2)}$. Although it did not account for 50% of the variability across studies for the $B_{0(5)}$, the confidence interval from Saum (2007) did not overlap with any of the confidence intervals from the other studies. A possible explanation for the unusually high recidivism rate in Saum's (2007) study is selective data retention. This study involved archival file coding from treatment settings, but based on the cohort of offenders and the completion date for the thesis, all files were likely 6 to 16 years old at the time of data collection. Saum (2007) noted that some treatment programs lost or destroyed some of their records in the process of changing locations. It is plausible that the purged records disproportionately included inactive files (i.e., nonrecidivists). Disproportionate data purging in Canadian national records has been found to artificially increase base rates when records are retrospectively sampled (Hanson & Nicholaichuk, 2000).

Analyses were rerun with Saum (2007) excluded. From Table 6, the B_1 results were virtually unaltered, although the aggregated B_0s were slightly lower. Most notably, the variability in the B_0s across studies was reduced but was still significant for $B_{0(0)}$, $B_{0(2)}$, and $B_{0(5)}$. The I^2 statistics changed from a range of 74% to 86% to 51% to 77%, indicating a clear reduction, but still moderate to large variability across studies. There were no cases from Saum (2007) for the 10-year follow-up analyses or for the Static-2002R analyses.

Fewer samples had follow-up information at 10 years ($k = 13$) and most of the samples had substantially fewer cases ($n = 2{,}373$). At 10 years, the weighted average odds ratio for Static-99R was 1.30 (95% CI of 1.23 to 1.36) for fixed-effect analyses, and 1.31 (95% CI of 1.23 to 1.39) for random-effects analyses ($k = 13$, $n = 2{,}373$; see Table 6). The variability across studies for relative predictive accuracy was small and nonsignificant ($Q = 15.10$, $p = .236$; $I^2 = 20\%$). One study (Wilson, Cortoni, & Vermani, 2007; Wilson, Picheca, & Prinzo, 2007) was excluded from the 10-year analyses because the logistic regression model (with 16 cases and only 1 recidivist) was misspecified.

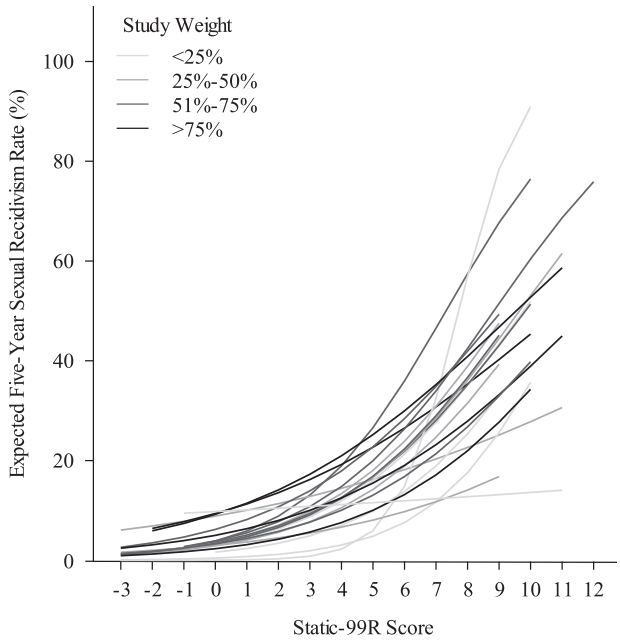

Figure 3: Logistic Regression Results From Each Sample
Note. Sample weight was computed using the inverse of the variance of the B_1 and divided into four categories: (a) studies with lowest weight (bottom 25%), (b) studies with low average weights (25%–50%), (c) studies with higher weight (51%–75%), and (d) studies with the largest amount of weight (top 75%).

