Open Access Journal of Forensic Psychology

http://www.forensicpsychologyunbound.ws/ -- 2012.4: 1-23

Problems with the Static-99R Prediction Estimates and Confidence Intervals

Theodore S. Donaldson, Ph.D., Independent Practice, Florence, OR, USA

Brian R. Abbott, Ph.D., Independent Practice, San Jose, CA, USA
Email: brian@dr-abbott.net
(Corresponding Author)

Christine Michie, B.Sc., Department of Psychology, Glasgow Caledonian University, Scotland, UK

Abstract

Estimating the probability of future sex offenses by released sex offenders has become a major issue in sexually violent predator civil commitment proceedings, and several risk prediction instruments are used for this purpose. The accuracy of the predictions is the central issue in determining the utility of these instruments. The accuracy of point estimates is addressed through analysis of the confidence intervals surrounding the point estimates. Herein we address problems with these confidence intervals due to ignoring underlying mathematical assumptions that are necessary for their validity. These problems are amplified when used in sexually violent predator evaluations because forensic experts must have a high degree of confidence in the results. We explore these problems in the case of the popular and most-used risk instrument, the Static-99R, and demonstrate that it is essentially impossible to establish a risk estimate with a high level of confidence. As a less important issue, we point out errors in the risk estimates for the Static-99R High Risk Need Group, such that estimates over 50% are rarely possible. We maintain that even the most optimistic expectation for risk instrument improvements will not produce risk predictions with the level of risk and confidence required by experts in SVP legal proceedings.

Keywords: Static-99R, sexually violent predators, sexual recidivism, risk prediction, Bayes Theorem

Introduction

The purpose of this paper is to further explore the underlying assumptions in calculating confidence intervals around risk estimates related to the Static-99R. In a previous article, Donaldson and Abbott (2011) elaborated the errors in using group confidence intervals for risk predictions about an individual. Herein we point out substantial errors in the Static-99R group confidence intervals, such that neither individual nor group confidence intervals allow predictions with a high degree of confidence in sexually violent predator civil confinement proceedings. As a lesser issue, we point out errors in

some of the Static-99R risk estimates primarily due to assuming that, if one cannot reject the null hypothesis, there are no meaningful differences.

Estimating violent sex-offense recidivism has become a central feature in sexually violent predator commitment hearings where the risk criterion is set at a statutorily defined threshold (e.g., more likely than not). Of the several actuarial instruments developed for the purpose of estimating that risk, the Static-99R is the most widely researched (Hanson & Morton-Bourgon, 2009) and currently the most often used in sexual offender risk assessments (Boccacinni, Murrie, Caperton, & Hawes, 2009), especially in the context of sexually violent predator civil confinement actions (Jackson, Travia, & Schneider, 2008). The Static-99R is a revision of the original Static-99 that is modified to accommodate and better account for the influence of two factors: changes in recidivism base rates that have occurred since the development of the original Static-99 and increasing age of the offender (Helmus, 2009). Currently, the Static-99R contains ten items, eight of which are scored 0 or 1, an age factor, with scores of -3 to +1, and the item "prior sex offenses," which can be scored 0 to 3. Possible total scores are from -3 to 12.

Scoring procedures and tables relating risk scores to recidivism estimates are provided in the Static-99 Coding Rules (Harris, Phenix, Hanson, & Thornton, 2003), the Static-99R Revised Age Weights (Helmus, Thornton, Hanson, & Babchishin, 2011), and the *Evaluator's Workbook for Static-99R* (Phenix, Helmus, & Hanson, 2012). The workbook provides logistic regression risk prediction tables for four reference groups, each comprised of data from between six and fifteen studies conducted by researchers around the world. Generally, most studies represent offenders released from prison with approximately 20% of the offenders having received community dispositions. The *Evaluator's Workbook for Static-99R* (Phenix et al., 2012) describes the four reference groups: Routine Corrections (8 studies), Preselected Treatment Need (PSTN) and Preselected High-Risk Need (PHRN), with each reference group consisting of six studies, and Non-routine Corrections (15 studies). The studies comprising the Non-routine Corrections reference group are the same studies constituting the PSTN and PHRN reference groups with three other studies added.

The Static-99R developers propose criteria by which to select a reference group to compare to the individual being assessed; however, this method has not been empirically validated nor established as reliable, and has not been subjected to peer review and publication. Nonetheless, there is a strong tendency among sexually violent predator evaluators to select the PHRN reference group to compare to the individuals being assessed. In states that require an estimate of more likely than not, only the highest risk scores from the PHRN reference group have risk estimates that even approach sufficiently high estimates to meet that criterion.

During research on their paper addressing individual confidence intervals for score-wise risk estimates from the Static-99R, Donaldson and Abbott (2011) noticed two significant issues related to the logistic regression estimates from the Static-99R experience

tables. First, the logistic regression estimates for the PHRN reference group consistently exceeded the observed values at scores greater than 2, a result that would not be possible if the regression were based only on the PHRN data. The regression line minimizes the deviations about the line, and cannot be consistently above or below the observed data. Second, the group confidence intervals reported by the Static-99R developers are based on accepting the premise that all members at the same score within a reference group are equivalent in terms of their risk potential. It is doubtful that this premise is true and, in fact, is undoubtedly seriously wrong. These two issues suggest that the absolute risk estimates from the PHRN reference group at a score greater than 2 may be inflated and the score-wise group confidence intervals may be underestimated in all reference groups. If true, these findings have important implications for the application of the risk data from the Static-99R.

We examine the assumptions and methods used to establish the logistic regression estimates for the PHRN reference group and describe the source of error in the score-wise logistic risk estimates. The second part of this article explains serious problems with the reported confidence intervals in the *Evaluator's Workbook for Static-99R* (Phenix et al., 2012). We demonstrate the problems and errors associated with group confidence intervals using the PHRN reference group. However, those problems and errors affect the other three Static-99R reference groups as well.

Errors of Risk Estimation

The *Evaluator's Workbook for Static-99R* (Phenix et al., 2012) provides tables indicating the associated risk at each score on the Static-99R for each of the four reference groups. The frequency data from which the estimates were derived can be obtained from experience tables published at: http://www.static99.org/pdfdocs/detailed_recid_ tables_ static99r_2009-11-15.pdf. We will refer to the 5-year estimates for the PHRN reference group but the same error trend is observed in the risk estimates at the 10-year follow-up period.

