Sexual Abuse: A Journal of Research and Treatment
http://sax.sagepub.com/

Development of Vermont Assessment of Sex Offender Risk-2 (VASOR-2) Reoffense Risk Scale

Robert J. McGrath, Michael P. Lasher, Georgia F. Cumming, Calvin M. Langton and Stephen E. Hoke

Sex Abuse 2014 26: 271 originally published online 29 April 2013
DOI: 10.1177/1079063213486936

The online version of this article can be found at:
http://sax.sagepub.com/content/26/3/271

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
ATSA
MAKING SOCIETY SAFER
Association for the Treatment of Sexual Abusers

Sexual Abuse: A Journal of
Research and Treatment
2014, Vol. 26(3) 271–290

DOI: 10.1177/1079063213486936
sax.sagepub.com
SAGE

Development of Vermont Assessment of Sex Offender Risk-2 (VASOR-2) Reoffense Risk Scale

Robert J. McGrath[1], Michael P. Lasher[1], Georgia F. Cumming[1], Calvin M. Langton[2,3], and Stephen E. Hoke[1]

Abstract

The present study aimed to revise the Vermont Assessment of Sex Offender Risk (VASOR) Reoffense Risk Scale, a commonly used sex offender risk assessment tool. The revised tool was named the VASOR-2. Among models tested to revise the scale, a logistic regression model showed the best balance between simplicity of use, goodness of fit, and internal validity (as tested with K-10 cross-validation), and maximized predictive accuracy. Predictive accuracy was tested using four meta-analytically combined data sets drawn from Canada and Vermont (N = 1,581). At 5-year fixed follow-up, the predictive accuracy for sexual recidivism for VASOR-2 (AUC = .74) was similar to the VASOR (AUC = .71). The findings show the VASOR-2 is well calibrated with observed recidivism rates for all but the highest risk sex offenders. The instrument showed good interrater reliability (ICC = .88). An advantage of the VASOR-2 is that it has fewer items and simpler scoring instructions than the VASOR. Norms are presented for a contemporary, nonselected, routine sample of Vermont sex offenders (n = 887).

Keywords

sex offender, risk assessment, Static-99R, VASOR-2

[1]Vermont Department of Corrections, Waterbury, VT, USA
[2]Ryerson University, Toronto
[3]University of Nottingham, UK

Corresponding Author:
Robert J. McGrath, Vermont Treatment Program for Sexual Abusers, 105 Happy Valley Road, Middlebury, Vermont 05753, USA.
Email: rmcgrath@sover.net

Introduction

Over the last two decades, use of structured risk instruments to predict sexual recidivism among identified sexual offenders has become routine professional practice (McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). In fact, it is commonly viewed as an essential component of effective sex offender management (Association for the Treatment of Sexual Abusers, 2005). Structured sex offender risk assessment approaches are consistently found to be more accurate than unstructured approaches (Hanson & Morton-Bourgon, 2009). Furthermore, interventions that take into account sex offenders' risk by providing more services to individuals judged more likely to reoffend are more often associated with reductions in recidivism than those that do not (Hanson, Bourgon, Helmus, & Hodgson, 2009; Lasher & McGrath, 2012; Lovins, Lowenkamp, & Latessa, 2009).

Structured risk instruments are composed of predetermined risk factors, specify how to combine factors into a total score, provide cutoff scores for risk levels, and often provide estimated recidivism rates (Dawes, Faust, & Meehl, 1989). Dawes and colleagues challenge instrument developers to periodically revise their scales to reflect advances in knowledge. Certainly, jurisdictions should regularly ensure that instrument norms are appropriate for local use.

Several current sex offender risk instruments are arguably improved versions of their earlier counterparts. Hanson, for example, developed the Rapid Risk Assessment for Sex Offense Recidivism static risk instrument in 1997, combined it with the Structured Anchored Clinical Judgment to produce the Static-99 (Hanson & Thornton, 2000), and revised it further with colleagues to create the Static-99R and Static-2002R (Helmus, Thornton, Hanson, & Babchishin, 2012). Hanson and colleagues also have developed what is now the third iteration of a dynamic risk assessment scheme (Hanson, Harris, Scott, & Helmus, 2007). Similarly, Minnesota Department of Corrections recently published the third version of their sex offender risk instrument (Minnesota Sex Offender Screening Tool-3; Duwe & Freske, 2012). At a psychiatric center in Canada, researchers modified the Violence Risk Appraisal Guide to better predict serious reoffending among sexual offenders (Sex Offender Risk Appraisal Guide; Quinsey, Harris, Rice, & Cormier, 2006).

In Vermont, modification of a dynamic risk scale, the Sex Offender Treatment Needs and Progress Scale, led to the development of the Sex Offender Treatment Intervention and Progress Scale (McGrath, Lasher, & Cumming, 2012). Earlier in Vermont, in response to a Department of Corrections request to develop a structured approach to guide sex offender sentencing and supervision decisions, McGrath and Hoke (1994/2001) developed the Vermont Assessment of Sex Offender Risk (VASOR). Although the VASOR has been used in Vermont since that time and increasingly in several U.S. jurisdictions in recent years (McGrath et al., 2010), the scale has undergone relatively little empirical examination.

The VASOR is a conceptual-actuarial instrument (Hanson & Morton-Bourgon, 2009). Literature reviews (McGrath, 1991, 1992) and clinical consensus among an expert panel (McGrath & Hoke, 1994/2001) guided item selection and weighting. The

VASOR is composed of a Reoffense Risk Scale, which has been the subject of four predictive validity studies and a Violence Scale (described in Measures section), which has not been subject to empirical examination.

The VASOR Reoffense Risk Scale predicted sexual recidivism (charges and child protective services sexual abuse substantiations) at 5-year fixed follow-up among sex offenders released from prison in Vermont (N = 172; AUC = .76; McGrath, Hoke, Livingston, & Cumming, 2001). Similarly, it predicted new sexual convictions at 3-year fixed follow-up among sex offenders released from prison in Canada (N = 176; AUC = .75; Langton, Barbaree, Harkins, Seto, & Peacock, 2002). Among two samples of Vermont offenders enrolled in community treatment, it did not predict new sexual charges at 5-year fixed follow-up in one sample (N = 208; AUC = .65; McGrath, Cumming, Hoke, & Bonn-Miller, 2007), but did predict new sexual charges at 3-year fixed follow-up in the other sample (N = 759; AUC = .73; McGrath, Lasher, & Cumming, 2011). The VASOR and VASOR-2 Reoffense Risk Scales are hereafter referred to, respectively as VASOR and VASOR-2.