Similar to the findings at 5 years, the B_0s showed significant variability across samples regardless of where they were centered on Static-99R (see Table 6), with I^2 statistics ranging between 64% and 78%, demonstrating large variability across studies. Among the individual studies, the predicted 10-year recidivism rate for a Static-99R score of 0 varied between 0.5% and 14.6%, with a weighted average of 7.7% (95% CI of 6.2 to 9.5) from fixed-effect analyses and 6.5% from random-effects analyses (95% CI of 4.2 to 9.8). The predicted recidivism rate for a Static-99R score of 2 varied between 2.5% and 20.2%, with a weighted average of 12.3% (95% CI of 10.7 to 14.1) from fixed-effect analyses and 10.8% from random-effects analyses (95% CI of 7.8 to 14.8). The predicted recidivism rate for a Static-99R score of 5 varied between 11.1% and 74.4%, with a weighted average of 24.6% (95% CI of 22.4 to 27.1) from fixed-effect analyses and 21.7% from random-effects analyses (95% CI of 16.9 to 27.4). For $B_{0(0)}$ and $B_{0(2)}$, the highest base rate was from Knight and Thornton (2007) and the lowest base rate was from Epperson (2003), both samples from the United States. For $B_{0(5)}$, the lowest base rate was from Boer (2003) and the highest was from Allan, Grace, Rutherford, and Hudson (2007).

Figure 3 graphs the B_1 and B_0 results across studies by plotting the predicted 5-year recidivism rates per score on the Static-99R. Saum (2007) was excluded. Consistent with the finding of nonsignificant variability in the B_1s, the shape of the slope appears relatively consistent across samples. In contrast, the slopes are clearly separated on the *y* axis, demonstrating different predicted recidivism rates even within the same Static-99R score. The variability was noticeably higher for the high Static-99R scores compared to the low scores, which is consistent with the higher I^2 values for $B_{0(2)}$ and $B_{0(5)}$, compared to $B_{0(0)}$.

TABLE 7: Fixed-Effect Meta-Analysis of Logistic Regression Coefficients for Static-2002R for Sexual Recidivism

	Fixed Effect		*Random Effects*						
	M	*95% CI*	M	*95% CI*	Q	I^2*(%)*	*95% CI*	k	n
5 years									
B_1 (*OR*)	1.29	1.22–1.37	1.32	1.22–1.43	7.46	33.0	0.0–73.0	6	1,613
B_0 (%)									
Centered 1	6.3	4.6–8.6	4.0	1.9–8.0	22.77***	73.6	43.6–87.7	7	1,644
Centered 3	10.1	8.2–12.4	6.8	3.9–11.8	29.82***	79.9	58.9–90.1	7	1,644
Centered 6	19.1	16.9–21.5	15.4	10.8–21.5	33.01***	81.8	63.6–90.9	7	1,644
10 years									
B_1 (*OR*)	1.27	1.19–1.35	1.30	1.18–1.42	5.13	41.5	0.0–80.3	4	1,057
B_0 (%)									
Centered 1	10.2	7.4–13.8	7.3	3.2–15.9	17.07***	82.4	57.8–93.2	4	1,057
Centered 3	15.6	12.8–19.0	11.8	6.1–21.8	25.00***	88.0	71.6–94.9	4	1,057
Centered 6	26.9	23.9–30.2	23.7	15.2–35.0	30.15***	90.0	77.4–95.6	4	1,057

Note. B_0s converted to percentage, B_1*S* converted to odds ratios. Total sample size and number of studies differ between B_0 and B_1 because B_1 cannot be computed without recidivists. *OR* = odds ratios.
***$p < .001$.

STATIC-2002R

The 5- and 10-year findings for Static-2002R (presented for individual samples in Table 5 and aggregated in Table 7) are similar to the pattern of findings for Static-99R, but with substantially fewer studies. At 5 and 10 years, there was no significant variability in the B_1s, with odds ratios for Static-2002R ranging between 1.27 (fixed-effect results at 10 years) and 1.32 (random-effects results at 5 years). The odds ratios for Static-99R and Static-2002R cannot be compared easily because they are influenced by the scale of the predictor. Static-2002R has a wider range of possible values, which would be expected to result in smaller odds ratios (because it represents the increase in the odds of recidivism for each one-unit increase on the Static scale).