The experience table for the PHRN reference group shows the observed score-wise risk estimates for the fixed follow-up period and for the logistic regression estimates, which are based on the fixed follow-up data. Table 1 reproduces some of that information: column 1 indicates the Static-99R scores, column 2 shows the total number of sex offenders with each score, column 3 reflects the number of recidivists for each score, and column 4 documents the observed recidivism rate based on the numbers in columns 2 and 3. Column 5 shows the predicted recidivism rates as calculated by the Static-99R developers. Note that the predicted recidivism rates in column 5 for scores 3 through 10 are consistently higher than the observed recidivism rates in column 4. These observed and predicted estimates are plotted in Figure 1. It is readily apparent in Figure 1 that the PHRN logistic regression estimates (Helmus, 2009) are consistently higher than the observed recidivism rates. It is impossible to obtain such difference between observed and regression estimates if the regression analysis is correct and based on the PHRN data. The estimated regression line is the one that minimizes the

(squared) deviations between the data points and the estimated points on the regression line, which necessarily requires that some of the data points are above the line, and some are below the line over the entire range of estimates.

Table 1: Static-99R Recidivism Information for The High Risk High Needs Group

1 Static-99R Scores	2 Score-Wise N	3 N of Score-Wise Recidivists	4 Observed Recidivism Rate (%)	5 Predicted Recidivism Rates Per Evaluator's Handbook	6 Predicted Recidivism Rates Using B_1 from High Risk Group Data
-3	1	0	0.0	-	4.1
-2	6	0	0.0	-	5.1
-1	32	2	6.3	5.4	6.4
0	44	3	6.8	7.2	8.0
1	91	5	5.5	9.4	9.8
2	88	16	18.2	12.2	12.1
3	161	23	14.3	15.8	14.8
4	220	42	19.1	20.1	18.1
5	230	46	20.0	25.2	21.8
6	189	53	28.0	31.2	26.1
7	131	42	32.1	37.9	30.9
8	72	24	33.3	45.0	36.1
9	35	14	40.4	52.4	41.7
10	11	6	54.5	59.7	47.5
11	2	0	0.0	-	53.4
12	0	0	0.0	-	59.0

In developing the Static-99R Helmus (2009) meta-analyzed 22 studies using a random-effects model in which the 22 studies are assumed to come from the same population. Two of the 22 studies were eliminated when calculating the B_1 coefficient[1] including one sample from the Routine Corrections (RC) reference group (Hanson, Harris, Scott, & Helmus, 2007). One of the remaining 20 studies (Hill, Habermann, Klusmann, Berner, & Briken, 2008) is not included in the samples comprising the RC, PSTN, and PHRN reference groups. Helmus' (2009) meta-analysis indicated that the regression slope

[1] The B_1 coefficient determines the slope of the regression line. As it relates to the Static99R data, B_1 represents the amount that the estimated recidivism rate changes for each 1 point increase in the Static-99R score.

coefficient B_1 did not differ significantly among 20 studies. Therefore, the predicted risk estimates for each reference group were determined by the average B_1 coefficient from all 20 studies. Table 2, column 1 lists each of the 20 studies with the corresponding reference group listed in column 2. The third column in Table 2 reflects the sample size for each study at the five-year follow up. The fourth column shows the base rate of each study. The fifth column shows the B_1 coefficient for each study along with the corresponding standard error in column 6.

Figure 1: Static-99R High Risk High Needs Observed and Predicted Sexual Recidivism Rates

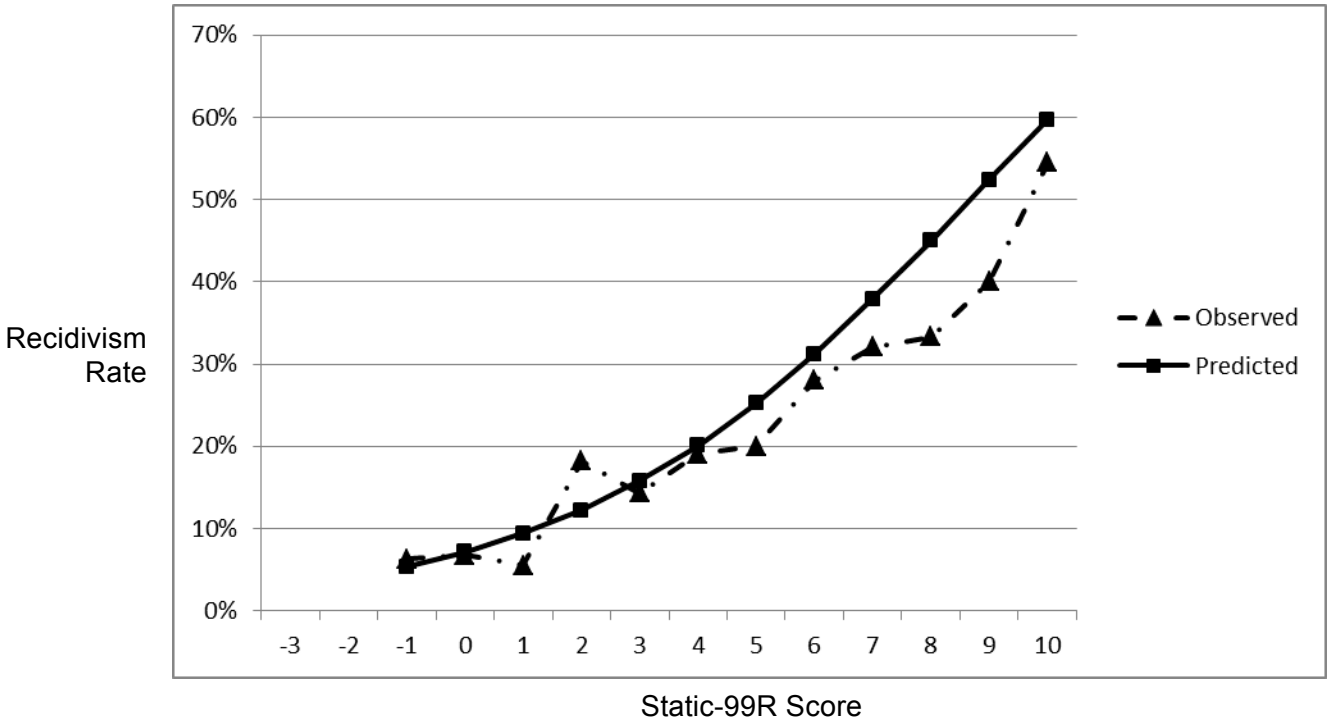


Table 2, column 5, demonstrates that the B_1 coefficient varies across the 20 studies from 0.035, for the Wilson data (Wilson, Cortoni, & Vermani, 2007a; Wilson, Picheca, & Prinzo, 2007b) to 1.013, for the Eher data (Eher, Rettenberger, Schilling, & Pfafflin, 2009). Helmus (2009) noted that these differences, while appearing quite large, were not significantly different. We do not think that testing the null hypothesis is the most appropriate test when one wants to show that meaningful differences do not exist between selected groups of studies.[2] However, after defining the three reference

[2] The issue here is one of statistical power or type 2 error, an issue largely ignored in most social science research (Cohen, 1988). Using a large value for the type 1 error (e.g. 30%) is a frequently used approach to avoid addressing the issue of power (i.e. the probability of detecting a meaningful difference of a specified size on a selected variable). In this approach one assumes that, if the null hypothesis cannot be rejected at a large value for the type 1 error, it is safe to act as if the null hypothesis is true. However, when the error variance is large it is nearly impossible to obtain differences among the selected variable

groups, Helmus (2009) failed to test for possible meaningful differences in B_1 among the reference groups. The differences revealed in columns 4 and 5 from Table 1 between observed and predicted risk estimates for the PHRN reference group suggest the problem could stem from using a B_1 value that is too large. We explored this by investigating the values of B_1 in the various reference groups.