Purpose of Study

Although VASOR predictive validity studies have been encouraging, we believed that several potential improvements in the scale and scoring manual were warranted. First, item weighting and construction of some VASOR items were needlessly complicated. Second, scale norms had not been examined in several years to determine if they required updating. Last, analyses of VASOR studies and results of a comprehensive meta-analysis (Hanson & Morton-Bourgon, 2005) caused concern that some VASOR items were unlikely risk predictors.

Accordingly, the primary aim of the present study was to construct a new version of the VASOR, named the VASOR-2, which would be more empirically grounded, easier to score, and accurately predict recidivism. Potential predictor variables for the new scale were VASOR and Static-99R items contained in four data sets (N = 1,581). We judged one of these data sets (n = 887) appropriate for developing local Vermont recidivism norms, and this was the secondary aim of the study.

Method

Measures

Static-99R. The Static-99R is a 10-item actuarial instrument designed to assess the recidivism risk of adult males who have been charged with or convicted of at least one sexual offense (Helmus, Thornton, et al., 2012). Items are identical to the Static-99 (Hanson & Thornton, 2000), with the exception of updated age weights. The 10 items pertain to sexual and nonsexual offense history, victim characteristics, and offender demographics. Total scores range from –3 to 12 points and are organized into four risk groups; low (–3 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6 to 12). A recent meta-analysis of 63 studies found a moderate relationship between Static-99

and sexual recidivism (Hanson & Morton-Bourgon, 2009). The authors of the Static-99 and Static-99R now recommend that evaluators use the revised version of the scale (Helmus, Thornton, et al., 2012).

VASOR. The VASOR is designed to assess sexual recidivism risk and offense severity of adult males who have been convicted of committing at least one sexual offense (McGrath & Hoke, 1994/2001). It is composed of two scales.

The 13-item Reoffense Risk Scale is composed of many of the same static risk factors found on the Static-99R as well as potentially changeable risk factors such as alcohol and drug use, residence and employment stability, and treatment amenability. Total Reoffense Risk Scale scores range from 0 to 125 and are organized into three risk groups; low (0 to 40), moderate (41 to 60), and high (61 to 125).

The six-item Violence Scale is composed of items measuring violence history and offense severity (e.g., force, sexual intrusiveness, and physical victim harm) and it is not a subject of the present study. In other research, we are examining the degree to which a revised version of the scale will predict the severity of sexual offenses committed by sexual recidivists.

Samples

Raw data were obtained for the four known VASOR data sets—three from Vermont and one from Canada—for which sufficient recidivism information was available to conduct logistic regression analyses for 5-year fixed follow-up periods. Raw data also included offender characteristics and Static-99R scores. Prior to merging, each data set was cleaned by correcting scoring errors and deleting cases that had an unacceptable number of missing items (see relevant scoring manuals) or had less than 5-year follow-up data. In addition, duplicate cases in Vermont data sets were removed. Consequently, sample sizes in the present study were sometimes smaller than those in the original studies. The total sample size was 1,581.

The 5-year follow-up period in the Vermont data sets was based on "calendar" time. Subtracting days participants were in prison during the 5-year fixed follow-up period, mean time-at-risk in the community was 53.8 months for the 1,015 nonsexual recidivist participants who comprised the Vermont 2007 and 2011 data sets. These data were not available for the Vermont 2001 data set. The 5-year follow-up period in the Canadian study used "street" time. Consequently, all participants in this data set were followed in the community for a full 1,825 days. Exceptions were that the follow-up period ended on the first date of a new offense for each type of recidivism (i.e., sexual or violent). In all data sets, only the first recidivism event for each recidivism type was counted as a new offense.

All sex offenders included in the data sets were male and age 18 or older at the time of placement in the community. They were convicted of at least one sexual offense against an identifiable child or nonconsenting adult victim (Category "A" sexual offense as defined in the Static-99 coding manual; Harris, Phenix, Hanson, & Thornton, 2003). Using this definition, individuals whose sex crimes were limited to offenses such as prostitution, statutory rape, or child pornography possession were excluded from the study.

Table 1. Characteristics of Studies.

	Vermont 2001	Canada 2002	Vermont 2007	Vermont 2011	Total
N	172	333	189	887	1,581
Age (*SD*)	38.0 (10.7)	39.0 (10.3)	35.3 (12.7)	34.2 (13.5)	35.7 (12.6)
Ethnicity White (%)	99.4	80.8	97.4	95.8	93.2
Offender type (%)					
Rapists	30.2	40.5	13.2	20.6	25.0
Child molesters	41.9	29.4	63.0	58.3	51.0
Incest	25.6	20.4	17.5	13.1	16.5
Noncontact	2.3	0.0	6.3	8.0	5.5
Mixed type	—	9.6	—	—	2.0
5-year recidivism rates					
Sexual	19.8	12.3	6.3	5.5	8.6
Violent	35.4	27.9	12.7	13.9	19.1
Recidivism criteria	Charge	Conviction	Charge	Charge	—
Risk scores *M* (*SD*)					
VASOR	43.9 (20.0)	48.5 (20.8)	29.2 (18.5)	25.3 (15.3)	32.7 (20.1)
VASOR-2	7.1 (3.7)	8.2 (3.5)	6.2 (3.7)	5.0 (3.1)	6.0 (3.6)
Static-99	2.9 (2.0)	3.4 (2.1)	2.6 (1.7)	2.7 (1.7)	2.8 (1.8)
Static-99R	2.8 (2.5)	3.3 (2.5)	2.4 (2.0)	2.5 (2.1)	2.7 (2.2)

Note: Vermont 2001 = McGrath, Hoke, Livingston, & Cumming, 2001; Canada 2002 = Langton, Barbaree, Harkins, Seto, & Peacock, 2002; Vermont 2007 = McGrath, Cumming, Hoke, & Bonn-Miller, 2007; Vermont 2011 = current study.

Table 1 shows the characteristics of the four samples. In terms of offender type, those who committed contact sexual offenses against extrafamilial children age 15 and younger were considered child molesters. Those who committed contact sexual offenses against victims age 16 or older were considered rapists. Incest offenders were individuals who sexually assaulted their biological children or stepchildren. Noncontact sex offenders committed offenses such as exhibitionism and voyeurism. Among the data sets, the primary difference in offender type definitions was that the Vermont studies categorized offenders by primary offense type, whereas the Canadian study had a mixed-type offender category. Samples are further described here.

