The Utility of Assessing "External Risk Factors" When Selecting Static-99R Reference Groups

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

Abstract

The Static-99 has been one of the most widely used sexual recidivism actuarial instruments. It has been nearly four years since the revised instrument, the Static-99R, has been released for use. Peer-reviewed literature has been published regarding the basis for changing the scoring system for the age-at-release item, the utility of relative risk data, and variability of sexual recidivism rates across samples. Thus far, the peer-reviewed literature about the Static-99R has not adequately addressed the reliability and validity of the system to select among four possible actuarial samples (reference groups) from which to obtain score-wise observed and predicted sexual recidivism rates to apply to the individual being assessed. Rather, users have been relying upon the Static-99R developers to obtain this information through a website and workshops. This article provides a critical analysis of the reliability and validity of using the level of density of risk factors external to the Static-99R to select a single reference group among three options and discusses its implications in clinical and forensic practice. The use of alternate methods to select Static-99R reference groups is explored.

Keywords: Static-99R; risk assessment; sex offenders; prediction; recidivism

---

The initial release of the Static-99R (Hanson & Thornton, 2000) provided a single sexual recidivism experience table for users to consult to obtain score-wise risk estimates to compare to the individual being assessed. The developers of the Static-99 ("developers") released multiple experience tables in 2008, ostensibly in an effort to better account for the significant variability in base rates observed across different samples of sexual offenders (Harris, Helmus, Hanson, & Thornton, 2008). The developers instructed users to select one among three experience tables ("reference groups") but this procedure was later modified to report a range of risk bounded by the sexual recidivism rates from two reference groups (Abbott, 2009). These procedures were abandoned one year later with the release of the Static-99R (Hanson, Phenix, & Helmus, 2009) and users were advised to follow a specific method by which to select a single reference group among four options (Phenix, Helmus, & Hanson, 2009).

In 2009, Phenix et al. (2009) instructed Static-99R users to select reference groups employing a method referred to as *cohort matching* where clinicians consider broad-brushed descriptions of the four Static-99R reference groups, devised by the developers without empirical validation, and attempt to match the characteristics from a single reference group that is most similar to the group representing the individual being assessed. The reliability and validity of the cohort-matching process has not been established and Wollert (2010) reported how classification error (the probability of

Abbott, B. R. (2013). The utility of assessing "external risk factors" when selecting Static-99R reference groups. *Open Access Journal of Forensic Psychology, 5*, 89-118.

selecting an erroneous reference group to compare to the individual being assessed) reduces the accuracy of the observed sexual recidivism rates. In a July 2012 revision of the Static-99R Evaluators' Handbook, Phenix, Helmus, & Hanson (2012) state the preferred method for selecting reference groups is to consider the density of risk factors external to the Static-99R. Studies of this procedure have not been conducted to test its reliability and validity. Phenix et al. (2012) constructed the reference-group-selection procedures based on two major untested hypotheses.

One, the developers assume, through post-hoc logical inference, that risk factors external to the Static-99R account for the differences in sexual recidivism base rates between reference groups (Phenix et al., 2012; Thornton, Hanson, & Helmus, 2010; Helmus, 2009). This assertion has never been tested empirically with the 20 samples comprising the three primary Static-99R reference groups; nor have the external risk factors believed to cause this differentiation ever been identified. In other words, it is unknown whether the density of risk factors external to the Static-99R is a valid conceptualization for explaining the differences in sexual recidivism base rates between the reference groups. Moreover, the developers have failed to operationalize the term "density," so it is unknown how to define, quantify, or measure it. In essence, the developers have introduced reference-group-selection procedures devised on a premise with unknown validity and unproven reliability. For the purposes of the following analysis, however, the author presumes that the density of risk factors external to the Static-99R account for discrimination of the reference groups.

Two, the current selection procedures (Phenix et al., 2012) are premised on the idea that the sexual recidivism base rate contained in the experience table that a user chooses is similar to that of the group representing the evaluee. The developers have not produced data supporting this assumption. The fit in sexual recidivism base rate between the two groups is essential in producing accurate score-wise risk estimates to compare to the individual being assessed (Donaldson & Wollert, 2008). Consequently, users who employ the current reference-group-selection procedures may unwittingly select an experience table that results in the inaccurate estimation of risk at the score assigned to the evaluee.

Since the inception of the Static-99R reference-group-selection procedure (Phenix et al., 2009), the developers have published peer-reviewed articles documenting the change in scoring the age-at-release item (Helmus, Thornton, Hanson, & Babchishin, 2012), the reporting of relative risk data (Hanson, Lloyd, Helmus, & Thornton, 2012), and the variability in base rates and score-wise risk estimates across the various samples comprising the Static-99R data set (Helmus, Hanson, Thornton, Babchishin, & Harris, 2012). In July 2012, the developers released a revised version of the Evaluators' Handbook (Phenix et al., 2012) that provides the current recommended procedures to select Static-99R reference groups. To date, the developers have not produced data informally or through peer-reviewed publication establishing the reliability and validity of the Static-99R reference-group-selection system, yet many clinicians have been using it for nearly four years as if its reliability and validity have been proven.

As readers will see, the following analysis of the Static-99R reference-group-selection procedure is based largely on unpublished material produced by the developers. Without peer-reviewed publication, Static-99R users must obtain information to implement the reference-group-selection procedures from disparate sources, including the Static-99 website and trainings conducted by the developers. In justifying the application of the reference-group-selection system in forensic and clinical practice, users are expected to rely on the developers' assurances that the hypothetical, empirically untested reference-group-selection system is reliable, works as intended, and produces accurate risk estimates for the individual being assessed.

The overarching aim of this article is to fill the gap in peer-reviewed literature about the Static-99R reference-group-selection method so that practitioners have a frame of reference to assess the reliability and validity of the procedures as applied in sexual recidivism risk assessments. Since Wollert (2010) has addressed the reliability and validity issues associated with the cohort-matching system, this article will focus on the preferred reference-group-selection procedure that assesses the density of risk factors external to the Static-99R. First, this article briefly describes background information about the statistical methods the developers employed to discriminate among the Static-99R reference groups. Second, the author examines the reliability and validity of using risk factors external to the Static-99R to select reference groups. Finally, the Discussion and Conclusions section explores alternative methods to select Static-99R reference groups.

Evolution of the Static-99R Reference Groups

Based on concerns raised by Abbott (2009), Helmus (2009) analyzed the possible influence of moderator variables (e.g., sample type, offender type, age at release, country, or era of study) effecting the discrimination of Static-99R reference groups. The meta-analyses of sample type (e.g., routine vs. preselected) involved eight iterations that produced a final solution (Preselected Version 3), as reported by Helmus (2009), Table 15, where observed differences in base rates of sexual recidivism distinguished three reference groups designated as Routine Corrections ("RC"), Preselected Treatment Need ("PTN"), and Preselected High Risk Needs ("PHRN"). The Non-Routine Corrections ("NRC") reference group is a compilation of the samples comprising the PTN and PHRN reference groups and three other studies (Phenix et al., 2012) and is considered only when applying the cohort-matching selection system.

After observing the statistically significant base rates of sexual recidivism across the three reference groups, the developers engaged in post-hoc, logical inference to explain the characteristics they hypothesized as distinguishing the three reference groups (Helmus, 2009). The developers assert that certain judicial, correctional, or administrative decisions were made that led offenders to be placed in different settings such as general prison populations, community supervision, outpatient or correctional treatment programs, and secure facilities for high-risk sex offenders. The developers posit that this decision-making process relied on the consideration of the density of risk factors external to the Static-99R and they characterize this as the "preselection effect."