Table 2: Slope Parameters (B_1), Standard Error, and 5-Year Base Rates For Static-99R Reference Groups from Appendix E Helmus (2009)

1 Study	2 Reference Group	3 N	4 Base Rate	5 B_1	6 SE
Bartosh	Routine Corrections	90	.133	0.137	0.114
Bigras	Routine Corrections	207	.092	0.354	0.111
Boer	Routine Corrections	299	.037	0.467	0.140
Craissati	Routine Corrections	200	.075	0.341	0.112
Eher	Routine Corrections	151	.020	1.013	0.401
Epperson	Routine Corrections	150	.107	0.347	0.107
Langstrom	Routine Corrections	1278	.054	0.308	0.050
Allan	Preselected Treatment	298	.117	0.439	0.087
Brouillette-Alaire	Preselected Treatment	199	.146	0.332	0.092
Harkins & Beech	Preselected Treatment	197	.096	0.350	0.097
Johansen	Preselected Treatment	272	.059	0.208	0.111
Swinburne-Romine	Preselected Treatment	569	.084	0.249	0.066
Ternowski	Preselected Treatment	247	.065	0.296	0.100
Bengtson	High Risk/High Needs	310	.197	0.208	0.065
Bonta & Yasine	High Risk/High Needs	101	.188	0.347	0.144
Haag	High Risk/High Needs	198	.197	0.299	0.090
Knight & Thornton	High Risk/High Needs	433	.247	0.240	0.052
Nicholaichuk	High Risk/High Needs	168	.227	0.362	0.099
Wilson	High Risk/High Needs	103	.117	0.035	0.136
Hill	Not included*	73	.110	.377	.232

* Study used to calculate average slope parameter but not included in any reference group

Table 3 shows the average B_1 (both weighted and unweighted) for the studies comprising each of the three reference groups. The PHRN reference group has a weighted B_1 coefficient of 0.249, compared to the Routine Corrections Group of 0.376. This indicates that the increase in risk as scores rise is less for the PHRN reference group than it is for the Routine Corrections reference group. According to Helmus

of interest large enough to reject the null hypothesis. The analysis has low power in that the probability of finding a “meaningful” difference is small.

(2009; Appendix H, page 187), the average B_1 coefficient of 0.290 was used to estimate the risk scores for each reference group. However, as noted in Table 3, B_1 for the PHRN reference group was 0.249 (weighted), indicating that the regression line is not as steep as it is for all 20 studies combined. Thus, the risk estimates for the PRHN reference group are excessive. Also note in Table 3 that the weighted average of B_1 for the RC reference group is 0.376 and for the PSTN reference group, it is 0.301. The average value of 0.290, based on all 20 samples, approximates (by accident) the slope parameter for the PSTN reference group but is not representative of the other Static-99R reference groups in that it inflates risk estimates in the PHRN reference group and underestimates the probability of risk in the RC reference group.

Table 3: Summary of Information Regarding Slope Parameters, Sample Sizes, and 5-Year Base Rates by Static-99R Reference Groups

Static-99R Reference Group	N	B_1 Unweighted Average	B_1 Weighted Average	Base Rate
Routine Corrections	2375	0.424	0.376	.060
Preselected Treatment	1783	0.312	0.301	.091
High Risk/High Needs	1313	0.249	0.249	.210

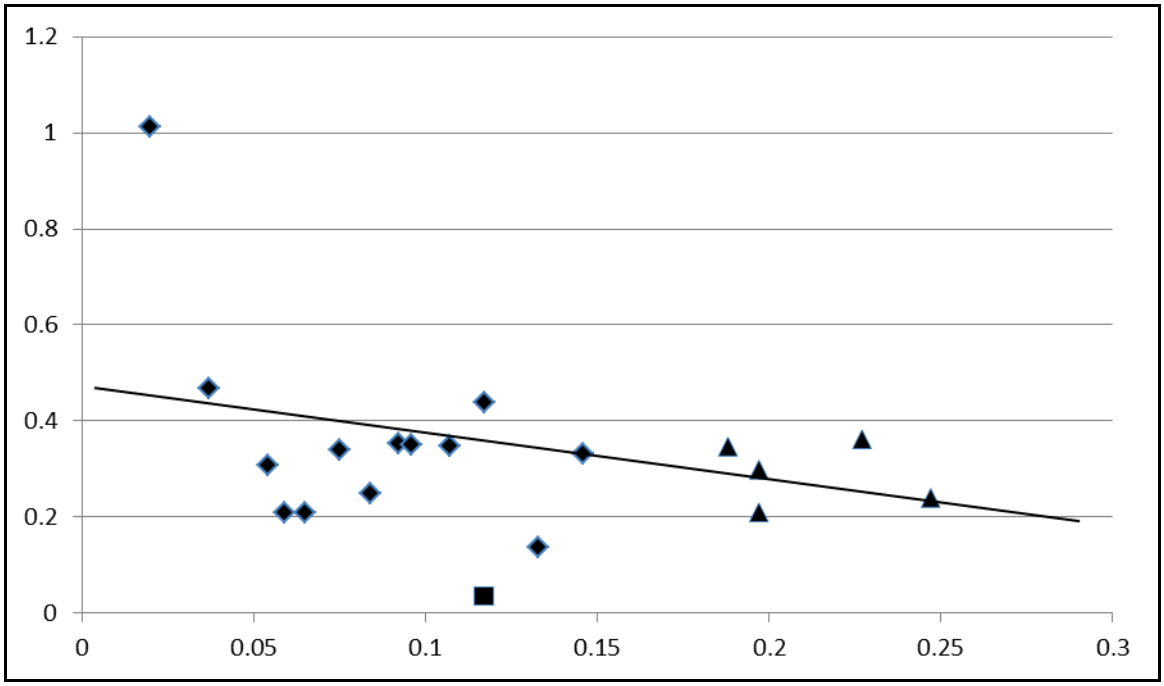
We explored the relationship between B_1 and the base rate further by regressing B_1 on the base rate for the 19 studies that make up the three reference groups. The results indicated a correlation between B_1 and the base rate of -0.387.[3] The pair-wise data points and the regression line are shown in Figure 2. The results indicate a fairly substantial reduction in the size of B_1 as the base rate increases. This result further supports the conclusion that the increase in risk, as scores increase, is less for the PHRN reference group than it is for the two other groups both of which have lower base rates. It appears that the three reference groups were defined—purposely or inadvertently—by their respective mean base rates and base rate for each sample comprising the reference groups is correlated with B_1, as seen in Figure 2. The negative correlation between B_1 and base rate caused the observed over-estimates of risk in the PHRN reference group.