Vermont 2001 (McGrath et al., 2001). This study followed 172 sex offenders who served a portion of a four or more year Vermont prison sentence between 1989 and 1993, and therefore, were eligible to enter the Vermont Department of Corrections prison sex offender treatment program. They were released to the community between 1989 and 1996. Of these men, 28.5% completed the treatment program, 24.4% entered but did not complete the program, and 47.1% refused treatment.

Canada 2002 (Langton et al., 2002). This study followed 468 sex offenders assessed at the Warkworth Sexual Behaviour Clinic between 1989 and 2000 while serving a

custodial sentence. The clinic was located in a medium security federal penitentiary in Ontario, Canada. All offenders were eligible to enter the Warkworth Sexual Behaviour Clinic sex offender treatment program. They were released to the community between 1990 and 2001. Of these men, 85.0% completed the treatment program, 8.1% entered but did not complete the program, and 6.8% refused treatment. This sample has been further described in Langton (2003).

Vermont 2007 (McGrath et al., 2007). This study followed 208 sex offenders who received community cognitive-behavioral treatment and correctional supervision in Vermont. One-half of the sample received periodic polygraph compliance exams and the other half did not. The two groups were exact pair-wise matched on Static-99 risk score, status as a completer of prison sex offender treatment, and year placed in the community. Men in the study were placed in the community between 1995 and 2001.

For current data analyses, men in this study ($n = 19$) who were also in the Vermont 2011 study, described hereafter, were counted only in the Vermont 2011 sample. None of these 19 men were charged for committing a sexual offense during follow-up periods in either the Vermont 2007 or 2011 data sets. Removing 19 individuals from the 2007 data set resulted in artificially increasing the sexual recidivism base rate among the remaining 189 individuals in this sample. We made this decision because we judged the Vermont 2011 sample appropriate for developing Vermont recidivism norms, so we did not want to remove any individuals from it.

Vermont 2011 (current study). This sample contained 97.6% of individuals ($n = 887$) among the exhaustive cohort of 909 sex offenders who were placed in the community in Vermont between 2001 and 2005. Lost to follow-up were the remaining 2.4% ($n = 22$) of individuals who otherwise met criteria for inclusion in the sample. Of the 887 individuals studied, 74.5% were on probation, 18.9% were on furlough, 2.4% were on parole, and 4.2% were released without follow-up correctional supervision after serving their maximum prison sentence. Of these individuals, 46.7% had served a prison sentence for their index sexual offense and 53.3% had not. Lastly, 17.5% received sex offender treatment in prison and 80.8% received at least some community sex offender treatment.

Outcome Measures

Recidivism data in the three Vermont samples were coded for each study participant for all new charges for sexual and violent (including sexual) offenses. The definition of a new sexual offense included a charge for a violation of supervision conditions if the incident could have been charged as a criminal sexual offense. In the Canadian sample, recidivism data were coded for new sexual and violent convictions only.

Scale Development

Scale development was informed by the work of several researchers (Blum, Kalai, & Langford, 1999; Harrell, Lee, & Mark, 1996, Kohavi, 1995; Vehtari & Lampinen, 2002; Worth & Cronin, 2003). It involved five major steps: (a) selecting scale items, (b)

developing potential item weighting models, (c) selecting a weighting model that showed the best balance between simplicity of use, goodness of fit, and maximized predictive accuracy, (d) determining relative risk categories and scores, and (e) comparing predictive accuracy of VASOR-2 with other scales. Development decisions aimed at making user transition from the VASOR to VASOR-2 as simple as possible.

Item Selection. Item selection began by examining the strength of the univariate predictive relationship between sexual recidivism and the 13 VASOR Reoffense Risk Scale items as measured by AUC (area under the curve of the receiver operating characteristic) analyses. The AUC statistic is a recommended index of predictive accuracy for relatively low base-rate phenomena such as sexual reoffending (Rice & Harris, 2005). It represents the probability that a randomly selected recidivist will have a higher score on a risk measure than will a randomly selected nonrecidivist. AUC values range from 0 to 1, with 0.5 representing chance-level prediction and 1 representing perfect prediction. Rice and Harris report minimum AUC values for small, medium, and large effects, which, respectively are .56, .64, and .71.

As expected (Hanson & Morton-Bourgon, 2005), two VASOR items—force used during index sex offense and amenability to treatment—did not show a statistically significant relationship ($p < .05$) to sexual recidivism and were dropped from the scale. The item "relationship to victim" did not quite reach statistical significance but was retained because victim relationship variables have consistently predicted sexual recidivism (Hanson & Bussière, 1998).

Three VASOR items—violations of community release conditions, address changes, and time employed—use trichotomous scales and these were transformed into dichotomous scales. Upon comparison, trichotomous and dichotomous versions of items predicted similarly, so we adopted the simpler scoring criteria for each item. Other similar modifications that simplified scoring rules are described elsewhere (McGrath, Hoke, & Lasher, 2013).

The Static-99R item "prior sentencing dates," a measure of general criminality, replaced the VASOR item "prior adult convictions," because it more accurately and simply predicted sexual recidivism. The single VASOR item "male victim and/or history of exhibitionism" was replaced with the Static-99R items "any male victim" and "any convictions for non-contact sex offenses." The later change resulted in increased item transparency over the original conflated item, as well as expanded the types of noncontact sexual offenses considered in the scale that have been statistically linked to sexual recidivism (Hanson & Bussière, 1998).

To account for the common finding that sexual recidivism declines with age (e.g., Helmus, Thornton, et al., 2012), an age item was added to the scale to decrease scale bias among older offenders. In the current data set, the distribution of sexual recidivists by age showed two trends of overall reductions in recidivism: first near age 35 and second near age 55.

Based on these analyses and criteria, we selected 12 items shown in Table 2 to comprise the VASOR-2 Reoffense Risk Scale. For economy of presentation, Table 2 also shows the univariate predictive relationship to sexual recidivism of each item in the final VASOR-2 model, whose development is described here.

Table 2. Predictive Accuracy of VASOR-2 Items for Sexual Recidivism at 5-Year Follow-up (N = 1,581).