The I^2 values for the B_1 were 33% at 5 years and 42% at 10 years, which represent small to moderate variability, which is higher than the amount of variability found for Static-99R (although these findings are based on only 6 and 4 studies, respectively). Nonetheless, the variability was still nonsignificant and substantially lower than the observed variability in the B_0s. There was significant variability in the $B_{0(1)}$, $B_{0(3)}$, and $B_{0(6)}$ for Static-2002R in both 5- and 10-year analyses. The I^2 values ranged between 74% and 90%, reflecting large variability across studies and slightly more variability than was found in the Static-99R B_0 analyses. No outliers were identified.

DISCUSSION

The current study demonstrates the value of separating relative risk from absolute risk for criterion-referenced measures of sex offender recidivism risk. The two measures examined in this study (Static-99R, Static-2002R) were remarkably consistent across diverse settings and samples when used to determine which sex offenders were riskier than others (i.e., discrimination, or relative predictive accuracy). In contrast, there was substantial

variation in the absolute recidivism rates associated with the same risk score (i.e., calibration) across the 23 samples examined in this meta-analysis. The range in absolute recidivism rates across studies was sufficiently large that values within the observed range could lead to meaningfully different conclusions concerning an offender's likelihood of recidivism.

The results also highlight the utility of using logistic regression to separate the discriminative properties of a measure from selection and base rate effects. Based on the standard AUC analyses (which do not control for selection effects), the between-sample variation in relative predictive accuracy for Static-99R was moderate (I^2 = 42.6%); once selection effects were controlled with logistic regression, the amount of between-sample variation was trivial (I^2 = 1.7%). Consequently, we believe that B_1 coefficients from logistic regression are a more robust effect size indicator for relative predictive accuracy than AUCs and should be routinely reported.

The stability of the relative risk estimates suggests that the same risk factors are important for diverse samples of sex offenders. This is not surprising given that the items for Static-99R and Static-2002R were developed based on previous meta-analyses of large, diverse samples (e.g., Hanson & Bussière, 1998). The stability of the relative risk estimates also supports the cumulative, stochastic model used to combine the items. Rather than considering complex interactions, items are assumed to be additive; each unit increase in the score is expected to be associated with approximately the same increase in the risk of recidivism (on a logit scale). In the current study, a one-unit increase was associated with an increase of approximately 1.3 in the odds of recidivism (for both Static-99R and Static-2002R).

The stability in relative risk was not matched, however, by a similar stability in absolute recidivism rates. For example, the predicted 10-year sexual recidivism rate for a Static-99R score of 2 was as low as 3% in some samples and as high as 20% in other samples. Similarly, the Static-99R score associated with a 15% predicted recidivism rate after 5 years ranged from 2 to 8 (a difference greater than 2 standard deviations). The reasons for this variation are not fully understood. Our preliminary analyses suggest that relatively small amounts of this base rate variation can be explained by cohort effects (i.e., year of release), country, recidivism criteria, quality of recidivism information, offender type, or treatment participation (Helmus, 2009). Of the possible explanations examined so far, the most promising appears to be systematic differences in the density of unmeasured risk factors external to Static-99R and Static-2002R. The highest base rates were found, for example, in samples that were explicitly preselected as high risk (e.g., Bengtson, 2008; Haag, 2005; Knight & Thornton, 2007); in contrast, the lowest recidivism rates were observed in routine, unselected samples (e.g., Boer, 2003; Långström, 2004). Further analyses of the sources of the base rate variability are ongoing.