Phenix et al. (2012; p. 35) state that the magnitude of the preselection effect (i.e., low, moderate, or high) is the basis by which to discriminate the RC reference group from the PTN and PHRN reference groups, as well as distinguishing differences across the two non-routine reference groups (PTN and PHRN). This is considered the preferred reference-group-selection method with the cohort-matching procedure serving as an alternative. The astute reader may recognize that the Evaluators' Handbook contains internally inconsistent instructions on reference-group-selection methods. Phenix et al. (2012: p. 32) also advise users who are considering using routine samples to select the RC reference group when a Static-99R experience table devised on a local sexual offender group is not available. This recommendation presumes that the RC reference group is representative of the typical sexual offender from the routine corrections population. If the eight studies comprising the RC reference group were sampled representatively from a larger predefined routine corrections population, then it would be assumed that the resulting sexual-recidivism base rate and score-wise risk estimates would generalize to other samples drawn from this population within a certain margin of error. Data from Helmus (2009) appear to contradict this assumption.

Table 1
Static-99R RC Reference Group:
Descriptive Data for 8 Studies

Routine Corrections	5-Year Base Rate %	n	n_r [+]
Bartosh et al. (2003)	13.3	90	12
Bigras (2007)	9.2	207	19
Boer (2003)	3.7	299	11
Craissati et al. (2008)	7.5	200	15
Eher et al. (2009)	2.0	151	3
Epperson (2003)	10.6	151	16
Hanson et al. (2007)	0	31	0
Langström (2004)	5.4	1,278	69
Total	6.0	2,407	145

n_r = number of recidivists

The RC reference group is comprised of eight studies as listed in Table 1. The Bartosh, Garby, Lewis, & Gray (2003) and Epperson (2003) studies are from the United States with the remaining samples coming from Canada, England, Austria, and Sweden. Using meta-analytic techniques, Helmus (2009) determined the sexual recidivism base rates among the eight samples varied more than would be expected by chance. This finding reflected that the samples comprising the RC reference group unlikely represented the same population of sexual offenders. After removing the two United States samples, Helmus (2009) determined the base rates for the remaining six samples appeared to vary only by chance. For reasons not clearly explained, Helmus (2009) decided to include the two United States

samples in the RC reference group experience table. These circumstances raise practical concerns when clinicians decide to select the RC reference group based on the assumption that the evaluee best matches this group.

The disparity in base rates among samples from the United States and other countries means that clinicians lack confidence that the 6.0% base rate for the RC reference group accurately represents the recidivism potential for sexual offenders from routine corrections populations. The 6.0% sexual recidivism base rate and corresponding score-wise recidivism rates from the RC reference group may underestimate the risk potential of offenders from the United States and inflate the recidivism potential of sexual offenders from other countries. The variability in sexual reoffense rates observed among the RC reference group studies in Table 1 suggests that the base rates among routine correctional samples may vary in meaningful ways from the aggregate base rate of 6.0%. Clinicians should be careful in accepting the 6.0% base rate and the resulting score-wise risk estimates as being generalizable to offenders who are presumed members of the routine corrections population. As demonstrated by Donaldson and Wollert (2008), the score-wise risk estimates are a direct function of the base rate of sexual recidivism and score-wise likelihood ratios. Consequently, clinicians who rely on the score-wise risk estimates calculated from the 6.0% sexual recidivism base rate would report inaccurate score-wise risk estimates if the group representing the evaluee had a sexual recidivism base rate of 2.0% or 13.3%, as seen, for example, in two studies reported in Table 1. This issue is further explored in the Discussion and Conclusions section.

Using a Dynamic Risk-Assessment Instrument For Selecting a Static-99R Reference Group

Phenix et al. (2012; p. 35) conceptualize the preselection effect as the density of risk factors external to the Static-99R and instruct users to quantify the density of external risk factors by applying a "dynamic risk-assessment scale." Phenix et al. (2012) further describe three levels of preselection corresponding to the Static-99R reference groups (as represented in parenthesis following each reference group): RC (low), PTN (moderate), and PHRN (high). Recommendations for selecting a specific dynamic-risk instrument to assess the preselection effect are not contained in the Evaluators' Handbook (Phenix et al., 2012); however, instructions for selecting a measure are found in other sources. Thornton, Hanson, and Helmus (2010) endorse the Stable-2007 (Hanson, Harris, Scott, & Helmus, 2007), the Forensic Structured Risk Assessment ("Forensic SRA;" Thornton & Knight, 2009), and the Violence Risk Scale- Sexual Offender ("VRS-SO;" Olver et al., 2007) to quantify the preselection effect, but they provide no guidance to users as to what scores from the instruments correspond to each preselection level. Thornton (2010) proposes the Structured Risk Assessment- Forensic Version ("SRA-FV"), a revision of the Forensic SRA, as a means to quantify the preselection effect for selecting Static-99R routine and non-routine reference groups.

If evaluators are to use a dynamic risk-assessment instrument to select a Static-99R reference group, there must be three ranges of scores that are mutually exclusive ("cut-

scores"), one for each Static-99R reference group (RC, PTN, and PHRN). Table 2 reports two Static-99R reference-group-selection models using cut-scores from the SRA-FV and Stable-2007.

In unpublished instructions, Thornton (2011) proposes a specific system for using an evaluee's score on the SRA-FV to select a Static-99R reference group. Although Thornton's (2011) proposal produces seven levels of SRA-FV scores, he instructs evaluators to use the evaluee's score to select one of three Static-99R reference groups: RC, PTN, or PHRN. Table 2, column 2, shows Thornton's (2011) recommendations for using an evaluee's SRA-FV score to choose a Static-99R reference group. The specific ranges of SRA-FV scores in Table 2 were devised by Thornton (2011).

While the developers have not proposed a reference-group-selection model for the Stable-2007, Phenix et al. (2012) describe three levels of preselection effect (low, moderate, and high) that can be quantified using dynamic-risk measures. Hanson and Harris (2008) provide Stable-2007 cut-scores designated as low, moderate, and high dynamic risk or need, as reported in Table 2, column 3. The reader should note that no one has proposed that the Stable-2007 mutually exclusive score ranges (0-3, 4-11, and ≥ 12) represent the ideal cut-off scores for choosing a Static-99R reference group. For the purposes of this analysis, however, the author assumes the Stable-2007 cut-off scores shown in Table 2, column 3 correspond to the three Static-99R reference groups: RC, PTN, and PHRN.

Table 2
Descriptive Statistics for Three Instruments Used to Measure Static-99R Preselection Effects

	SRA-FV	Stable-2007	VRS-SO[c]
N of items	10	13	17
N of risk domains	3	5	3
Possible item points	0 - 2	0 – 2	0 - 3
Range of total score	0 - 6	0 – 26	0 - 52
Scores associated with preselection effect (Static-99R reference group):			
Low (RC)	≤ 2.3[a]	0 – 3[b]	N.R.[d]
Moderate (PTN)	2.4 – 3.2[a]	4 – 11[b]	N.R.[d]
High (PHRN)	≥ 3.3[a]	≥ 12[b]	N.R.[d]

[a] From Thornton (2011); [b] From Hanson & Harris (2008); [c] Pretreatment dynamic risk factor; [d] Not reported by instrument developer

For the VRS-SO pretreatment dynamic-risk factor, Olver, Wong, Nicholaichuk, & Gordon (2013) report scores in four mutually exclusive score bands. No one has proposed a procedure by which to segregate the VRS-SO pretreatment dynamic risk total scores

into three mutually exclusive score bands to select among the three Static-99R reference groups.