	Reoffense Risk Scale		
		AUC	95% CI
1.	Age at community placement	.57*	[0.52, 0.61]
2.	Prior sex offense convictions	.66***	[0.61, 0.71]
3.	Prior sentencing dates[a]	.56*	[0.51, 0.61]
4.	Any violations of community release during past 5 years	.59***	[0.54, 0.65]
5.	Any convictions for noncontact sex offenses[a]	.63***	[0.58, 0.68]
6.	Any male victims[a]	.56*	[0.50, 0.61]
7.	Relationship to victims	.54	[0.48, 0.59]
8.	Offense-related sexual fixation	.68***	[0.63, 0.73]
9.	Substance abuse during past 5 years in community	.58**	[0.53, 0.63]
10.	Address changes during past year	.56*	[0.51, 0.61]
11.	Time employed or in school during past year in community	.56*	[0.51, 0.62]
12.	Sexual recidivist after treatment or treatment dropout	.60***	[0.55, 0.65]

Note: AUC = area under the curve; CI = confidence interval.
[a]Static-99R items.
*p < .05. **p < .01. ***p < .001.

Model Development. Following Worth and Cronin (2003), we developed six potential item weighting schemes for the new 12-item scale. Two models were unweighted, two models were weighted using logistic regression (using the odds ratios estimated by logistic regression), and two models were weighted using discriminant analysis (using linear discriminant function coefficients). Weighted item scores were rounded to the closest whole number and depending on the model and risk factor, ranged from 1 to 5.

Model Selection. From the six weighting scheme models developed, we aimed to select the one that showed the best balance between simplicity of use, best goodness of fit, and maximized predictive accuracy. Harrell et al. (1996) has summarized and compared three major model selection approaches. These are data-splitting, K-fold cross-validation, and bootstrapping. Data-splitting separates a database into two portions, one large and the other smaller. Modeling is conducted on the larger portion of the database and tested (i.e., cross-validated) on the smaller held out portion. The K-fold cross-validation method uses the entire database for model development. The database is split into k equal segments. Modeling is conducted on all but one segment, tested on the extracted segment, and repeated k times with the results averaged into a cross-validation model. Bootstrapping is similar to K-fold cross-validation except that it involves taking n (often 200 or more) samples with replacements from the database.

Several researchers opine that data-splitting is the weakest of the three approaches (e.g., Blum et al., 1999; Harrell et al., 1996) Among its weaknesses, it sacrifices data in developing and testing the model. It also introduces more variation in the model testing process than K-fold or bootstrapping techniques because the sample is split and

analyzed only once. K-fold cross-validation with a moderate number of data-splits is more efficient to conduct than a similar analysis with a high number of folds or bootstrapping and achieves similar test utility (Blum et al., 1999; Kohavi, 1995; Vehtari & Lampinen, 2002). Consequently, we chose to use K-10 cross-validation for our model validation tests and to select the one model that best fit the previously described criteria.

The K-10 cross-validation process involved splitting the data into 10 equal segments and conducting 10 sets of modeling 90% of the data and calculating model error and AUC values on the remaining 10%. The ten new model error values were averaged together ($Error_{cv}$) and compared to the error of the original, or apparent, model ($Error_{app}$), and a paired-samples *t*-test examined differences between the two. Model errors were calculated from each regression analysis' predicted category probability (PCP). The PCP is the probability that a VASOR-2 score correctly predicted whether a participant was a reoffender. Model error equals the average of one minus the PCP for each case. This calculation was conducted with each modeling analysis for the apparent and cross-validated models.

The ten new AUCs were averaged together into a cross-validated AUC value (AUC_{cv}) and compared to the apparent AUC value (AUC_{app}) and differences between the two were analyzed using Integrated Discrimination Improvement (IDI; Pencina, D'Agostino, D'Agostino, & Vasan, 2008). Better models showed small and insignificant differences between apparent and cross-validated model error and AUC values. Comparisons of AUC values for each potential VASOR-2 model and the VASOR utilized IDI. AUC comparisons used combined data sets with all cases (n = 1,581).

Risk Categories. To select risk categories (e.g., low, moderate-low, moderate-high, and high) and score ranges for each category, multiple configurations were examined on three criteria. First, we identified configurations of scores with the best goodness of fit using the Hosmer–Lemeshow χ^2 statistic, which examines the overall difference between observed and expected values (Hosmer & Lemeshow, 2000). Second, we tested for predictive accuracy as measured by the AUC value. Third, we examined the 95% confidence interval of each category to test for greatest exclusivity between groups by minimizing the overlap of confidence intervals. The goal of this process was to create risk categories based on estimated recidivism rates (generated by logistic regression analysis), which adequately fit the observed data, effectively predicted recidivism, and established categories that had meaningfully different recidivism rates.

Comparisons Across Samples. Finally, we compared the predictive accuracy (as measured by AUCs) of the newly constructed VASOR-2 with the original VASOR, as well as Static-99 and Static-99R. Furthermore, subset analyses compared AUCs both among offender types (rapists, child molesters, and noncontact offenders) and the four data sets used to construct the scale. A number of methods exist to compare predictive accuracy of AUCs among scales (Stephan, Wesseling, Schink, & Jung, 2003). Here, both the DeLong DeLong Clarke-Pearson Difference (DeLong, DeLong, & Clarke-Pearson, 1988) and IDI statistics were used for these comparisons.

The DeLong Difference statistic examines differences in AUC curves based on the individual curve variances and the curves' covariance. IDI is a newer rank order statistic that examines the probability of differences in discrimination slopes. IDI is arguably an improvement over earlier methods used to compare differences in discrimination of AUCs among scales (Woodman, Thompson, Kim, & Hakendorf, 2011). In the current study, IDI will be positive and significant when the VASOR-2, compared to another risk instrument, assigns a greater predicted probability of reoffense to reoffenders and a lesser predicted probability of reoffense to nonreoffenders.

To test the variability of AUC values across subsets of the database (i.e., the four studies used to construct the scale and offender types), we used Cochran's Q statistic (Hedges & Olkin, 1985) and the I^2 statistic (Higgins & Thompson, 2002). A significant Q statistic indicates that more variability across studies exists than would be expected by chance. When Q is significant, an I^2 statistic indicates the magnitude of the variability. I^2 scores around 25% can be interpreted as low, around 50% as moderate, and around 75% as high (Higgins, Thompson, Deeks, & Altman, 2003).

VASOR-2 Final Model Evaluation

Logistic regression equations examined estimated reoffense rates for individual scores and relative risk categories based on fitting scores to a logistic distribution. Logistic regression coefficients (β_0 and β_1) were calculated by aggregating the constant and score coefficients within the regression equation of each subset (Hasselblad & Hedges, 1995) to better account for variability of offense rates across samples (Hanson, Helmus, & Thornton, 2010). Q and I^2 statistics were also calculated for the constant and score coefficients to test for significant variability across the four meta-analytically combined databases. Following the recommendations of Helmus, Hanson, Thornton, Babchishin, and Harris (2012), β_0 coefficients for the VASOR-2 recentered at the median score (7) and one standard deviation below and above the median score (3, 11) were also meta-analytically combined to examine variability across samples at these different scores.