Column 6 of Table 1 shows the risk prediction estimates based on the logistic regression analysis by using the slope parameter calculated only on the studies comprising the PHRN reference group. These estimates appear much closer to the observed recidivism rates in that they are sometimes higher and sometimes lower than the observed values, which is what one expects from regression estimates. These

[3] A central t-test of the significance of r = 0.387, with 17 degrees of freedom yields t = 2.04, which is just short of the p = .05 level for rejecting the null hypothesis.

recalculated regressions estimates are consistently lower than the logistic regression estimates provided by Helmus (2009), as seen in Table 1, column 5. The B_1 weight based on the PHRN group data accounts for the differences in the risk estimates we calculated and those cited by Helmus (2009), as well as showing that our estimates more closely approximate the observed data in the PHRN reference group.

Figure 2: B_1 Regressed on Base Rate



B_0 =0.468; B_1 = -1.131; y = 0.468 – 1.136x; r = -0.387

As noted in Table 1, the Helmus logistic regression estimates for scores of 9 and 10 exceed the 50% level, whereas our recalculated estimates only exceed 50% for scores of 11 and 12. There were no data in the Static-99R PHRN experience table for scores 11 and 12 and the estimates we provide at these scores in Table 1, column 6 result from the extrapolation of the regression. Our analysis indicates that it would be exceedingly rare to find a group risk estimate that exceeds 50% even when using the PHRN reference group for comparison. However, in most cases it is poor practice to compare an offender to this group, as the characteristics of the offenders in the reference groups are essentially unknown.

Finally, we point out one curious fact about the makeup of the PHRN group. As noted in Table 2, the combined Wilson et al. (2007a & 2007b) studies result in a considerably smaller base rate and slope coefficient, B_1, than the other samples in that group. This is

most obvious in Figure 2, where the position of the Wilson data, relative to the other groups in the regression, is noted by the square symbol. The other five studies of the PHRN group are the five data points to the extreme right of the Wilson data, as illustrated by the triangle symbols in Figure 2. Obviously, the Wilson data are outliers, and one wonders why it was included in this group in the first place. Even when calculating the risk estimates and confidence intervals (CIs), both with and without the Wilson data, the resulting risk estimates are very similar, although the CIs are somewhat smaller without the Wilson data. We have not addressed this issue, primarily because we are focusing on the Static-99R, as it exists and as it is used in practice today. There are also serious questions regarding the composition of the reference groups, especially the PHRN group, including the fact that the largest sample (Knight & Thornton, 2007) is very old and was based on data collected from 1959 to 1984. In our opinion, it is basically inappropriate to apply norms from this era to contemporary groups of sexual offenders.

Confidence Intervals

Forensic experts who address the likely-to-recidivate criterion when conducting sexually violent predator evaluations generally report not only risk estimates but also a confidence interval (CI) about those estimates. Authors of risk prediction instruments such as the Static-99R provide tables showing the relationship between instrument scores and estimates of risk. These tables also show a 95% CI for each risk estimate. These CIs are based on the score-wise risk estimates and are usually interpreted as the range of uncertainty about the risk estimate.

A related problem is that group CIs do not indicate the degree of confidence related to making predictions about individuals. Recently, Cooke and Michie (2010) reported on the CIs for an individual taken from the population on which the instrument was developed and showed that those CIs are much wider than CIs for the group means. Donaldson and Abbott (2011) provided evidence that it is nearly impossible to ever be 90% confident that an offender has a risk over a few percentage points, no matter which of the reference groups from the Static-99R (Phenix et al., 2012) were used.

In written reports and testimony, SVP evaluators do not address the differences between group and individual CIs. They continue to erroneously apply group CIs as reported in risk instrument manuals to the individuals they evaluate. As addressed below, we show that there are problems so significant with group CIs that it is impossible to have a high degree of confidence in the risk estimates in that the lower bound of a reasonable CI rarely indicates a substantial risk, and never over 50%.

A serious deficiency in the reported CIs for actuarial instruments such as the Static-99R and Static-2002R (Hanson, Phenix, & Helmus, 2009), and MnSOST (Epperson, et al., 2003), is that the calculations of the group CIs are based upon the unstated, and obviously wrong, assumption that every individual with the same risk score has the

same risk. This is a complex issue involving statistical analysis of proportions. There are at least two models that underlie the analysis of proportions.

In model one, a sample is made up of elements identified by one of two characteristics. In the classic example of proportions, the sample elements are marbles that differ only with respect to their color. In the simplest case, there are two colors. The proportion of marbles with one color in a sample of marbles depends only on the relative number of marbles with each color in the population from which the sample is taken. Applying model one to risk assessment, the proportion would represent the number of recidivists relative to the number of non-recidivists in a sample. In order to calculate the sample CI the offenders in the sample must differ only in terms of being recidivists versus nonrecidivists. In order to use model one, it must be assumed that the probability is either 1 or 0 for everyone in the sample, and the proportions of 1's (or 0's) in a sample estimates the probability of 1's in the population.

In model two, each member of the sample has some potential probability for recidivism that is a characteristic of that individual, and the observed proportion of recidivists in the sample is the average of the potential probabilities for recidivism. In model two, each individual in the sample has some probability for recidivism—based on risk relevant factors unique to the individual. Calculation of CIs is based on the proportion of recidivists in the sample but the proportion does not take into account potential differences in the unique risk associated with each individual's probability to recidivate. The resulting CIs are thus necessarily based on the average of the potential risk across all individuals in the sample and do not take into account individual *differences* in risk. That is, the variance estimate for the CIs are based on the proportion (average) and do not take into account the variance among the individuals who are pooled in order to calculate the average.

Given an observed proportion for a sample, both models lead to the same probability estimate. However, the CIs calculated by the formula for model one is not accurate for model two. Most practical applications of proportion statistics likely involve a mixture of the two models, but this also leads to construction of inaccurate CIs.