It is not surprising that the Evaluators' Handbook (Phenix et al., 2012) lacks specificity regarding the selection of instruments to assess risk factors external to the Static-99R because research in this area is limited. At the time of the release of the Evaluators' Handbook (Phenix et al., 2012), the VRS-SO had not been administered to any of the 20 studies comprising the Static-99R reference groups. The Stable-2007 had been administered to 31 participants from a single study within the RC reference group (Hanson et al., 2007). The SRA-FV was devised on the Bridgewater sample (Knight & Thornton, 2007), which is one among six studies comprising the PHRN reference group. Later, Thornton (2011, 2010) used the Bridgewater data to statistically contrive a Static-99R reference-group-selection model as reflected in Table 2, column 2. Readers should note that Thornton (2010) did not administer the SRA-FV to individuals from any of the other 19 samples comprising the three Static-99R reference groups.

Reliability of Quantifying the Preselection Effect

In completing the Stable-2007, SRA-FV, and VRS-SO, raters assign a specific numerical value to each item according to the standardized rating or coding guidelines provided by each instrument developer (Cf., **Stable-2007**: Hanson & Harris, 2008; **SRA-FV**: Thornton, 2012 & 2011; **VRS-SO**: Wong, Olver, Nicholaichuk, & Gordon, 2003; Olver et al., 2013). The item scores are summed to obtain a total score. Determining the selection of an appropriate Static-99R reference group is therefore dependent on raters reliably ascertaining total instrument scores that fall within the critical region defining each level of the preselection effect.

It is apparent from reading the three instrument manuals that the rating criteria for items are inherently subjective which, in turn, is likely to result in users inconsistently applying the rating instructions. The score for each risk factor is determined by the fit of the evaluee's behavior to the scoring guidelines. In arriving at this decision, users must consider a wide array of complex behavioral patterns displayed by an evaluee that occur at specific rates over the duration of the specified assessment time frames (e.g., within one year of assessment on the Stable-2007 or a lifetime for the SRA-FV). The reliability of the assessment procedure is obviously affected by such factors as the ambiguity of terms, users assigning differing meaning to the rating criteria, the extent and quality of available information to rate factors, and rating biases (e.g., anchoring bias, negative and positive halo effects, or recency bias). Moreover, the subjectivity of the scoring methods opens the door to an allegiance effect diminishing reliability (Murrie, Boccacinni, Caperton, & Rufino, 2012; Boccaccini, Murrie, Caperton, & Hawes, 2009; Murrie, Boccaccini, Turner, Meeks et al., 2009). No matter what the source, variability in scorer judgment is likely the largest threat to the reliability of the results obtained from the Stable-2007, SRA-FV, and VRS-SO.

The Specialty Guidelines for Forensic Psychology includes, "Forensic practitioners seek to provide opinions and testimony that are sufficiently based upon adequate scientific

foundation, and reliable and valid principles and methods that have been applied appropriately to the facts of the case" (Guideline 2.05, American Psychological Association, 2011). How reliable is reliable enough? Qualitative labels provided by Cicchetti (1994) for inter-rater reliability are "poor" for Intra-class Correlation (ICC) values less than .40, "fair" for values between .40 and .59, "good" for values between .65 and .74, and "excellent" for values between .75 and 1.0. More relevant to the developers' proposals (Phenix et al., 2012; Thornton, 2011, 2010) to use dynamic risk-assessment instruments to guide the choice of a Static-99R reference group, Heilbrun (1992) recommends a minimum floor of .80 in the reliability coefficient when selecting which measures to use in forensic applications. Marshall (2006) argues for a higher inter-rater agreement of not less than .90.

One study examines the inter-rater agreement for the SRA-FV total score. In unpublished research, Sachsenmaier, Thornton, and Olson (2011) report an ICC of .55 among 19 psychologists who completed the SRA-FV on 69 individuals committed to a sexually violent predator civil confinement center in Wisconsin. Cicchetti (1994) considers this ICC value as a fair level of agreement among the raters. The ICC value of .55 represents the proportion of true-score variance (e.g., the extent to which the SRA-FV measures long term vulnerabilities). Subtracting this value from 1.0 reveals the proportion of error variance,[1] which is .45. When the error variance approximates the true-score variance, it becomes obvious that the SRA-FV total score will inform as often as it will mislead.

The Stable-2007 was developed and validated on samples of sexual offenders under community supervision (Hanson et al., 2007). Inter-rater reliability data was not provided for the developmental sample but it is interesting that two studies of sexual offenders in custody report it. Eher, Matthes, Schilling, Hauber-MacLean, and Rettenberger (2012) randomly selected 15 Stable-2007 protocols obtained from male prisoners in Austria and found excellent inter-rater reliability for the total score (ICC = .90). Fernandez (2008) studied 55 Stable-2007 ratings completed by correctional program officers for incarcerated Canadian sexual offenders. The ICC for the total score was .92.

The VRS-SO validation research (Olver, Wong, Nicholaichuk, & Gordon, 2007) studied participants from a high-intensity sexual offender treatment program operated in a Canadian prison. Thirty-five randomly selected cases were studied for inter-rater reliability. The ICC (single measure) was .74 for pretreatment dynamic risk total score. Beggs and Grace (2010) reported VRS-SO inter-rater reliability derived from a study of child molesters incarcerated at a New Zealand special treatment unit. ICC coefficients were computed on 23 cases scored by two raters. The pretreatment dynamic risk total score achieved an ICC of .90.

[1] In the true-score model, error variance is attributable to random sources irrelevant to the measurement of the trait or ability the instrument purports to measure in an observed score or distribution of scores. Common sources of error variance include those related to test construction (including item or content sampling), test administration, and test scoring and interpretation (Cohen & Swerdlik, 2001).

Available research reveals a level of inter-rater reliability for the SRA-FV that is inadequate for any forensic application. In contrast, the studies mentioned in this section suggest that inter-rater reliability is not an inherent barrier to the use of the Stable 2007 or the VRS-SO. The reliability of the procedure, however, is not sufficient to determine whether the proposed reference-group-selection methods work as intended. The validity of the preselection effect models is therefore explored in the next section.

Do Instrument Scores Accurately Quantify the Preselection Effect?

Validity determines the extent to which the Static-99R reference-group-selection procedures perform as intended. The validity of the Static-99R reference-group-selection system is premised on three untested assumptions. One, the preselection effect can be categorized into low, moderate, and high levels. Two, a single Static-99R reference group is associated with only one level of preselection effect. Three, each level of preselection effect can be quantitatively determined based on non-overlapping cut-scores from the Stable-2007, SRA-FV, and VRS-SO. For example, these assumptions would lead to the hypothesis that members from the PHRN population only evidence high levels of preselection as measured by SRA-FV scores equal to or greater than 3.3. The assumptions underlying the validity of the preselection effect model have not previously been tested, despite its widespread use. This section describes the methods used to test the three assumptions undergirding the Static-99R reference-group-selection system and reports the results of the analysis.