Finally, to assess model calibration (Altman, Vergouwe, Royston, & Moons, 2009), we used calibration plots with Lowess smoothed curves and Hosmer–Lemeshow χ^2 goodness of fit tests. A Lowess curve plots a fit line, which is compared to a reference line representing a perfect correlation of observed and estimated reoffense rates, otherwise referred to as perfect calibration of the assessment tool.

Vermont Norms

Of the four data sets included in this study, we used the Vermont 2011 data set ($n = 877$) to develop contemporary Vermont sexual recidivism norms. We judged the Vermont 2011 data set ideal for this purpose because it contained the near exhaustive cohort of sex offenders who were placed in the community in Vermont between 2001 and 2005. As such, the Vermont 2011 sample is considered an unselected (i.e., consecutive cases) routine correctional sample of sex offenders, which could be viewed as roughly

representative of all adjudicated sex offenders (Phenix, Helmus, & Hanson, 2012). This is opposed to samples preselect, for example, on treatment need, psychiatric disorder, or risk level.

Data Analyses

Analyses were conducted using SPSS 17.01, except for the DeLong Difference statistic, which was conducted using Analyse-It 2.20 for Microsoft Excel (Analyse-it Software, 2009), and meta-analytical combinations, Q statistics, and I^2 statistics, which were calculated manually in Microsoft Excel.

Results

As shown in Table 1, the sexual recidivism rate for the entire sample (N = 1,581) at fixed 5-year follow-up from the date of placement in the community was 8.6% and for any violent (including sexual) recidivism was 19.1%. For the nonselected routine Vermont 2011 sample (n = 887) on which Vermont norms were based, the sexual recidivism rate was 5.5% and the violent recidivism rate was 13.9%

Model Validation

Of six models tested to establish optimal VASOR-2 item weights, a logistic regression model showed the best balance among simplicity of use, goodness of fit, and maximized predictive accuracy. Cross-validation testing indicated no significant decrease from the apparent to cross-validated AUC values (AUC_{app} = .77, $p < .001$; AUC_{cv} = .76, $p < .01$; IDI = –.001, p = .48). Also, no significant increase in model error was found ($Error_{app}$ = .09; $Error_{cv}$ = .09; $t(1,580) = .36$, p = .72).

Table 3 shows that the VASOR-2 predicted sexual recidivism over the 5-year fixed follow-up period (AUC = .77, $p < .001$; 95% CI [0.73, 0.81]) as did the VASOR (AUC = .74, $p < .001$; 95% CI [0.70, 0.79]). Table 3 also highlights that for the total score, the DeLong Difference score was significant and the IDI was insignificant. This indicates that although the individual AUC analyses produced significantly different results, the IDI analyses found the difference between predictive abilities were not significantly different. This lack of statistical consistency indicates some ambiguity as to whether the VASOR-2 performs better than the VASOR. DeLong Difference scores and IDIs were both significant for the 2007 and 2011 subgroups, in the direction of better predictive accuracy for the VASOR-2 than VASOR.

Among the four meta-analytically combined VASOR-2 samples, the distribution of individual AUC values was no greater than would be expected by chance (Q = 1.60, p = .21). Predictive accuracy of VASOR-2 was similar when the samples were meta-analytically combined (AUC = .74, $p < .001$, 95% CI [0.69, 0.79]). However, as shown in Table 3, comparison of VASOR and VASOR-2 scores using meta-analytically combined samples found no significant difference in the predictive accuracy of the two tools.

Table 3. Comparative Predictive Accuracy of VASOR and VASOR-2 Reoffense Risk Scales for Sexual Recidivism by Study at 5-Year Follow-up.

	VASOR		VASOR-2			
Study	*AUC*	95% CI	*AUC*	95% CI	Difference	*IDI*
Vermont 2001 (*n* = 172)	.80***	[0.72, 0.88]	.77***	[0.68, 0.87]	–.023	–.024
Canada 2002 (*n* = 333)	.67***	[0.58, 0.75]	.70***	[0.62, 0.78]	.037	.017*
Vermont 2007 (*n* = 189)	.63	[0.47, 0.78]	.73**	[0.58, 0.87]	.071*	.038*
Vermont 2011 (*n* = 887)	.71***	[0.63, 0.79]	.76***	[0.68, 0.83]	.048**	.022*
Total (*N* = 1,581)	.74***	[0.70, 0.79]	.77***	[0.73, 0.81]	.023*	.007
Meta-analytical combination (*k* = 4)	.71***	[0.66, 0.76]	.74***	[0.69, 0.79]	.022	.003

Note: Vermont 2001 = McGrath, Hoke, Livingston, & Cumming, 2001; Canada 2002 = Langton, Barbaree, Harkins, Seto, & Peacock, 2002; Vermont 2007 = McGrath, Cumming, Hoke, & Bonn-Miller, 2007; Vermont 2011 = current study. AUC = Area under the curve. CI = confidence interval; Difference = DeLong DeLong Clarke-Pearson ROC Difference; IDI = integrated discrimination improvement.
*p < .05. **p < .01. ***p < .001.

Table 4. Meta-Analysis of Logistic Regression Coefficients for VASOR-2 for Sexual Recidivism (*k* = 4, *N* = 1,581).

	Coefficent	95% CI	*Q*	I^2 (%)
β_1	0.252	[0.20, 0.31]	2.33	—
β_0	–4.61	[–5.16, –4.05]	10.17*	70.5
β_0 (Centered 3)	–3.85	[–4.26, –3.43]	15.74**	80.9
β_0 (Centered 7)	–2.85	[–3.11, –2.60]	28.13***	89.3
β_0 (Centered 11)	–2.01	[–2.23, –1.80]	16.19**	81.5

Note: CI = confidence interval; Q = Cochran's Q statistic.
*p < .05. **p < .01. ***p < .001.

The VASOR-2 also predicted sexual recidivism across three subtypes of offenders. Based on meta-analytically combined AUCs of the four data sets used in the study, it predicted sexual recidivism among child offenders (n = 1,067; AUC = .74; p < .001; 95% CI [0.67, 0.81]), rapists (n = 395; AUC = .77; p < .001; 95% CI [0.67, 0.87]), and noncontact offenders (n = 87; AUC = .69; p = .02; 95% CI [0.50, 0.88]). The 32 mixed-type offenders in the Canada 2002 sample were not counted in offender type analyses. Variability of AUCs across the four data sets was no more than would be expected by chance among child molesters (Q = 2.19; p = .53) and noncontact offenders (Q = 5.31, p = .07), but it was greater than expected by chance among rapists (Q = 12.51, p = .006; I^2 = 76.0%).