The MnSOST-R, for example, combines several scores yielded by the instrument into a bin, then provides an estimate of risk for individuals in a particular bin along with the CI around that prediction estimate. Because the individuals in the bin do not each have the same score, they obviously do not have the same probabilities of recidivism and the CIs based on the *average* risk across multiple scores underestimate the size of the CI. Even when risk estimates are provided for each score, as in the Static-99R, the underlying assumption that every individual with the same score has the same risk is most improbable. The variability in risk among individuals with the same risk score partially results from the static risk factors on actuarial instruments not fully accounting for risk variables unique to the individual (Hanson & Harris, 2001) so that all sex offenders are not equally likely to reoffend (Hanson, Morton, & Harris, 2003). The variability of risk among individuals violates the assumption of individuals having equal

risk at equal scores that underpins probabilistic prediction models such as actuarial risk measures. Woodworth and Kadane (2004) emphasize this critical issue by explaining how the individual being assessed must be considered conditionally exchangeable with members of the actuarial class at the individual's obtained score. In other words, there should be no way to identify a subset of members in the actuarial group that recidivate at a different rate and no known factors external to the actuarial measure can be used to distinguish members of the class.

The risk of reoffense does not only depend on static risk factors included in the risk instruments and those not included, or even dynamic factors that are currently in vogue (Thornton, Hanson, & Helmus, 2010), but also is always partially determined by the environments into which offenders are released and the situations that develop for offenders in those environments. Situational variables yet to be encountered by the offenders undoubtedly contribute to risk but are not identified in any risk instrument and, by their very nature, cannot be. These factors, however, vary widely among offenders with the same risk instrument score.

It is axiomatic that all sex offenders with the same score on a risk-prediction instrument do not have the same probability of reoffense. Despite this important fact, little or no information is available regarding the sources of variability in risk among offenders with identical scores, and there has been almost no discussion in the literature regarding the effects of variable probabilities of risk on reported CIs.

Table 4: Static-99R PHRN Confidence Intervals for Scores 6 – 9 Compared to Recalculated Confidence Intervals

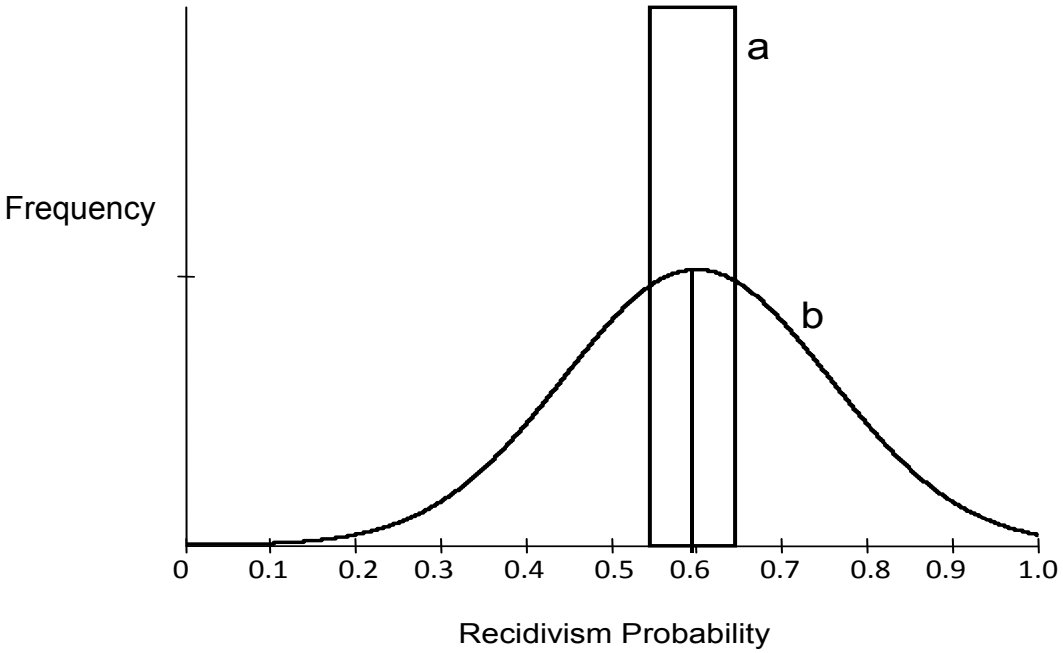
1 Static-99R Score	2 95% CI Helmus Average Across Studies	3 Recalculated 95% CI	4 Recalculated CI At 80% Level	5 Recalculated CI At 40% Level
6	27.2 - 35.5	23 – 30	24 - 28	25 - 28
7	33.1 – 42.9	27 – 36	28 - 34	28 - 34
8	34.3 – 50.8	30 – 42	32 - 40	32 - 40
9	45.9 – 58.8	34 – 50	37 - 47	37 - 47

Clinicians can use the CI to determine a level of confidence of an individual having a risk of at least some value, say P_c, where P_c would be the lower bound of the CI. While the CIs for sexual recidivism actuarials are expressed at the 95% level for risk estimates, an expert using one of the instruments in court proceedings would probably be more interested in a lower CI. At the 95% CI, an expert would be 97.5% certain that the individual had a risk of at least as big as the lower bound of the 95% CI. Using the lower limit of the 80% CI, an expert could express a 90% certainty of an individual at a

specific score on the actuarial measure. Similarly, a 70% level of confidence can be achieved by considering the lower bound of the 40% CI. Table 4 reflects the three confidence interval levels at the Static-99R scores 6-9.

We will use a hypothetical situation with which to demonstrate the effect on the CI when each person in a group of individuals with the same risk score has a different probability of recidivism. Suppose that a group of offenders with a particular score has a risk of 0.6 for recidivism. Assume that everyone in the group has the same risk, which is shown in Figure 3 with the bar marked "a"; that is, there is no individual variability in risk. Thus, the entire group is represented by the bar graph, indicating a risk of 0.6. In this case (depending on sample size), the 80% CI might extend from 0.52 to 0.67. One would then be 90% confident that individuals in this group had a risk of greater than or equal to 0.52. However, suppose that the probability of individual risk is actually distributed as shown by the curve marked "b." The group still has an average of 0.6, but now suppose the 80% CI extends from 0.40 to 0.80; in which case, one is 90% confident this group has a risk of greater than or equal to .40.

Figure 3: Hypothetical Distribution of Recidivism Probabilities for a Group of Individuals at the Same Score



The above example is for illustrative purposes and is not based on any actuarial data regarding the distribution of probabilities within a score. It demonstrates, however, how the CI is seriously affected by variability of individual probabilities. In this example, one goes from being confident that the group has a risk of greater than 50%, to not being

confident that it is greater than 50%. This difference may seriously impede a forensic expert's attempt to address civil commitment requirements.