Methods and data analysis. In a conference workshop, Hanson and Thornton (2012) reported results from the administration of the SRA-FV, Stable-2007, and VRS-SO among samples that were considered representative of RC, PTN, and PHRN populations. The study examined 15 samples, two of which were part of the 20 samples comprising the three Static-99R reference groups (Hanson et al., 2007; Knight & Thornton, 2007). Appendix A provides the references for the studies. The research by Hanson and Thornton (2012) explored whether the level of preselection, as measured by the mean scores for the three measures, accounted for the discrimination of the base rate differences for the three Static-99R reference groups. The researchers found that the mean scores from the three instruments suggested that certain levels of preselection corresponded with each Static-99R reference group. The results further indicated that it might be possible to test whether Static-99R reference groups could be selected using mutually exclusive cut-off scores from the SRA-FV, Stable-2007, and VRS-SO.

To obtain the data necessary to test the reference-group-selection system, the author contacted the developers to request the following information for each reference group by each dynamic-risk measure: mean score, standard deviation, range of scores, number of participants, and number of studies. Dr. Hanson graciously provided the data with the exception of the range of scores, which was not available. The data was used for three purposes, including to test Thornton's (2011) SRA-FV selection model, to determine whether the Stable-2007 levels of dynamic risk (low, moderate, and high) correspond to the selection of Static-99R reference groups as reflected in Table 2, and

to explore the feasibility of constructing a selection model based on the VRS-SO pre-treatment dynamic total score.

Table 3
Stable-2007, SRA-FV, and VRS-SO: Number of Participants (N) and Studies (k) and Descriptive Statistics by Static-99R Reference Groups

Reference Group (Preselection Effect)	N (k)	Mean Score	95% Confidence Interval	Standard Deviation	Expected Range of Scores[a]
Stable-2007 (maximum score = 26)					
RC (low)	1,198 (2)	7.06	6.7 – 7.4	5.15	0 – 22.51
PTN (moderate)	646 (6)	10.99	10.1 – 11.8	3.92	0 – 22.75
PHRN (high)	189 (1)	14.70	14.2 – 15.2	3.40	4.5 – 24.9
SRA-FV (maximum score = 6)					
PTN (moderate)	439 (3)	2.22	2.1 – 2.3	.922	0 – 4.99
PHRN (high)	513 (2)	3.26	3.2 – 3.3	.761	0.98 – 5.54
VRS-SO (maximum score = 52)					
PTN (moderate)	481 (3)	20.74	20.1 – 21.4	7.23	0 – 42.43
PHRN (high)	510 (2)	27.23	26.7 – 27.8	6.02	9.17 – 45.29

[a] Computed as ± 3 standard deviations from the mean score

Table 3 reports the descriptive data for each measure in three panels, with each panel representing a specific instrument. The number of studies and sample sizes for each reference group are reported in column 2. The mean scores and standard deviations provided by Dr. Hanson are reported in columns 3 and 5, respectively. The 95% confidence intervals for the mean values for each instrument are presented in column 4 and the author computed these data using the following formula:

$$\bar{x} \pm z\alpha_{/2} \left(\frac{\hat{s}}{\sqrt{n}}\right) \quad (1)$$

Where,

$\bar{x}$ = Estimate of population mean
$\hat{s}$ = Estimate of population standard deviation
α = 1 – (Confidence Level/100)
$z\alpha_{/2}$ = z-table value
n = Sample Size

Since Hanson and Thornton (2012) did not report actual ranges of scores, the author computed expected ranges of instrument scores by adding and subtracting three stan-

dard deviation units from the mean score. These data are shown in the last column of Table 3. Based on the statistical properties of a normal distribution of scores around the mean, it is expected that 99% of all scores in a population fall within three standard deviations below and above the mean. The lower limits for some of the ranges of scores fell below zero. In such instances, the lower limit was truncated at zero since the three measures do not report negative score values. Inspection of the last column in Table 3 reveals that the upper limit of the score range does not exceed the maximum score for each measure. This suggests that the standard deviations for scores corresponding to each reference group adequately estimate the range of scores for each instrument.

Results. Column 4 in Table 3 demonstrates that the 95% confidence interval about the mean instrument scores does not overlap between preselection levels. Since Helmus (2009) has shown that each reference group is statistically independent, as reflected by sexual recidivism base rates, a rule-of-thumb method can be applied to determine whether the Stable-2007, SRA-FV, and VRS-SO mean scores are statistically different across the three levels of preselection effect. Cumming and Finch (2005) indicate that non-intersecting confidence intervals among independent groups reflect statistically significant differences at $p < .01$. This indicates that mean instrument scores distinguish one reference group from another and this finding is consistent with the developers' hypothesis that higher base rate groups exhibit a greater density of risk factors external to the Static-99R. As discussed in the concluding section of the article, significant differences in Stable-2007, SRA-FV, and VRS-SO mean scores are insufficient to devise a valid system to select Static-99R reference groups.

The results summarized in Table 3 further reveal a wide dispersion of Stable-2007, SRA-FV, and VRS-SO scores around the mean values at each level of preselection. In fact, scores span nearly the entire range of values for each instrument. These data reflect that the mean dynamic-risk scores may be statistically different but sexual offenders within each reference group show remarkably similar variation in their dynamic-risk scores. Because the variation in dynamic-risk scores for each instrument is homogeneous across reference groups, clinicians cannot expect sexual offenders who are assessed on the Stable-2007, SRA-FV, or VRS-SO to demonstrate restricted ranges of dynamic-risk scores corresponding to low, moderate, or high levels of preselection. Taken together, these findings suggest that clinicians cannot rely upon the evaluee's total dynamic-risk score to select a single Static-99R reference group. This hypothesis is explored further next.

Table 4 reports the data used to assess the validity of the Stable-2007 and SRA-FV selection models. For the purposes of this comparison, the author rounded the mean SRA-FV scores reported in Table 4, column 3, to one decimal place, consistent with Thornton's (2011) recommendation for reporting the level-of-need scores. The following table summarizes the results from this analysis.

The results reported in Table 4, columns 2 and 3 reveal that Thornton's (2011) model for selecting Static-99R reference groups failed to replicate in the five studies where the

SRA-FV was administered. The results contradict that a single Static-99R reference group can be chosen based on a restricted range of SRA-FV total scores. This is apparent in two ways when inspecting the results from Table 4.

Table 4
Testing SRA-FV and Stable-2007 Selection Models

Reference Group (Preselection Effect)	SRA-FV Selection Model[a]	SRA-FV Results Mean (Range)	Stable-2007 Selection Model[b]	Stable-2007 Results Mean (Range)
RC (low)	≤ 2.3	-	0 - 3	7.06 (0 – 22.51)
PTN (moderate)	2.4 – 3.2	2.2[c] (0.00 – 4.99)	4 – 11	10.98 (0 – 22.74)
PHRN (high)	≥ 3.3	3.3[c] (0.98 – 5.54)	≥ 12	14.70 (4.5 – 24.9)

[a] Thornton (2011); b Hanson & Harris (2008); [c] mean scores from Hanson and Thornton (2012) rounded to single decimal place according to SRA-FV scoring instructions (Thornton, 2011)

One, the mean SRA-FV score of 2.2 for the PTN reference group (moderate preselection) falls in the critical region of ≤ 2.3 that defines the RC reference group (low preselection). Assuming a normal distribution of total scores, it is reasonable to conclude that more than one-half of the SRA-FV scores from the PTN reference group are likely consistent with the scores defining samples thought to have less preselection effect. Even though the mean SRA-FV score of 3.3 is at the lower limit for the cut-off associated with the PHRN reference group, it is likely that nearly half the scores below the mean fall within the lower risk PTN group.