Table 4 shows the meta-analytically combined β_0 and β_1 logistic coefficients. Also shown are β_0 coefficients centered at median VASOR-2 score of 7 and one standard deviation below and above the median (3, 11), as well as Q and I^2 statistics. The results show that estimated recidivism rates at various risk levels, as measured by VASOR-2

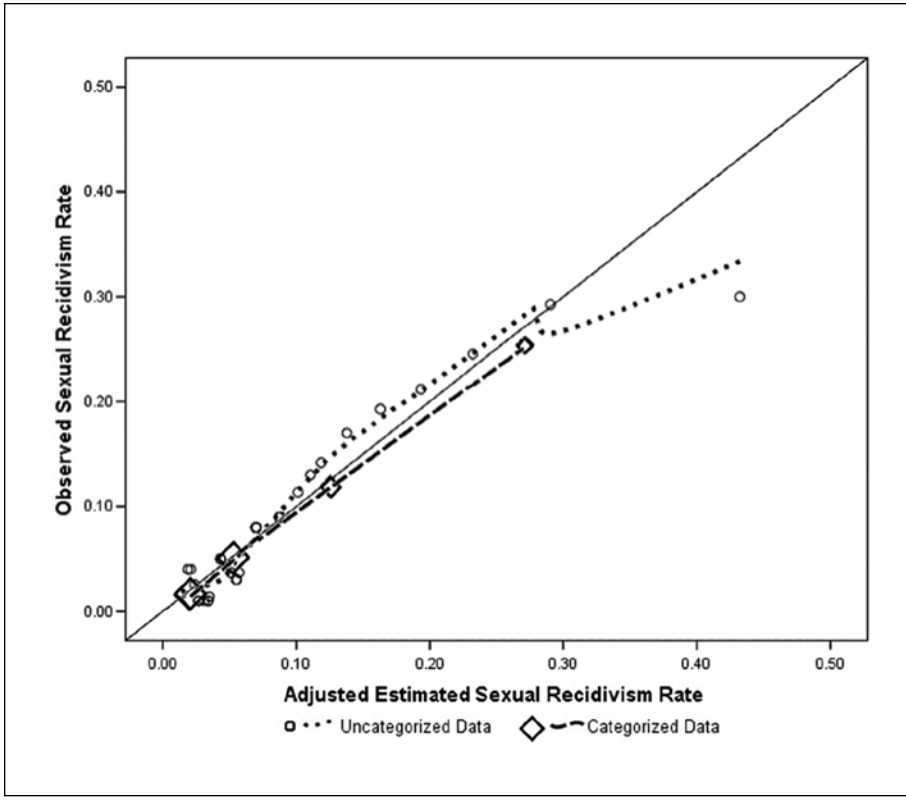

Figure 1. Calibration plot with Lowess smoothed curve for VASOR-2 adjusted estimated sexual recidivism rates and observed sexual recidivism rates (N = 1,581).
Note: Uncategorized data plots represent n = 51.

scores, demonstrated large and significant variability across the four data sets. For example, the meta-analytically combined estimated reoffense rate at the median score of 7 was 5.5%, but the estimated reoffense rates among the four individual databases at a score of 7 ranged from 4.3% to 10.7%.

Figure 1 shows the Lowess plot for uncategorized and categorized (low, moderate-low, moderate-high, and high) scores adjusted sexual recidivism rates. Uncategorized plots represent 31 equal groups of 51 cases and the categorized plots are sized relative to their sample distribution. As shown by the dotted line, estimated recidivism rates using uncategorized scores closely match the observed recidivism rates for all but the highest scoring cluster of participants, which represented 3.2% of the sample. These overall findings are consistent with the Hosmer–Lemeshow χ^2 goodness of fit test results (χ^2 (8, N = 1,581) = 2.71, p = .95). As shown by the dashed line, estimated recidivism rates using categorized scores are closely associated with observed rates, which is also consistent with the Hosmer–Lemeshow χ^2 goodness of fit test results (χ^2 (2, N = 1,581) = .31, p = .86).

Table 5. VASOR-2 Reoffense Risk Scale Categories and Observed and Estimated Sexual Recidivism Rates at 5-Year Follow-up for Nonselected Routine Vermont Sample (n = 887).

Score	Risk category	Percent of sample	Observed (recidivists/total N)	Estimated	95% CI
0-5	Low	41.0	1.4 (5/364)	1.7	[1.0, 2.8]
6-8	Moderate-low	35.4	4.5 (14/314)	4.2	[2.9, 6.0]
9-11	Moderate-high	15.9	11.3 (16/141)	10.2	[6.1, 16.4]
12-22	High	7.7	20.6 (14/68)	22.6	[15.0, 32.7]
Total		100.0	5.5 (49/887)	5.5	[2.1, 11.3]

Note: CI = confidence interval.
Area under the curve = 0.75, $p < .001$; CI [0.68, 0.82].

Vermont Norms

Table 5 shows the 5-year observed and estimated sexual recidivism rates for VASOR-2 categorized scores for the nonselected routine Vermont 2011 sample (n = 887) on which contemporary Vermont norms were based. The uncategorized VASOR-2 scores, ranging from 0 to 22, were grouped into categories and assigned a value from 1 to 4, respectively, representing low, moderate-low, moderate-high, and high-risk categories.

Incidental Findings

Incidental study findings were that the VASOR-2 predicted violent recidivism (AUC = .71, $p < .001$, 95% CI [0.68, 0.74]) slightly better than the VASOR (AUC = .69, $p < .001$, 95% CI [0.66, 0.72]; DeLong Difference = .02, p = .005; IDI = .07, $p <$ .001). Similar to sexual recidivism, a slight decrease in AUC for violent recidivism is found when meta-analytically combining subgroup AUCs (AUC = .69, $p < .001$). The Q statistic showed that the distribution of individual AUC values is no greater than chance (Q = .96, p = .81). Additional recidivism tables for sexual and violent reoffenses for individual and categorized VASOR and VASOR-2 scores are available from the authors.