We do not believe that the base rate variation observed in the current study is unique to Static-99R or Static-2002R, or to the original Static-99 (see Helmus, 2009) or Static-2002 (see Hanson et al., 2010). To our knowledge, the only other scale with multiple validations of the absolute recidivism estimates is the VRAG. Two studies found nonsignificant differences between the estimated and observed recidivism rates (G. T. Harris et al., 2002; G. T. Harris et al., 2003); however, both studies used samples of offenders who were similar in jurisdiction and setting to the samples on which the scale was developed (G. T. Harris, Rice, & Quinsey, 1993). Two other studies found that VRAG estimates were higher than the observed rates, although the follow-up periods in both studies were shorter than VRAG norms. Neither of these independent, external replications used goodness-of-fit statistical

tests (Mills et al., 2005; Snowden et al., 2007). Further examination of the absolute risk properties of other actuarial scales is sorely needed.

Despite the variation in base rates, the absolute recidivism rate for the typical sex offender (defined as the median value) did not exceed 15% after 5 years for all samples (with the exception of one outlier previously discussed). Based on the most extreme values of the 95% confidence intervals for Static-99R or Static-2002R (random effects), a plausible range for the 5-year recidivism rate for the typical sex offender would be between 4% and 12% and, for 10 years, 6% to 22%. Most sex offenders would be expected to have 5-year sexual recidivism rates of 7% or less. These values are lower than the 10% to 15% sexual recidivism rate after about 5 years found in previous meta-analyses (e.g., Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005; A. J. R. Harris & Hanson, 2004).

A plausible explanation for the higher estimates in previous studies is that they oversampled from higher risk settings. This could be an issue of convenience; good research requires comprehensive information, which is more readily available for offenders serving long sentences or subject to special measures (e.g., high intensity treatment, civil commitment). For example, criminal court statistics for Canada suggest that no more than 11% of all convicted sex offenders receive prison sentences greater than 2 years (typically reserved for serious offenses or offenders with lengthier criminal histories) or are selected for specialized measures for high-risk offenders (Canadian Centre for Justice Statistics, 2008), yet 60% of the offenders from the 10 Canadian samples in the current study fell into this category. In support of this interpretation, readers will notice that the unadjusted recidivism average across all the samples in the current study was 11% after 5 years and 17% after 10 years (see Table 2). These values are very similar to the unadjusted averages found in previous research summaries. Nevertheless, even the high end of the range of plausible recidivism rates is considerably lower than the public commonly believes (Levenson, Brannon, Fortney, & Baker, 2007).

IMPLICATIONS FOR RESEARCHERS

For researchers specifically concerned with predicting sex offender recidivism, the next obvious step is to explain the between-sample variation in base rates. More generally, the stability of absolute recidivism rate estimates is an important topic for offender risk prediction. Almost all previous research in offender risk assessment has focused solely on relative risk (e.g., *r*, Cohen's *d*, AUC). Conventions have yet to be developed for evaluating the accuracy of recidivism rate estimates. Our research team (e.g., Helmus et al., 2012) has drawn some inspiration from the research on predicting breast cancer, where the distinction between relative and absolute prediction is considerably advanced (Bondy, Lustbader, Halabi, Ross, & Vogel, 1994; Costantino et al., 1999; Gail et al., 1989; Spiegelman, Colditz, Hunter, & Hertzmark, 1994). In particular, the E/O index (the ratio of expected to observed values) has considerable intuitive appeal (Gail & Pfeiffer, 2005; Rockhill et al., 2003) and can be used to examine the prediction of recidivism (for an example, see Helmus et al., 2012).

The current results demonstrate the value of using logistic regression to separate the discriminative properties of a measure from the sample base rate. The generalizability of each of these parameters can then be tested using conventional meta-analytic techniques (Borenstein et al., 2009; Hanson & Broom, 2005). As well, there may be circumstances in which researchers would want to simultaneously test both parameters using multivariate meta-analysis (Arends, 2006).

Further research is needed to determine the extent to which the general results are unique to these measures. Other fields of study have similarly found stability in relative risk, but not absolute rates (e.g., car accidents [Ingre et al., 2006; Rosén, Stigson, & Sander, 2011]; suicide risk [Paterson et al., 2008]). Consequently, it is possible that this pattern is based on more general psychometric principles. Certain criterion-referenced measures may intrinsically index relative risk, with base rates being largely determined by external factors, such as site-specific variation in the difficulty of preventing (or detecting) the outcome of interest. Alternately, another possibility is simply that the incorporation of additional risk factors in these scales may adequately account for this variability.