Although we know essentially nothing about the distribution of individual probabilities for a given risk instrument score, we can explore the variability within a Static-99R reference group. For two reasons, we focus on this variability as it relates to the PHRN Reference Group. One, in our experience, it is the reference group to which SVP evaluators most often compare the individual being assessed. Two, the PHRN reference group has the highest base rate of any reference group, and therefore, the highest prediction estimates. Therefore, if it can be shown that the expert cannot have sufficient confidence in the estimates for the PHRN Reference Group then certainly it follows that the other reference groups would have similar problems as they were derived using the same methodology.

Analysis

The PHRN reference group is comprised of 6 studies (shown in Table 5). The sex-offense base rates for the six studies in this group ranged between 11.5% and 24.7%. Since risk estimates depend on the base rate (Donaldson & Wollert, 2008; Mossman,

Table 5: Risk Estimates for Static-99R
PHRN Reference Group Scores 6-9

	Static-99R Score-wise Risk Estimates				
Sample	**6**	**7**	**8**	**9**	**Base Rate**
Wilson	16.7	18.6	19.7	25.0	11.5
Bonta & Yessine	25.6	29.1	30.5	37.1	18.8
Bengtson	26.8	30.0	31.6	38.3	19.7
Haag	26.8	30.0	31.6	38.3	19.7
Nicholaichuk	30.5	34.4	35.3	42.6	22.6
Knight & Thornton	32.9	36.7	38.3	45.5	24.7
Group Estimate					
Helmus predicted	31.2	37.9	45.0	52.4	21.0
Bayesian computation	26.0	31.0	36.1	41.7	21.0

2006), different base rates necessarily produce different score-wise risk estimates in the various studies. Helmus (2009) having found no significant differences between the base rates and regression slope parameter, pooled the samples from the 6 studies comprising the PHRN reference group. As pointed out previously, the problems of pooling based only on the failure to reject the null hypothesis obscure meaningful differences among studies and raise two important questions related to CIs. One, do the differing base rates of the studies comprising the PHRN reference group effecthe size of the Static-99R prediction estimates? Two, assuming those differences in the

prediction estimates are meaningful, what effect do differing base rates have in terms of the conclusions about the size of the CI?

In order to provide an estimate of the variability within a group, we calculated the probability of recidivism for scores 6, 7, 8 and 9 for each of the studies within the PHRN reference group. The probability of risk associated with each score for the various study samples was calculated using Bayes Theorem, which depends only on the base rate and the score likelihoods. The score likelihoods are derived by taking the number of recidivists and nonrecidivists at each score and dividing these values by the total number of recidivists and nonrecidivists, respectively. The likelihoods for the combined studies comprising the PHRN reference group are shown in Table 6 for each score. The score-wise likelihood ratio is computed by dividing the likelihood of recidivists by the likelihood of nonrecidivists.

Table 6: Score-wise Likelihoods for Static-99R PHRN Reference Group

Static-99R Score	Likelihood of Recidivists	Likelihood of Nonrecidivists	Likelihood Ratio
-3	0.0	.0009	0.0
-2	0.0	.0057	0.0
-1	.0072	.0289	.250
0	.0108	.0395	.274
1	.0181	.0829	.218
2	.0579	.0694	.835
3	.0833	.1330	.626
4	.1521	.1716	.887
5	.1666	.1774	.939
6	.1920	.1311	1.46
7	.1521	.0858	1.77
8	.0869	.0462	1.87
9	.0507	.0202	2.50
10	.0217	.0048	4.51
11	0.0	.0019	0.0

We then computed the Bayesian estimates using the group likelihoods associated with scores of 6 to 9 for each sample. As seen in Table 6, at a score of 6, the likelihood for recidivists is 0.192, and the likelihood for nonrecidivists at that score is 0.131. For a score of 7, the likelihood for recidivists is 0.152, and for a nonrecidivist is 0.086, and so forth, for scores 8 and 9. Using Bayes's Theorem, one can then calculate the group probabilities associated with scores of 6 to 9 from the likelihoods and the base rates for the sample. Using Bayes Theorem, the probability of recidivism, given a score, S_j, is:

$$P(R \mid S_j) = \frac{P(S_j \mid R)P(R)}{P(S_j \mid R)P(R) + P(S_j \mid R^{'})P(R^{'})} \quad (1)$$

Where:

$P(S_j \mid R)$ is the likelihood of score S_j for recidivists R,
$P(S_j \mid R^{'})$ is the likelihood of score S_j for nonrecidivists, $R^{'}$,
$P(R)$ is the sample base rate, and
$P(R^{'})$ is 1- $P(R)$.

The score-wise risk probabilities computed from Bayes Theorem associated with each of the six studies comprising the PHRN reference group are shown in Table 5. For the Static-99R scores of 6, 7, 8, and 9, the results reveal a wide range of estimates. For example, at the five-year follow up, the risk estimates for a score of 6 are 16.7% in the Wilson et al. (2007a & 2007b) data to 32.9% in the Knight and Thornton (2007) sample. For a score of 7, the risk estimates bounded by these two studies ranges from 18.6% to 36.7%. The bottom panel in Table 5 shows the group-based risk estimates obtained by Helmus (2009) and those we calculated using Bayes Theorem. The range of risk estimates, as noted in Table 5, for a score of 6 is 16% to 33%. It is apparent from these calculations that the average predicted risk estimate for the PHRN reference group (31.2%) is not a good estimate of the risk for the various studies that made up that score group. The same trend is seen at all other scores in Table 5.

The 95% CIs for the Static-99R scores of 6, 7, 8, and 9, as reported in the *Static-99R Evaluator's Workbook* (Phenix et al., 2012), are shown in Table 4, column 2. However, as reflected in Table 1, the calculations of the risk estimates in the Evaluator's Workbook (Phenix et al.) are not the same as those calculated by us. Consequently, the CIs will also be different between our data and that of Helmus (2009). These differences are evident by comparing the 95% CIs in columns 2 and 3 from Table 4. The third column in Table 4 provides the 95% CIs based on our calculations using the slope parameter for the 6 studies comprising the PHRN reference group. Columns 4 and 5 in Table 4 reflect the 80% and 40% CIs for scores of 6 through 9 based on our logistic regression estimates for the PHRN reference group.

As illustrated in Table 4, the CIs based on the average risk prediction estimates for the various studies comprising PHRN reference group do not begin to capture the variability existing within that group. The CI in Table 4 for our data for a score of 6 is 23% to 30%, resulting in the risk estimates from two studies (values of 16.7% and 32.9%, as seen in Table 5) falling outside the 95% CI. We observed similar results for scores of 7, 8, and 9. The CI based on the group mean is an underestimate, that is, narrower than it should be, and the true CI would have to include the lower-bound estimate of the CI for the 16% estimate to the upper limit of the CI for the sample with the 33% estimate.