Incidental study findings also indicated that the Static-99 and Static-99R predicted outcomes; both sexual recidivism (Static-99: AUC = .69, $p < .001$, 95% CI [0.64, 0.73]); (Static-99R: AUC = .69, $p < .001$, 95% CI [0.65, 0.74]) and violent recidivism (Static-99: AUC = .69, $p < .001$, 95% CI [0.65, 0.72]); (Static-99R: AUC = .70, $p <$ .001, 95% CI [0.67, 0.73]). In comparison, the VASOR-2 predicted sexual recidivism better than the Static-99 (DeLong Difference = .08, $p < .001$, IDI = .05, p = .001) and the Static-99R (DeLong Difference = .08, $p < .001$, IDI = .04, $p < .001$) and for violent recidivism better than Static-99R (DeLong Difference = .01, p = .48, IDI = .01, $p <$ .001). However, while statistically similar to the comparison with the Static-99R, the VASOR-2 did not significantly predict violent recidivism better than the Static-99 (DeLong Difference = .03, p = .05, IDI = .01, p = .34).

Lastly, the VASOR-2 showed good interrater reliability based on two independent ratings of 30 consecutive cases evaluated in Vermont's prison sex offender treatment program by pairs of six master's level mental health professionals. The total VASOR-2 score single measure interclass correlation coefficient was .88.

Discussion

The present study reports the results of efforts to revise the VASOR. The new scale, the VASOR-2, has fewer items and simpler scoring instructions than the VASOR, and consequently, is easier to use. The VASOR-2 predicted at least as accurately as the earlier version, and there was some evidence of increased accuracy.

The predictive accuracy of the VASOR-2 total score is similar to that of other sex offender risk-assessment instruments (Hanson & Morton-Bourgon, 2009). However, because the present results are based on scale development with a construction sample in which selection of items, weighting schemes, and cutoff scores were maximized for the sample, fairer comparisons of VASOR-2 to other instruments require replication studies. In the present study, we did not hold out a portion of the data set to conduct a separate cross-validation because we were convinced a K-fold cross-validation scale development model was superior to a data-splitting approach (Harrell et al., 1996). This meant that we used the entire data set in scale development.

The VASOR-2 also predicted sexual recidivism among three offender subtypes: rapists, child offenders, and noncontact sex offenders. Caution is indicated, however, when using the scale with noncontact offenders due to the low sample size and statistical power associated with these estimates.

Given the importance of developing local norms and our goal of establishing VASOR-2 Vermont norms, an important finding was that the instrument predicted sexual recidivism ($AUC = .76$) in the Vermont 2011 sample. This was an ideal normative sample for making local sex offender management decisions as it was composed of the near exhaustive cohort (98%) of sex offenders who were placed in Vermont communities between 2001 and 2005. As well, the sample was relatively large. The opportunity to collect data on this cohort was enhanced by the fact that the same authority, the Vermont Department of Corrections, manages prisons, jails, probation, and parole in the state.

Despite the fact that a focus of the present study was on developing a scale for use in Vermont, there is good reason to believe that the VASOR-2 will show ability to rank order risk for sexual recidivism risk in other jurisdictions as well. Reasons for optimism about the generalizability of the instrument in other jurisdictions are that the original VASOR and the VASOR-2 predicted sexual recidivism in a Canadian sample. In the present study, VASOR-2 predicted slightly better than Static-99 and Static-99R, and these instruments have consistently predicted elsewhere. Individual risk factors that comprise the VASOR-2 have been linked in multiple studies to sexual recidivism (Hanson & Bussière, 1998). Lastly, even risk instruments that are randomly generated from pools of established risk factors can predict as well as the original instruments from which the pooled risk factors were taken (Kroner, Mills, & Reddon, 2005).

Nevertheless, it is not uncommon for instruments to predict better on data on which they were developed than on new data (Bleecker et al., 2003). How well the VASOR-2 will perform in other samples is an empirical question that needs to be studied.

Overall, absolute recidivism estimates reported in the present study, as well as for the full sample and subsets of the full sample (which are available from the authors), should be used with caution. Whereas established sex offender risk instruments commonly show good predictive discrimination across studies, little research has examined the stability of absolute recidivism estimates across studies. Of the few studies to examine this issue, estimated recidivism rates of Static-99R and Static-2002R scores among 23 samples showed large and significant variability within each score (Helmus, Hanson, et al., 2012), and this is consistent with findings examining data sets in the present study shown in Table 4. Variability across studies can be influenced by study definitions of recidivism (e.g., arrests, charges, convictions, child protective services substantiations) as well as the characteristics of the offenders being studied (e.g., probationers, parolees, offenders prescreened for civil confinement) and local incident rates. Variability can also be influenced by local reporting, investigation, and prosecution practices, all of which may vary across time.

Absolute recidivism estimates are particularly important in certain contexts, such as civil commitment hearings where courts must determine whether an individual's risk to sexually reoffend reaches a specific threshold (e.g., 51% or greater lifetime estimated sexual recidivism rate). The small number of very high-risk sex offenders in the present study and the instrument's relatively poor calibration for this type of offender should dissuade practitioners from using the VASOR-2 for civil commitment evaluations at this time. Jurisdictions may also use absolute recidivism estimates to identify very low risk offenders (e.g., 1% or less 5-year estimated sexual recidivism rate) who might be deemed to not need sex offender treatment. Even if an absolute probability estimate is not required for decision making, it can be important for communicating risk, as nominal risk labels (e.g., low, moderate, or high) carry more meaning when anchored to specific recidivism rates (Hilton, Carter, Harris, & Sharpe, 2008).

Unfortunately, most jurisdictions do not have locally calibrated risk instruments. The calibration process requires considerable resources, technical expertise, and adequate follow-up time. Nevertheless, relative risk information can be useful for a variety of purposes. For example, if a jurisdiction wants to implement the risk principle by providing the most intensive services to those offenders at highest risk to reoffend; whether the highest risk group has a 20% or 35% 5-year sexual recidivism rate may not matter. The greatest attention, relative to other offenders, should still be placed on the higher risk group.

Vermont agencies and professionals now regularly use the relative risk categories and the associated recidivism rates shown in Table 5 to inform local decisions concerning sentencing, supervision, treatment, and community notification. Others may find the VASOR-2 useful in applied settings as well, particularly with routine correctional samples in jurisdictions with similar sexual recidivism base rates as those in the Vermont 2011 sample. Of course, development of local norms is always ideal.

Additionally, as Mossman (2006) cautions, even among jurisdictions with similar sexual recidivism base rates, absolute recidivism rates associated with a relative risk categories can still differ significantly.