IMPLICATIONS FOR PRACTITIONERS

The finding of meaningful variation in absolute recidivism rates complicates the interpretation of these Static risk measures. The original study of Static-99 did not find significant differences in the overall recidivism rates across three development samples (Hanson & Thornton, 2000); consequently, there was only one recidivism estimates table linking the scores to recidivism rates (N = 1,086). This turned out to be a gross simplification. The current findings indicate that evaluators cannot, in an unqualified way, associate a single reliable recidivism estimate with a single score on the Static-99/R or Static-2002/R risk scales. Evaluators interested in reporting absolute recidivism rate estimates must not only calculate a Static score but also make a separate professional judgment concerning which sample the offender most closely resembles (resources for this task are currently available from www.static99.org).

For the lower scores, the observed differences in the recidivism rates were not huge and may result in the same applied decisions regardless of the reference group used (all estimates were less than 10% at 5 years). For the higher scores, however, the absolute differences were often large enough to matter. For example, the 5-year sexual recidivism for a score of 5 was 10% in a large, representative sample of Swedish sex offenders (Långström, 2004), compared to a rate of 25% in a large U.S. civil commitment sample (Knight & Thornton, 2007). The absolute differences for more extreme scores were even larger.

One approach to addressing this variability would be to collect local norms. Given that many jurisdictions routinely collect (and digitize) Static-99 scores for administrative purposes, the research could be conducted with relatively little cost. Local norms, however, may not be superior to those derived from meta-analytic averages. For researchers constructing local norms, we recommend that revisions to the norms be based on at least 100 sexual recidivists from a clearly defined sample and that the reliability of both the Static scoring and the recidivism information be verified.

Another approach to addressing base rate variability is to ignore it. In many contexts, precise estimates of absolute risk are not needed because decisions can be based on relative risk (which was found to be stable). For example, probation officers may have the resources to conduct home visits for only 20% of the highest risk offenders. In this context, other quantitative metrics for communicating risk (such as percentile ranks; Hanson et al., in press) may have more utility than absolute recidivism rate estimates.

Our view, however, is that no single Static table will ever fully capture the potential range of recidivism risk. The Static measures were never intended to measure all relevant risk factors. These external risk factors need to be considered when estimating absolute recidivism rates. Review of the recidivism base rates suggests that samples can be clustered

into three qualitatively different groups (see www.static99.org). The highest recidivism rates are found in samples that have been explicitly preselected on risk relevant variables. The lowest recidivism rates are found in routine correctional samples with little or no preselection (e.g., consecutive cases). Samples of offenders referred for treatment demonstrate recidivism rates intermediate between these two groups (i.e., routine and high risk/high need).

Our interpretation is that this between-sample variability can be largely explained by unmeasured, external risk factors (Thornton, Hanson, & Helmus, 2010). These risk factors could include those discussed by Mann and colleagues (2010; e.g., sexual preoccupation, deviant sexual interest, offense-supportive attitudes, emotional congruence with children, self-regulation problems) as well as the "Central 8" risk factors identified for general offenders (e.g., substance abuse, personality disorder; Andrews & Bonta, 2010). The incremental contribution of external risk factors to Static-99 is well established (e.g., Allan et al., 2007; Beech, Friendship, Erikson, & Hanson, 2002; Eher, Rettenberger, Matthes, & Schilling, 2010; Thornton, 2002), and it is possible to create large within-sample differences in absolute recidivism rates based on the density of risk factors external to Static-99 (e.g., Eher, Matthes, Schilling, Haubner-MacLean, & Rettenberger, 2012; Hanson, Harris, et al., 2007; Olver, Wong, Nicholaichuk, & Gordon, 2007; Thornton, 2010).