Two, the SRA-FV cut-scores for the PTN and PHRN reference groups overlap substantially. As a result, there is a high probability that the SRA-FV score assigned to an evaluee will be consistent with two Static-99R reference groups instead of one reference group as Thornton's (2011) model posits. While the SRA-FV was not tested on samples representing routine correctional settings, it is reasonable to assume that the SRA-FV scores from the RC reference group would likely span a large portion of the range of possible SRA-FV values, thus overlapping the other two reference groups by wide margins.

The reader should note that Hanson and Harris (2008) designed the Stable-2007 cut-scores to be associated with low, moderate, and high dynamic risk, as shown in Table 4, column 4. For the purposes of this analysis, the author assumed that the three Stable-2007 cut-score ranges correspond to the respective Static-99R reference groups. The results in the last column of Table 4 reveal that the dispersion of Stable-2007 scores around the mean values for each Static-99R reference group is remarkably similar. Like the SRA-FV results, it is apparent that sexual offenders at any level of preselection demonstrate wide variation in Stable-2007 scores that run the gamut of dynamic-risk levels. As a result, clinicians will encounter the quandary of an evaluee's Stable-2007 score being representative of two, if not all three, Static-99R reference groups. Moreover, the homogenous variation in Stable-2007 scores across reference

groups bodes poorly for developing a reference-group-selection model that relies on mutually exclusive cut-off scores.

A goal of this study was to explore whether the VRS-SO pretreatment dynamic-risk cut-scores could be developed to select a single Static-99R reference group. Inspection of the last column in Table 3, panel 3, casts substantial doubt that such an endeavor would be successful. Like the two other instruments, the dispersion of VRS-SO pre-treatment dynamic-risk total scores around the mean values for the reference groups are so homogeneous that it would be virtually impossible to establish mutually exclusive cut-off scores to select a single Static-99R reference group.

Discussion and Conclusions

In developing the Static-99R, Helmus (2009) found that twenty convenience samples sorted into three groups, based on statistically significant differences in sexual recidivism base rates. The three groups were designated as RC, PTN, and PHRN, and the developers issued separate sexual recidivism experience tables for each reference group (Phenix et al., 2012). The developers used post-hoc logical inference to hypothesize that the density of risk factors external to the Static-99R explain the statistically significant differences in sexual recidivism base rates among all the samples and the three reference groups. Based on this conceptualization, the developers recommend using SRA-FV, Stable-2007, or VRS-SO to quantify risk factors external to the Static-99R. The reference-group-selection procedures have been advanced in clinical and forensic practice without virtually any empirical examination as to the reliability or validity of the methods.

In their initial effort to capture the preselection effect, the developers devised a qualitative system for selecting reference groups based on cohort matching. This procedure introduces a fourth reference group (non-routine corrections) into the selection mix (Phenix et al., 2012; p. 19). Wollert (2010) points out that the reliability for this procedure has not been established and it is fraught with misclassification error that degrades the accuracy of score-wise risk estimates. To ameliorate misclassification error, Wollert (2010) recommends that users consider a single-experience table representing the aggregate of the four Static-99R reference groups, but the developers have not provided this data.

More recently, the developers appear to have moved in a direction to quantify the preselection effect using instruments that purport to measure risk factors external to the Static-99R. This selection model hinges on the conceptualization that the total scores derived from measures of risk factors external to the Static-99R quantify the preselection effect. Moreover, it is speculated that the total instrument scores can be grouped into mutually exclusive cut-scores that discriminate each reference group. Phenix et al. (2012) describe this as the preferred method for reference group selection and the cohort-matching procedure is applied when users do not complete an instrument that measures other risk factors. Phenix et al. (2012) do not recommend a specific instrument to assess the preselection effect but other publications by members of the Static-

99R team suggest the use of the Stable-2007, SRA-FV, and VRS-SO (Thornton et al., 2010; Thornton, 2010 & 2011). This study appears to be the first attempt to critically analyze the reliability and validity of applying the three instruments to quantify the hypothesized preselection effect when choosing Static-99R reference groups.

The limited number of inter-rater agreement studies for the Stable-2007 and VRS-SO demonstrate that these measures appear to achieve a minimum level of reliability for forensic use (Heilbrun, 1992). It is conceivable that the Stable-2007 and VRS-SO total scores could be used to establish a Static-99R reference-group-selection system. On the other hand, data from Sachsenmaier et al. (2011) show the reliability for the SRA-FV total score falls far below minimum standards for forensic practice. There is an urgent need to revise the SRA-FV scoring system to increase its reliability, if this is even possible. Until then, clinicians who choose to conduct SRA-FV assessments would be hard pressed to justify its use statistically (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2003; Anastasi & Urbina, 1997) and ethically (American Psychological Association, 2011 & 2010) when selecting Static-99R reference groups and, for that matter, in other forensic applications. Even if the reliability of the Stable-2007, SRA-FV, and VRS-SO achieve acceptable levels for forensic practice, this does not ensure that the resulting models to select Static-99R reference groups are valid.

Validity of Selection Models

Hanson and Thornton (2012), Phenix et al. (2012), and Thornton et al. (2010) contend that that hypothesized preselection effect can be quantified based on total scores from measures of risk factors external to the Static-99R. Using data obtained from Hanson and Thornton (2012), the results from the analysis were consistent with the developers' hypotheses in part. On the one hand, the Stable-2007, SRA-FV, and VRS-SO mean total scores were in the expected direction (e.g., higher mean scores for higher base rate groups), as predicted by Hanson and Thornton (2012). On the other hand, the Stable-2007 and SRA-FV selection models overlapped by such wide margins that it would be virtually impossible to select a single Static-99R reference group.

The Stable-2007 and SRA-FV selection models assume that mutually exclusive cut-scores define the levels of preselection (i.e., low, moderate, or high) associated with the Static-99R reference group (RC, PTN, or PHRN, respectively). The total instrument score should result in the selection of a single Static-99R reference group. The results of this analysis refute this hypothesis, as the total instrument score assigned to an evaluee would likely be consistent with choosing at least two if not all three of the Static-99R reference group options. Only the extreme upper ends of the score distribution for the Stable-2007 and SRA-FV were unaffected by score duplication. This finding appears of limited utility forensically as it affects a tiny proportion of the score distribution. It is clear that these data directly contradict the contention of the developers (Phenix et al., 2012; Thornton, 2011; Thornton et al., 2010) that the density of risk factors external to the Static-99R can be used to select a single reference group.

The results from the VRS-SO studies were similar to the findings from the Stable-2007 and SRA-FV. As seen in Table 3, the range of pretreatment dynamic risk total scores overlapped substantially for the PTN and PHRN reference groups. The participants from the PHRN reference group were distinguished from the PTN group by the fact that they did not score below nine, but this garners little forensic utility because scores of nine or greater are also representative of the PTN reference group score distribution. The members of the PHRN reference group also scored slightly higher overall on the pretreatment dynamic risk total score than their PTN counterparts, but such discrimination has little utility since it affects a very small proportion of the score distribution. Like the other two instruments, it would be expected that the range (proxy of score distribution) of pretreatment dynamic risk total scores for the RC reference group would overlap substantially with the PTN and PHRN reference groups. Taken together, these results suggest that the prospect is poor for constructing a valid selection model based on mutually exclusive VRS-SO pretreatment dynamic-risk cut-scores.