Under this circumstance, the true 95% CI is considerably wider than that based on the mean value of the group estimate.

The above analysis is based on the assumption that the score-wise likelihoods based on the group estimate are applied to each of the studies comprising the PHRN reference group. This assumption is probably not true, but has little effect on the calculations and certainly not on the conceptual issues. The likelihoods do not depend on base rate (Donaldson & Wollert, 2008), so that whatever effect the differences in likelihoods would have on the predictions is independent of the effect of the base rates. This analysis shows that the CIs for a group that is made up of multiple studies with different base rates yield CIs that are substantially too narrow, thus overestimating the accuracy and applicability of the estimate, even to the group itself.

Discussion

We have demonstrated two problems in using the Static-99R to make predictions about sex-offense recidivism. The first has to do with errors in the score-wise risk estimates reported in the *Static-99R Evaluator's Workbook* (Phenix et al., 2012), due to having used the slope regression coefficient B_1 that was established on regression analysis of all studies combined and then applying that same coefficient to selected groups of studies (Helmus, 2009). This error led to inflation of predicted recidivism rates beyond the observed rate at the Static-99R scores above 2 in the PHRN reference group. The second issue relates to the violation of the assumptions of equal probability for subjects with the same score, which results in CIs that are too narrow.

Helmus (2009) used the random-effects model, but we found no evidence that the various sex-offender studies relied upon by Helmus were, in fact, randomly selected. To the contrary, every available sample was used. There is much uncertainty about the features of the populations that were sampled. In using the random-effects model, the only appropriate development of the score-wise risk estimates would necessarily include all studies.

The most important issue, however, is that selecting specific groups from those available precludes there being a random-effects model. The use of the random-effects model implies that there is no interest in the effects of any *particular* sample or group of studies. Focusing on a particular group results in a fixed-effects model and inferences from the results of the fixed-effects model are made about a specific group. Because the actual slope coefficient for the PHRN reference group was lower than the *average* slope coefficient based on all studies, our analysis has shown that the risk estimates for the PHRN reference group are overestimates. As a result, the prediction estimates for the High Risk Group are higher than the actual observed score-wise recidivism rates for all scores greater than 2. Using corrected predicted estimates based on the slope parameter for the PHRN group data shows that a risk estimate of more likely than not, i.e., greater than 50%, is almost never obtained when based on the actuarial information.

Our second finding shows that the CIs reported for the Static-99R for each reference group are greatly underestimated by some unknown amount. The use of score-wise CIs is based on the untested and certainly untrue assumption that every member with the same risk score has the same potential of recidivism. If that assumption cannot be supported, the CIs are underestimates. We investigated the effects of violating the equal-probability assumption by considering a non-equal distribution of probabilities for members of a group with the same risk score and determined that deviations from equal probabilities increase the size of the CI. In fact, almost any distribution of group score-wise risk probabilities makes it impossible to have a high degree of confidence in group estimates.

Research has demonstrated that the CIs for a group cannot be applied to a single individual (Cooke & Michie, 2010), and applying this to the Static-99R (Donaldson & Abbott, 2011) indicates that one can never have high confidence in any risk estimate (greater than a few percentage points). The basic reason for this is that the correlations between the Static-99R scores and sex-offense recidivism are low and account for too little of the total variance. In this article, we have shown that even group CIs do not allow for high confidence in the estimate, and the lower bound of the CI rarely, if ever, exceeds 50%, even at the 70% level of confidence.

There have been a number of attempts to increase the magnitude of the prediction estimates of the Static-99R. These attempts have included using dynamic or other psychological risk factors in addition to the static actuarial items (Hanson et al., 2007; Mann, Hanson, & Thornton, 2010); combining the risk estimates with the Structured Risk Assessment-Forensic Version (Thornton, 2010); combining several risk instruments (Doren, 2010); and considering undetected sexual offenses. The research into increasing prediction estimates, though, is in its infancy and results are inconsistent. However, given the low correlation between risk instrument scores and recidivism, and a substantial history of psychological research that rarely shows high correlations between psychological constructs and outcomes (Cohen & Cohen, 1983), even the most optimistic outlook about the potential improvements in risk assessments could not produce large risk estimates with a high degree of confidence (Donaldson & Abbott, 2011). The research on risk estimation may have credible scientific value, but it is difficult to conceive of situations in which risk estimation can achieve results about an individual of any substantial value, and rarely, if ever, above 50%. This observation is consistent with previously published research in that the highest recidivism rate for the full-sample MATS-1 was 38% (Wollert, Cramer, Waggoner, Skelton, & Vess, 2010; Table 6, n = 9,305) and it was 26% for the Static-99R (Helmus et al., 2011; Appendix, n = 8,106). Therefore, it would impossible to ever have the lower bound of the group or individual confidence interval exceeding 50%.

Customary forensic practice by SVP evaluators is to select an actuarial reference group considered as best representing the population from which the individual being assessed comes and reporting the score-wise risk estimate and corresponding 95% CI.

Donaldson and Abbott (2011) have shown the significant degree of conceptual and actual error that exists when using the Static-99R group-based risk estimates for application to individuals. In our study, errors were identified in group-based predicted-risk estimates and CIs yielded by the Static-99R. These same conceptual errors are present in other sexual recidivism actuarial instruments such as the Static-2002R (Hanson et al., 2009), and MnSOST-R (Epperson et al., 2003). Under the circumstances described in this paper, a high degree of confidence in the score-wise actuarial risk estimates cannot be achieved.

We should point out that the findings presented in this paper, while confirmatory, are not new. Our findings are consistent with the conclusions of Woodworth and Kadane (2004; p. 240):

> It is, therefore, necessary to set a high standard for expert testimony about the probability of recidivism. We have reviewed existing instruments for estimating the probability of recidivism, and find little evidence that they meet such standards.
>
> It is not possible now, and may not be possible in the foreseeable future, to predict violent sexual recidivism with sufficient sensitivity and specificity to protect the public without putting a high burden on nonviolent offenders.

Our assessment of the utility of actuarial risk-assessment instruments (ARAIs) in the legal arena will be difficult to accept for those who develop or rely on these tools in practical situations. ARAIs are valuable for developing models that explain sexual recidivism among groups of sexual offenders. Yet, like most (perhaps all) explanatory models in the social sciences, the correlation between the prediction model and outcome is too low to result in highly accurate predictions about groups and, especially, individuals.