The implication of this interpretation is that evaluators seriously interested in estimating absolute recidivism rates must extend their assessment beyond these Static measures to consider other important, risk-relevant factors. Given the superiority of structured judgment over unstructured professional judgment, we recommend that decisions concerning the density of external risk factors also be based on structured, empirically validated risk tools (e.g., Eher et al., 2011; Olver et al., 2007; Thornton, 2010).

REFERENCES

References marked with an asterisk indicate studies included in the analyses. Wilson, Cortoni, et al. (2007) and Wilson, Picheca, et al. (2007) both have an asterisk, although they were combined into one sample for the current study. As well, Eher et al. (2008) and Eher et al. (2009) represented only one sample for the current study.

*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

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Newark, NJ: LexisNexis/Anderson.

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

Arends, L. R. (2006). *Multivariate meta-analysis: Modelling the heterogeneity*. Haveka, Netherlands: Alblasserdam. Retrieved from http://repub.eur.nl/res/pub/7845/Proefschrift%20Lidia%20Arends.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, ON: Public Safety Canada.

*Bartosh, D. L., Garby, T., & Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. *International Journal of Offender Therapy and Comparative Criminology*, *47*, 422-438. doi:10.1177/0306624X03253850

Beech, A., Friendship, C., Erikson, M., & Hanson, R. K. (2002). The relationship between static and dynamic risk factors and reconviction in a sample of U.K. child abusers. *Sexual Abuse: A Journal of Research and Treatment*, *14*, 155-167. doi:10.1177/107906320201400206

*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*, *68*(09), NR30941.

Boccaccini, M. T., Murrie, D. C., Caperton, J. D., & Hawes, S. W. (2009). Field validity of the Static-99 and MnSOST-R among sex offenders evaluated for civil commitment as sexually violent predators. *Psychology*, *Public Policy*, *and Law*, *15*, 278-314. doi:10.1037/a0017232

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

Bondy, M. L., Lustbader, E. D., Halabi, S., Ross, E., & Vogel, V. G. (1994). Validation of a breast cancer risk assessment model in women with a positive family history. *Journal of the National Cancer Institute*, *86*, 620-625.

*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: Public Safety and Emergency Preparedness Canada]. Unpublished raw data.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

*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.

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

Canadian Centre for Justice Statistics. (2008). [*Adult Criminal Court Survey*. Ottawa, ON: Author]. Unpublished raw data.

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

*Cortoni, F., & Nunes, K. L. (2007). *Assessing the effectiveness of the National Sexual Offender Program* (Research Rep. No. R-183). Unpublished report, Correctional Service of Canada, Ottawa.

Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., & Wieand, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*, *91*, 1541-1548.

*Craissati, J., Bierer, K., & South, R. (2011). Risk, reconviction, and "sexually risky behaviour" in sex offenders. *Journal of Sexual Aggression*, *17*, 153-165. doi:10.1080/13552600.2010.490306

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

Dobash, R. P., & Dobash, R. E. (1995). Reflections on findings from the Violence Against Women survey. *Canadian Journal of Criminology*, *37*, 457-484.

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

Doren, D. M. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 25-36.

Eher, R., Matthes, A., Schilling, F., Haubner-MacLean, T., & Rettenberger, M. (2012). Dynamic risk assessment in sexual offenders using STABLE-2000 and the STABLE-2007: An investigation of predictive and incremental validity. *Sexual Abuse: A Journal of Research and Treatment, 24*, 5-28. doi: 10.1177/1079063211403164

Eher, R., Rettenberger, M., Matthes, A., & Schilling, F. (2010). Stable dynamic risk factors in child sexual abusers: The incremental predictive power of narcissistic personality traits beyond the Static-99/Stable-2007 priority categories on sexual reoffense. *Sexual Offender Treatment*, *5*(1), 1-12. Retrieved from http://www.sexual-offender-treatment.org/2-2010_02.html

Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2008). Failure of Static-99 and SORAG to predict relevant reoffense categories in relevant sexual offender subtypes: A prospective study. *Sexual Offender Treatment*, *8*(1), 1-20. Available from http://www.sexual-offender-treatment.org/sot-1-2008.html

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

Eisenhart, C. (1947). Inverse sine transformation of proportions. In C. Eisenhart, M. W. Hastay, & W. A. Wallis, (Eds.), *Selected techniques of statistical analysis for scientific and industrial research and production and management engineering* (pp. 395-416). New York, NY: McGraw-Hill.