Possible Reasons for Invalidity of the Selection Models

While the Stable-2007, SRA-FV, and VRS-SO data from Hanson and Thornton (2012) did not test inter-rater agreement, this potential source of error would affect the validity of the reference-group-selection methods. Low levels of inter-rater agreement for the total SRA-FV, Stable-2007, and VRS-SO scores may explain the variability in score ranges that caused the wide overlap in cut-scores associated with each Static-99R reference group. The necessity for precise measurement cannot be over-emphasized. However, adequate reliability does not ensure that the Static-99R selection system is valid. It appears that the three selection models tested in this study failed to perform as intended because, likely, the magnitude of risk factors external to the Static-99R was an incorrect assumption to explain the statistically significant differences in sexual recidivism base rates between the three reference groups.

The failure to establish the validity of the selection models is not surprising when considering that the developers (Helmus, 2009) employed post-hoc logical inference to hypothesize the preselection effect. Devising a selection system based on non-overlapping score ranges runs counter to the expectation that the range of observed instrument scores is likely to be homogeneous among different groups. For instance, the selection system that Thornton (2011; 2010) devised is akin to the expectation that Static-99R scores within each reference group would be restricted to a certain range without overlapping other reference groups such that the Static-99R scores for members of the RC reference group would be ≤ 2, 3 to 4 for the PTN reference group, and ≥ 5 for the PHRN reference group. To the contrary, in all Static-99R reference groups, the participants demonstrate a full range of scores. This observation mirrors the results from this study demonstrating similar dispersion of Stable-2007, SRA-FV, and VRS-SO total scores across the three reference groups.

The mean instrument scores observed by Hanson and Thornton (2012) produce statistically significant differences in the levels of preselection effect across all three instruments. The significant differences in mean scores for the three measures suggest that,

as the base rate increases across reference groups, there is a concomitant escalation in levels of external risk factors. This observation might reflect a true difference between reference groups based on total average scores from the Stable-2007, SRA-FV, and VRS-SO, or it might represent a high correlation between the Static-99R and the measures of external risk factors. If the Static-99R and external risk measures account for a high proportion of the same variance associated with sexual recidivism, then the significant differences in mean instrument scores across preselection levels may only represent the high degree of co-variation. Alternatively, if a sufficient number of external risk factors redundant with the Static-99R produce error of sufficient magnitude, then it would confound the sensitivity of a few items contributing to the hypothesized preselection effect. The extent of shared variance among measures needs further study but, nevertheless, it is obvious that the homogeneous dispersion of instrument scores across the three levels of preselection presents an insurmountable obstacle in designing a Static-99R reference-group-selection model that relies on non-overlapping critical regions as proposed by Thornton (2011) or when using the existing Stable-2007 cut-off score system (Hanson & Harris, 2008).

Finally, there is a fundamental problem with proposing a selection model that is premised on post-hoc logical inference. It is simply unknown whether risk factors external to the Static-99R even explain the differences in base rates among the RC, PTN, and PHRN reference groups. The developers used post-hoc logical inference to hypothesize the preselection effect as discriminating different Static-99R reference groups without empirical validation. The selection models tested in this analysis may not have performed as hypothesized because of the speculative nature of their foundation.

Before advocating that clinicians employ untested selection models to choose Static-99R reference groups, it seems reasonable to expect the developers to prove the validity and reliability of their methods. As a first step, the developers could generate a list of possible risk factors external to the Static-99R. Next, the intercorrelation between the Static-99R and identified external risk factors could be computed to determine which factors add unique variance above the Static-99R. Any risk factors found to have statistical independence from the Static-99R could then be tested to determine if a reliable and valid selection model could be developed. If such a model were established, it would need to undergo sufficient replication studies (Schmidt, 2009) before it is deployed for forensic use. Until then, it seems scientifically and ethically indefensible for clinicians to apply speculative reference-group-selection methods in forensic practice.

Implications for Clinical and Forensic Practice

The most obvious conclusion for clinicians to draw from this analysis is to abandon the use of the Static-99R because it lacks a reliable or valid method by which to choose one of three or four reference groups to compare to the individual being assessed. Clinicians cannot have any level of confidence in their predictions of sexual recidivism risk when it is uncertain whether the correct reference group was selected. Short of scuttling the Static-99R reference-group-selection system, this article concludes by explor-

ing two alternate selection procedures. The first method is referred to as *base-rate matching* and relies on using the current Static-99R reference groups. The second selection procedure rejects choosing among four options in favor of a single reference group along the lines recommended by Wollert (2010). It is obvious that an alternate method for selecting a reference group is unnecessary when clinicians have a Static-99R experience table developed and validated on a local group of sexual offenders that is representative of the individual being assessed.

Base-rate matching. The base-rate-matching approach is premised on the concept of selecting a Static-99R reference group that has a base rate of sexual recidivism most similar to the group representing the individual being assessed. This method is grounded in mathematics showing that the recidivism estimate at each score on an actuarial instrument is a function of the observed base rate of sexual recidivism for the group (Donaldson & Wollert, 2008). Matching the base rate of sexual recidivism between the group representing the individual being assessed and a Static-99R reference group will produce the most accurate score-wise estimate to compare to the evaluee.

Donaldson and Wollert (2008) show how the risk estimate at a given score is determined largely by the observed base rate of sexual recidivism in the offender group. This is a direct effect whereby the risk estimate at each score on the actuarial measure increases as the base rate rises. The recidivism rate at each score is also further determined by the score-wise likelihood ratio. The risk estimates for the selected Static-99R reference group will be inaccurate as applied to the group representing the evaluee when the base rates of sexual recidivism differ substantially between the local population of sexual offenders and the selected Static-99R reference group. Helmus (2009) and Helmus, Hanson et al. (2012) document significant variability in base rates across the samples comprising the Static-99R experience tables. As a result, users cannot assume the base rate of sexual recidivism expressed in an experience table is consistent with the group representing the evaluee when applying either of the two recommended selection procedures. Rather, it makes statistical sense to select a Static-99R reference group that reflects a base rate of sexual recidivism similar to the local group representing the evaluee.

The base-rate-matching approach involves identifying a sexual recidivism rate representative of the local group to which the evaluee belongs. The local sexual recidivism base rate is relied upon to select a Static-99R reference group that has the most similar base rate. An exact match in base rates may not occur; however, a variation of a few percentage points will not make a meaningful difference in the validity of the score-wise risk estimate as applied to the evaluee. An example involving sexually violent predators illustrates this point.

The author has observed the practice of government evaluators who assess sexually violent predators in California by typically choosing the PHRN reference group when using either of the two Static-99R reference-group-selection procedures. The five-year sexual recidivism base rate of the PHRN reference group is 21%. Data from several

studies in California contradict this base rate as representative of the sexual recidivism potential for individuals being considered for involuntary civil confinement. The California Department of Corrections (2010) and the California Sexual Offender Management Board (2008) studied two separate cohorts of sexual offenders released from prison in 2006 and 2003, respectively, with an aggregate sample size of 11,298. The average three-year sexual recidivism rate was 3.4%. This figure suggests a five-year rate near 6%. A third research project by Padilla (2006) and as reported by Zavodny, DeClue, and Cohen (2012), found that that a group of 93 petitioned sexually violent predators released from custody at the commitment center in California reoffended sexually at a rate of 6% over the 4.7 year follow-up. The rate of sexually violent reoffense, as defined by the California Welfare and Institution Code 6600 (b), was 4.3% over the same period. These base-rate data support the selection of the Static-99R RC reference group that has a 6% sexual recidivism rate over a five-year follow-up. The comparability of sexual recidivism base rates between the local offender group and the RC reference group would produce the most accurate score-wise sexual recidivism rates to compare to the SVP candidate being assessed. Choosing the PHRN reference group would instead result in reporting score-wise risk estimates that are approximately three times greater than would be expected in a lower base rate population.