As with other primarily static risk measures, the VASOR-2 does not provide a comprehensive assessment of risk. Contemporary sex offender research has increasingly focused on adding dynamic factors to risk assessment schemes to expand their utility. Dynamic measures within combined static and dynamic instruments give providers direction about which potentially changeable problems to target in treatment and supervision. Some evidence indicates that combined instruments predict recidivism better than single focused ones (e.g., Babchishin, Hanson, & Helmus, 2012; McGrath et al., 2012). Indeed, recent research found that the VASOR-2, when combined with a dynamic measure, the Sex Offender Treatment Intervention and Progress Scale (SOTIPS; McGrath et al., 2012), predicted better than when used alone (McGrath et al., 2013).

A further limitation of this study was the fact that only one of the four data sets used in the present study, the Canadian sample, was from outside Vermont. It is encouraging that all four data sets predicted sexual recidivism, but replication studies using more diverse populations are warranted. These include populations with more racial diversity than the present study, in which over 93.2% of participants were Caucasian, as well as more geographic diversity.

Evaluators and jurisdictions have several validated sex offender risk-assessment instruments from which to choose. These instruments, including the VASOR-2, focus primarily on predicting sexual recidivism. Although the VASOR-2 predicted recidivism in the present study, and is composed of risk factors that have consistently been linked to sexual recidivism in other studies, replication studies are needed. An aspect of the original VASOR, which was not examined here, concerns the classification and prediction of offense severity, and this is a focus of our ongoing research.

Acknowledgments

The authors thank Howard Barbaree for permission to use the Canadian data set and Karl Hanson for his helpful comments on this research. The views expressed are those of the authors and not necessarily those of the Vermont Department of Corrections.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by National Institute of Justice grant 2008-DD-BX-0013 to the Vermont Department of Corrections.

References

Altman, D., Vergouwe, Y., Royston, R., & Moons, K. G. M. (2009). Prognosis and prognostic research: Validating a prognostic model. *British Medical Journal*, *338*, 1432-1435.

Analyse-it Software. (2009). Analyse-it for Microsoft Excel (Version 2.20) [Computer software]. Leeds, UK.

Association for the Treatment of Sexual Abusers. (2005). *Practice standards and guidelines for the evaluation, treatment, and management of adult male sexual abusers*. Beaverton, OR: Author.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012). Even highly correlated measures can add incrementally to predicting recidivism among sex offenders. *Assessment*, *19*, 442-461.

Bleecker, S. E., Moll, H. A, Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. M. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, *56*, 826-832.

Blum, A., Kalai, A., & Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In S. Ben-David (Ed.), *Proceedings of the 12th Annual Conference on Computational Learning Theory* (pp. 203-208). New York, NY: ACM.

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

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.

Duwe, G., & Freske, P. J. (2012). Using logistic regression modeling to predict sexual recidivism: The Minnesota Sex Offedner Screen Tool-3 (MnSOST-3). *Sexual Abuse: A Journal of Research and Treatment*, *24*, 350-377.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders. *Criminal Justice and Behavior*, *36*, 865-891.

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. L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project*. Ottawa, Ontario, Canada: Public Safety Canada.

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.

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., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119-136.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*, 361-387.

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

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167-178.

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

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.

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.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558.

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

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.

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

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2* (pp. 1137-1143). San Francisco, CA: Morgan Kaufmann.

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*, 360-374.

Langton, C. M. (2003). Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy. *Dissertations Abstracts International, 64*, 1907B. (UMI No. NQ78052).

Langton, C. M., Barbaree, H. E., Harkins, L., Seto, M. C., & Peacock, E. J. (2002, October). *Evaluating the predictive validity of seven risk assessment instruments for sex offenders*. Paper presented at the 21st annual conference of the Association for the Treatment of Sexual Abusers (ATSA), Montreal, Quebec, Canada.

Lasher, M. P., & McGrath, R. J. (2012). The impact of community notification on sex offender reintegration: A quantitative review of the research literature. *International Journal of Offender Therapy and Comparative Criminology*, *56*, 6-28.

Lovins, R., Lowenkamp, C. T., & Latessa, E. J. (2009). Applying the risk principle to sex offenders: Can treatment make some sex offenders worse. *The Prison Journal*, *89*, 344-357.

McGrath, R. J. (1991). Sex-offender risk assessment and disposition planning: A review of empirical and clinical findings. *International Journal of Offender Therapy and Comparative Criminology*, *35*, 329-351.

McGrath, R. J. (1992). Five critical questions: Assessing sex offender risk. *Perspectives*, *16*, 6-9.

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

McGrath, R. J., Cumming, G. F., Hoke, S. E., & Bonn-Miller, M. O. (2007). Outcomes in a community sex offender treatment program: A comparison between polygraphed and matched non-polygraphed offenders. *Sexual Abuse: A Journal of Research and Treatment*, *19*, 381-393.

McGrath, R. J., & Hoke, S. E. (2001). *Vermont Assessment of Sex Offender Risk Manual*. Middlebury, VT: Author. (Original work published 1994)

McGrath, R. J., Hoke, S. E., & Lasher, M. P. (2013). *Vermont Assessment of Sex Offender Risk-2 (VASOR-2) Manual*. Middlebury, VT: Author.

McGrath, R. J., Hoke, S. E., Livingston, J. A., & Cumming, G. (2001, November). *The Vermont Assessment of Sex Offender Risk (VASOR): An initial reliability and validity study*. Paper presented at the 20th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, San Antonio, TX.

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2011). *A model of static and dynamic sex offender risk assessment* (Document No. 236217). Washington, DC: United States Department of Justice.

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse: A Journal of Research and Treatment*, *24*, 431-458.

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

Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B., Jr., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, *27*, 157-172.

Phenix, A., Helmus, L., & Hanson, R. K. (2012, July 26). *Static-99R and Static-2002R evaluators workbook*. Retrieved from http://www.static99.org/pdfdocs/Static-99RandStatic-2002R_EvaluatorsWorkbook2012-07-26.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.

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.

Stephan, C., Wesseling, S., Schink, T., & Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clinical Chemistry*, *49*, 433-439.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, *14*, 2439-2468.

Woodman, R. J., Thompson, C. H., Kim, S. W., & Hakendorf, P. (2011, September). *Comparison of the C-statistic with new model discriminators in the prediction of long versus short hospital stay*. Paper presented at the 2011 Australia and New Zealand Stata Users Group Meeting, Fremantle, Western Australia.

Worth, A. P., & Cronin, M. T. D. (2003). The use of discriminant analysis, logistic regression, and classification tree analysis in the development of classification models for human health effect. *Journal of Molecular Structure*, *622*, 97-111.