*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, Bismarck.

Epperson, D. L., Kaul, J. D., Huot, S., Goldman, R., & Alexander, W. (2003). *Minnesota Sex Offender Screening Tool–Revised (MnSOST-R) technical paper: Development*, *validation*, *and recommended risk level cut scores*. St. Paul: Minnesota Department of Corrections. Retrieved from http://blueshifthome.com/recidivism/scales/minn/Minn%20-%20MnSOST-R%20Technical%20Paper%20Development%20validation%20and%20Recommended%20Risk%20Level%20Cut%20Scores%20-%20Dec%202003.pdf

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, *81*, 1879-1886.

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

*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. *Dissertations Abstracts International*, *66*(08), 4531B. (UMI No. NR05662)]. Unpublished raw data.

Hanson, R. K. (2008). What statistics should we use to report predictive accuracy. *Crime Scene*, *15*(1), 15-17.

Hanson, R. K., & Broom, I. (2005). The utility of cumulative meta-analysis: Application to programs for reducing sexual violence. *Sexual Abuse: A Journal of Research and Treatment*, *17*, 357-373. doi:10.1007/s11194-005-8049-1

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, *66*(2), 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, ON: Public Safety Canada.

Hanson, R. K., Helmus, L., & Bourgon, G. (2007). *The validity of risk assessments for intimate partner violence: A meta-analysis* (Corrections Research User Rep. No. 2007-07). Ottawa, ON: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/vra_ipv_200707_e.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., Helmus, L., & Thornton, D. (in press). 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*.

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.

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.

Hanson, R. K., & Nicholaichuk, T. (2000). A cautionary note regarding Nicholaichuk et al. (2000). *Sexual Abuse: A Journal of Research and Treatment*, *12*, 289-293.

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, ON: Department of the Solicitor General of Canada.

*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. (2004). *Sex offender recidivism: A simple question* (User Report 2004-03). Ottawa, ON: Public Safety and Emergency Preparedness Canada.

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

Harris, G. T., & Rice, M. E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior*, *34*, 1638-1658. doi:10.1177/0093854807307029

Harris, G. T., Rice, M. E., & Cormier, C. A. (2002). Prospective replication of the *Violence Risk Appraisal Guide* in predicting violent recidivism among forensic patients. *Law and Human Behavior*, *26*, 377-394.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior*, *20*, 315-335. doi:10.1177/0093854893020004001

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.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.

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). Also available from www.static99.org

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: Journal of Research and Treatment, 24*, 64-101 doi:10.1177/1079063211409951.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557-560.

*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

Hilton, N. Z., Carter, A., Harris, G. T., Sharpe, A. J. B. (2008). Does using nonumerical terms to describe risk aid violence risk communication? *Journal of Interpersonal Violence*, *23*, 171-188. doi:10.1177/0886260507309337

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario Domestic Assault Risk Assessment. *Psychological Assessment*, *16*, 267-275.

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

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, G., & Pickles, A. (2006). Subjective sleepiness and accident risk: Avoiding the ecological fallacy. *Journal of Sleep Research*, *15*, 142-148. doi:10.1111/j.1365-2869.2006.00517.x

Interstate Commission for Adult Offender Supervision. (2007). *Sex offender assessment information survey* (ICAOS Doc. 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. *Dissertations Abstracts International*, *68*(03), 3255527B.

*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). Washington, DC: 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

Levenson, J. S., Brannon, Y. N., Fortney, T., & Baker, J. (2007). Public perceptions about sex offenders and community protection policies. *Analyses of Social Issues and Public Policy*, *7*, 137-161.