In practical application, a local base rate of sexual recidivism may not be available to aid in the selection of a Static-99R reference group. In this event, evaluators could turn to nationally representative studies of sexual recidivism base rates to assist in selecting a Static-99R reference group. For instance, Wollert and Waggoner (2009) report a five-year sexual recidivism base rate of 6.5% among 17,697 participants from the United States. Of the 9,691 sexual offenders released from prisons in 15 states during 1994, Langan, Schmitt, and Durose (2003) found that 5.3% were rearrested for another sexual offense within three years. This would suggest an 8.8% sexual reoffense rate over five years.

Clinicians relying upon national studies must be sensitive as to how sexual recidivism base rates vary across samples or jurisdictions (Helmus, Hanson et al., 2012). The average base rate of sexual recidivism determined by the aggregated data may not accurately represent the probability of sexual reoffense for some of the independent groups comprising the combined sample. As a result, the base rate may be incompatible with that of the local jurisdiction representing the evaluee. This potential source of inaccuracy could be alleviated by conducting sexual recidivism studies at the local level (i.e., state-wide). Whenever possible, it is preferable to produce Static-99R experience tables for local jurisdictions.

A limitation of the base-rate-matching approach is the assumption that the score-wise likelihood ratios are similar between the selected Static-99R reference group and the local group representing the evaluee. The accuracy of the score-wise risk estimate as applied to the individual would be affected adversely should the score-wise likelihoods vary substantially between the selected Static-99R experience table and the local population representing the evaluee. The same problem exists when using any method to select an actuarial experience table to compare to the evaluee. Clinicians should rec-

ognize and consider this source of potential error when rendering conclusions and make it known in reports and testimony.

Aggregate reference group. Consistent with the recommendation of Wollert (2010), Table 5 reports data for a single experience table comprising all 23 Static-99R samples and explores its application in risk assessments. The developers have not released this data; however, the author compiled it, as described below, and refers to the data-set as the *Aggregate reference group*. The five-year base rate of sexual recidivism for the Aggregate reference group is 11.1% and the AUC is .714 (95% CI = .694, .734), which reflects moderate selection accuracy. Table 5 provides the observed and predicted sexual recidivism rates for the Aggregate reference group. The method by which the data in Table 5 were produced is explained in the next paragraph.

The observed recidivism rates for the Aggregate reference group, at each score and for the entire sample, were compiled by combining the frequency data from the NRC group (15 studies) and the RC group (8 studies). These data are contained in the detail recidivism tables provided by the developers.[2] Each score-wise observed recidivism rate was calculated by taking the number of recidivists at a given score divided by the corresponding total number of participants. Predicted estimates of sexual recidivism for the Aggregate reference group were computed using Logistic Regression. The slope parameter (B_1) to compute the predicted risk estimates is based on the frequency data for the aggregate sample. This computation differs from the one used by the developers. The Static-99R developers calculated the average slope parameter from twenty studies and used it in the regression analysis to arrive at the predicted score-wise risk estimates for each reference group (Helmus, 2009). Using the average slope parameter for twenty studies for the Logistic Regression has been found to produce spurious predicted risk estimates within reference groups, which is especially pronounced in the PHRN reference group (Donaldson, Abbott, & Michie, 2012). Precision of the observed score-wise risk estimate was computed using the 95% binomial exact CI (Clopper & Pearson, 1934), which is a preferred method when dealing with dichotomous outcome data. The accuracy of the risk estimate as applied to a presumed individual from the actuarial sample is represented by the 95% confidence interval for the individual ("CII"). Further details regarding the methods for calculating the predicted risk estimates and accuracy data about them have been previously explained by Donaldson and Abbott (2011).

[2] Available at http://www.static99.org/pdfdocs/detailed_recid_tables_static99r_2009-11-15.pdf

Table 5
Static-99R Aggregate Reference Group Experience
Table: Five-Year Observed and Predicted Risk Estimates

Static-99R Score	Total n of Recidivists	Total n	Observed Recidivism Rate (%)	95% CI	Predicted Recidivism Rate (%)+	95% CI+	95% CII+
-3	0	78	0.0	0.0, 4.0	2.0	1.0, 2.0	0.0, 93.0
-2	4	113	3.5	0.97, 8.8	2.0	2.0, 3.0	0.0, 95.0
-1	14	497	2.8	1.6, 4.7	3.0	3.0, 4.0	0.0, 96.0
0	27	619	4.4	2.9, 6.3	4.0	4.0, 5.0	0.0, 97.0
1	42	735	5.7	4.2, 7.7	6.0	5.0, 6.0	0.0, 98.0
2	51	743	6.9	5.2, 8.9	7.0	7.0, 8.0	0.0, 98.0
3	75	770	9.7	7.7, 12.1	10.0	9.0, 11.0	0.0, 99.0
4	88	729	12.1	9.8, 14.7	13.0	12.0,14.0	0.0, 99.0
5	107	593	18.0	15.0, 21.4	16.0	15.0,18.0	0.0, 99.0
6	93	404	23.0	19.0, 27.4	21.0	19.0,24.0	0.0, 100
7	66	259	25.5	20.2, 31.3	27.0	24.0, 30.0	0.0, 100
8	44	137	32.1	24.4, 40.6	33.0	29.0, 37.0	0.0, 100
9	21	57	36.8	24.5, 50.7	40.0	35.0, 45.0	0.0, 100
10	10	22	45.5	24.4, 67.8	47.0	41.0, 54.0	0.0, 100
11	0	3	0	0.0, 70.8	55.0	50.0, 60.0	0.0, 100
Total	642	5759	11.1	10.4, 12.0	-	-	-

+ Data from Donaldson and Abbott (2011).

There appear to be some advantages when using the Aggregate reference group only to compare to individuals being assessed, such as reducing potential classification error associated with selecting among a variety of reference groups (Wollert, 2010) and the greater variability in the data results in increased reliability (Anastasi & Urbina, 1997). The benefits of using a single actuarial table are outweighed by two limitations. One, the sexual recidivism base rate of 11% may not be accurate as applied to the group representing the evaluee and this would produce an inaccurate estimation of risk. Second, the Aggregate reference group combines data from 23 convenience samples. The membership of this group was not selected using representative sampling methods so the use of a single experience table raises concerns as to the generalizability of the risk data to other groups of sexual offenders. As a stand-alone experience table, it cannot be assumed the risk data from the Aggregate reference group accurately represents the recidivism potential of sexual offenders who are scored on the Static-99R. For these reasons, it seems most prudent for clinicians to consider the experience table from the Aggregate reference group as another option to select when applying the base-rate-matching approach.

Recommendations for Selecting Static-99R Reference Groups. For clinical and forensic practitioners who conduct risk assessments with the Static-99R, it is always preferable to consider an experience table containing score-wise risk estimates

that were developed and validated on a local group of sexual offenders representing the evaluee. When this choice is unavailable, it is recommended that users follow the base rate-matching approach, as described previously. The following paragraph summarizes how to employ this procedure.