Attempts are ongoing to improve the accuracy of ARAIs primarily by considering risk factors beyond those captured by the ARAI (Abbott, 2011). Given the nature of the technical problems associated with such efforts (Donaldson & Abbott, 2011; Mokros, Stadtland, Osterheider, & Nedopil, 2010; Vrieze & Grove, 2010; Woodworth & Kadane, 2004), it is unrealistic to expect that meaningful improvements will result in risk-assessment methods that produce sufficient accuracy in the absolute risk estimates as applied to groups or individuals. As a result, clinicians will continue to lack the necessary confidence in the risk estimate to conclude in SVP legal proceedings that an individual meets the statutorily defined risk threshold. So, what are clinicians to do?

We propose an alternative to the customary practice of risk assessment in SVP cases. An option is to make risk estimation contingent on the presence in the offender of the "mental abnormality" itself; wherein proof of the existence of the mental abnormality establishes the requisite risk. If there is valid and sufficient mental-health information (First & Halon, 2008; Wollert, 2007) to conclude that the offender currently suffers a

condition that predisposes him to engage in sexual violence, and he currently has serious difficulty controlling the sexually violent behavior associated with that condition, then the offender is presumed to meet the dangerousness criterion as defined by the SVP statute. The alternative we propose is consistent with the United States Supreme Court decision in Kansas v. Crane (2002) upholding the constitutionality of SVP laws. The Crane court conceptualized the issue of dangerousness as a function of the conjunction between serious difficulty controlling behavior and the nature and severity of the mental abnormality so that an individual with this condition could be distinguished from the dangerous but typical recidivist convicted in the ordinary criminal case.

Acknowledgements

The authors thank Robert Halon, Ph.D., for his considerable contributions and assistance in preparing this article.

Received March 25, 2012; accepted June 8, 2012.

References

Abbott, B. R. (2011). Throwing the baby out with the bath water: Is it time for clinical judgment to supplement actuarial risk assessment? *Journal of the American Academy for Law and Psychiatry*, *39*, 222-230.

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(4)*, 278-314.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Psychology Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34 (2010)*: 259-274.

Donaldson, T. S., & Abbott, B. R. (2011). Prediction in the individual case: An explanation and application of its use with the static-99r in sexually violent predator risk assessments. *American Journal of Forensic Psychology*, *29*, 1-31.

Donaldson, T., & Wollert, R. (2008). A mathematical proof and example that Bayes's Theorem is fundamental to actuarial estimates of sexual recidivism risk. *Sexual Abuse: A Journal of Research and Treatment*, *20(2)*, 206-217.

Doren, D. M. (2010). Single instrument vs. multidimensional model risk assessment—resolution and the risk profile. In A. Schlank (ed.)*, The Sexual Predator: Legal Issues, Assessment, Treatment* (pp. 10-2-10-18), N.J.: Civic Research Institute.

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

Epperson, D. L., Kaul, J. D., Huot, S. J., Hesselton, D., Alexander, W., & Goldman, R. (2003). *Minnesota sex offender screening tool-revised (Mn-SOST-R): Development, performance, and recommended risk level cut scores*. Unpublished paper.

First, M. B., & Halon, R. L. (2008). Use of DSM paraphilia diagnoses in sexually violent predator commitment cases. *Journal of the American Academy of Psychiatry and the Law*, *36*, 443-454.

Hanson, R. K., & Harris, A. J. R. (2001). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *13 (2)*, 105-122.

Hanson, R. K., Harris, A. J. R., Scott, T. L., & Helmus, L. (2007). *Assessing risk of sexual offenders on community supervision: The dynamic supervision project*. Retrieved March 23, 2008, from http://www.publicsafety.gc.ca/res/cor/rep/_fl/crp2007-05-en.pdf

Hanson, R. K., Phenix, A., & Helmus, L. (2009). Static-99 and static-2002: *How to interpret and report scores in light of recent research*. Workshop presented at the ATSA 28th Annual Research and Treatment Conference on September 30, 2009, Dallas: TX.

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)*, 1-21.

Hanson, R. K., Morton, K. E., & Harris, A. J. R. (2003). Sexual offender recidivism risk: What we know and what we need to know. *Annals of the New York Academy of Sciences*, *989*, 154-166.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 Coding Rules Revised- 2003*. Retrieved January 10, 2010, from http://www.static99.org/pdfdocs/static-99-coding-rules_e.pdf

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples*. Retrieved January 10, 2010, from http://www.static99.org/pdfdocs/helmus2009-09static-99normsmathesis.pdf

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2011). Assessing the Risk of older sex offenders: Developing the Static-99R and Static-2002R. Retrieved June 2, 2011, from http://www.publicsafety.gc.ca/res/cor/rep/2011-01-aroso-eng.aspx.

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.

Jackson, R., Travia, T., & Schneider, J. (2008). Annual survey of sex offender civil commitment programs. *Sexual Offender Civil Commitment Network Research Committee.* Retrieved October 25, 2009, from: www.soccpn.org/ATSA%202008%20presentations/2008%20Annual%20Survey%20of%20Programs.pdf

Kansas v. Crane, 534 U.S. 407 (2002).

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

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.

Mokros, A., Stadtland, C., Osterheider, M., & Nedopil, N. (2010). Assessment of risk for violent recidivism through multivariate Bayesian classification. *Psychology, Public Policy, and Law*, *16(4)*, 418-450.

Mossman, D. (2006). Another look at interpreting risk categories. *Sexual abuse: A Journal of Research and Treatment*, *18(1)*, 41-63.

Phenix, A., Helmus, L., & Hanson, R. K. (2012). *Static-99R evaluator's workbook*. Retrieved February 2, 2012, from: http://www.static99.org/pdfdocs/ Static-99RandStatic-2002REvaluatorsWorkbook2012-01-09.pdf

Thornton, D. (2010). *Using the forensic version of the SRA in sex offender risk assessment*. Presented at workshop sponsored by Central Coast Clinical and Forensic Psychology Services, Atascadero, CA (December 2 & 3, 2010).

Thornton, D., Hanson, R. K., & Helmus, L. (2010). *Moving beyond the standard model for actuarial risk assessment for sexual offenders*. Retrieved August 3, 2010, from http://ccoso.org/newsletter.php

Vrieze, S. L., & Grove, W. M. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological Assessment*, *22(2)*, 382-395.

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

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

Woodworth, G. G., & Kadane, J.B. (2004). Expert testimony supporting post-sentence civil incarceration of violent sexual offenders. *Law, Probability and Risk, 3,* 221-241.

Wollert, R. (2007). Poor diagnostic reliability, the null-bayes logic model, and their implications for sexually violent predator evaluations. *Psychology, Public Policy, and Law*, *13(3)*, 167-203.

Wollert, R., Cramer, E., Waggoner, J., Skelton, A., & Vess, J. (2010). Recent research (n = 9,305) underscores the importance of using age-stratified actuarial tables in sex offender risk assessments, *Sex Abuse: Journal of Research and Treatment, 22*: 471-490.