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.* Brandon, VT: Safer Society Foundation.

Mills, J. F., Jones, M. N., & Kroner, D. G. (2005). An examination of the generalizability of the LSI-R and VRAG probability bins. *Criminal Justice and Behavior*, *32*, 565-585. doi:10.1177/0093854805278417

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal of Forensic Mental Health*, *2*, 1-6.

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

Mossman, D. (2006). Another look at interpreting risk categories. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 41-63.

Nafekh, M., & Motiuk, L. L. (2002). *The Statistical Information on Recidivism–Revised 1 (SIR-R1) scale: A psychometric examination* (Research Rep. No. R-126). Ottawa, ON: Correctional Service of Canada.

*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.

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

Paterson, B., Dowding, D., Harries, C., Cassells, C., Morrison, R., & Niven, C. (2008). Managing the risk of suicide in acute psychiatric inpatients: A clinical judgement analysis of staff predictions of imminent suicide risk. *Journal of Mental Health*, *17*, 410-423. doi:10.1080/09638230701530234

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *Series 5*, *50*, 157-175. doi:10.1080/14786440009463897

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

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.

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.

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

Rosén, E., Stigson, H., & Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis and Prevention*, *43*, 25-33. doi:10.1016/j.aap.2010.04.003

*Saum, S. (2007). A comparison of an actuarial risk prediction measure (Static-99) and a stable dynamic risk prediction measure (Stable-2000) in making risk predictions for a group of sexual offenders. *Dissertations Abstracts International*, *68*(03), 3255539B.

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

Spiegelman, D., Colditz, G. A., Hunter, D., & Hertzmark, E. (1994). Validation of the Gail et al. model for predicting individual breast cancer risk. *Journal of the National Cancer Institute*, *86*, 600-607.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1-26. doi:10.1111/1529-1006.001

*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*, *66*(06), 3428B.

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment*, *14*, 139-153. doi:10.1023/A:1014620214905

Thornton, D. (2010, October). *Can variation in known psychological risk factors account for the effect of pre-selection on recidivism base rates?* Paper presented at the 29th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Phoenix, AZ.

Thornton, D., Hanson, R. K., & Helmus, L. (2010, Spring). Moving beyond the standard model for actuarial assessment for sexual offenders. *Perspectives: Quarterly Newsletter of the California Coalition on Sexual Offending*, p. 1-4. Retrieved from http://www.ccoso.org/newsletter.php

*Wilson, R. J., Cortoni, F., & Vermani, M. (2007). *Circles of support and accountability: A national replication of outcome findings* (Rep. No. R-185). Ottawa, ON: Correctional Service of Canada.

*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. *Howard Journal*, *46*, 327-337. doi:10.1111/j.1468-2311.2007.00480.x

Leslie Helmus, MA, is a PhD student at Carleton University in Ottawa, and completed this study while working in the Corrections Research Unit at Public Safety Canada. Her research has focused on sex offender risk assessment, where she has been involved in developing and renorming several risk assessment tools.

R. Karl Hanson, PhD, is a senior research scientist with Public Safety Canada, and adjunct professor in the Psychology Department of Carleton University, Ottawa. He has been researching and developing assessment measures for sexual offenders for more than 20 years.

David Thornton, PhD, works part-time as treatment director at Sand Ridge Secure Treatment Center in Wisconsin and is also a professor in the department of clinical psychology in Bergen (Norway). He has been researching and developing assessment measures and treatment programs for sexual and violent offenders since 1990.

Kelly M. Babchishin, MA, is a graduate student in forensic psychology at Carleton University and a research assistant at Public Safety Canada. Her research interests include pedophilia, online sexual offending, and more applied areas, such as sex offender risk assessment and treatment.

Andrew J. R. Harris, PhD, CPsych, is director of the Forensic Assessment Group, Ottawa, Canada. He has worked in both clinical and research positions for the province of Ontario and in the federal correctional system. He teaches and consults on research methodology and sexual and violent risk assessment.