The base rate matching approach requires users to determine a sexual recidivism base rate at either the five-year or ten-year follow ups that are most consistent with the group representing the evaluee. The user should be confident that the evaluee is a member of the selected base rate group, which may be found among the local population of sexual offenders or in another representative sample. As stated earlier, the base rate match does not need to be exact, as a variation of a few percentage points would not make a significant difference in the accuracy of the score-wise risk estimates as applied to the group representing the evaluee. Table 6 provides a reference by which to select a Static-99R reference group when applying the base rate matching procedure. Users can obtain the experience table for the Aggregate reference group from Table 5 and the experience tables for the remaining reference groups are located at the following link.[3]

Table 6
Guidelines for Selecting a Static-99R Reference Group Using the Base Rate Matching Approach

Reference group to select	When 5-year sexual recidivism base rate is approximately:	When 10-year sexual recidivism base rate is approximately:
RC	6.0%	N.A.[+]
PTN	9.0%	13.0%
Aggregate	11.0%	N.A.[+]
NRC	15.0%	20.0%
PHRN	21.0%	29.0%

[+] Not available

The following example illustrates how to employ Table 6 to select a Static-99R reference group. A clinician is tasked with determining the sexual recidivism potential of an individual before he is released from prison on parole as the outcome will determine the level of community supervision methods to which the offender will be subjected. The clinician completes the Static-99R and assigns the evaluee a score of four. The clinician is aware that the local corrections agency recently conducted a sexual recidivism study for sex offenders released from prison and found a five-year rate of 5.2%. The clinician determines the evaluee is represented by the membership of the study sample. Comparing the local base rate of 5.2% to the values contained in Table 6 informs the clinician to select the RC reference group. The clinician then accesses the current Static-99R RC experience table from the Static99.org website to obtain the five-year predicted estimate at the score of four.

[3]Available at http://www.static99.org/pdfdocs/detailed_recid_tables_static99r_2009-11-15.pdf

When reporting the score-wise risk estimate, forensic and clinical practitioners are reminded that the base-rate-matching approach assumes similarity in the score-wise likelihood ratios between the selected Static-99R reference group and the group representing the individual. To the extent the score-wise likelihood ratios diverge substantially, the Static-99R sexual recidivism rates may over or under estimate the risk potential of the group representing the evaluee. This limitation should be described in reports or testimony. Readers who are unfamiliar with the interaction between the sexual recidivism base rate and score-wise likelihood ratios are referred to Donaldson and Wollert (2008) for a full explanation.

Acknowledgements

The author would like to thank Theodore S. Donaldson, Ph.D. for his guidance and thoughtful comments that assisted greatly in the development of this article.

Received April 3, 2013; revision submitted May 8, 2013; accepted May 14, 2013

References

Abbott, B. R. (2009). Applicability of the new Static-99 experience tables in sexually violent predator risk assessments. *Sexual Offender Treatment*, 1, 1-24.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2003). *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

American Psychological Association. (2011). *Specialty guidelines for forensic psychology*. Washington, D.C.: American Psychological Association.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct*. Washington, D.C.: American Psychological Association.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing: Seventh Edition*. New Jersey: Prentice-Hall.

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

Beggs, S. M. & Grace, R. C. (2010). Assessment of dynamic risk factors: An independent validation study of the Violence Risk Scale: Sexual Offender Version. *Sexual Abuse: A Journal of Research and Treatment*, 22(2), 234-251. doi: 10.1177/ 1079063210369014.

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

Boccaccini, M. T., Murrie, D. C., Caperton, J. D., & Hawes, S. W. (2009). Field validity of the Static-99 and MnSOST-R among sex offenders evaluated for civil commitment as sexually violent predators. *Psychology, Public Policy, and Law*, 15(4), 278-314. doi: 10.1037/a0017232.

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

California Department of Corrections (2010). *2010 Adult Institutions Outcome Evaluation Report*. Retrieved on November 10, 2010, from: http://www.cdcr.ca.gov/adult_research_branch/research_documents/arb_fy0506_outcome_evaluation_report.pdf

California Sexual Offender Management Board. (2008, January). *An Assessment of Current Management Practices of Adult Sex Offenders in California: Initial Report*. Retrieved March 3, 2008, from www.cdcr.ca.gov/News/docs/SOMB%20Report022108.pdf.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290. doi: 10.1037/1040-3590.6.4.284.

Clopper, C. & Pearson, S. (1934). The use of confidence or fiducial limits illustrated in the case of the Binomial. *Biometrika* , 26, 404-413. doi: 10.1093/biomet/26.4.

Cohen, R. J. & Swerdlik, M. (2001). *Psychological testing and assessment: An introduction to tests and measurement*- Fifth Edition. Columbus, OH: McGraw Hill.

Craissati, J., Bierer, K., & South, R. (2008). *What do sex offenders really get up to? Risk prediction, community failure and "sexually risky behaviours" in a nine-year follow-up study.* Unpublished manuscript.

Cumming, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170-180. doi: 1037/0003-066X.60.2.170.

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), 5-35.

Donaldson, T. S., Abbott, B. R., & Michie, C. (2012). Problems with the Static-99R prediction estimates and confidence intervals. *Open Access Journal of Forensic Psychology*, 4, 1-23.

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. doi: 10.1177/1079063208317734.

Eher, R., Matthes, A., Schilling, F., Hauber-MacLean, T. & Rettenberger, M., (2012). Dynamic Risk Assessment in Sexual Offenders Using Stable-2000 and the Stable-2007: An Investigation of Predictive and Incremental Validity. *Sexual Abuse: A Journal of Research and Treatment*, 24, 5-28. doi: 10.1177/1079063211403164.

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

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

Fernandez, Y. (2008, October). *An examination of the inter-rater reliability of the Static-99 and Stable-2007*. Poster presentation at the 27th Annual Research and Treatment conference of the Association for the Treatment of Sexual Abusers, Atlanta, GA. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7.

Hanson, R. K. & Harris, A. J. R. (2008, February). *Stable-2007 master coding guide*. Unpublished document.

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.static99.org/pdfdocs/hansonharrisscottandhelmus2007.pdf.

Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. *International Journal of Forensic Mental Health*, 11(1), 9-23.

Hanson, R. K., Phenix, A., Helmus, L. (2009, September). *Static-99R and Static-2002R: How to interpret and report in light of recent research*. Pre-Conference workshop at the 28th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Dallas, TX. Retrieved on January 2, 2011, from: http://www.static99.org/pdfdocs/atsapre-conparti_ii_static-99r02r20090930.pdf.

Hanson, R. K. & Thornton, D. (2012, October). *Preselection effects can explain group differences in sexual recidivism base rates in Static-99R validation studies*. Workshop presented at ATSA 31st Annual Research and Treatment Conference, Denver, CO. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7

Hanson, R. K. & Thornton, D. (2000). Improving risk assessment for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, 24(1), 119-136.

Harris, A. J. R., Helmus, L, Hanson, R. K., & Thornton, D. (2008). *Are new norms needed for Static-99?* Workshop presented at the ATSA 27th Annual Research and Treatment Conference on October 23, 2008, Atlanta: GA. Retrieved January 10, 2009: http://www.static99.org/pdfdocs/arenewnormsneededforstatic-99october2008.pdf

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257-272. doi: 10.1007/BF01044769.

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

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

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Assessing risk of older sex offenders: Developing the Static-99R and Static-2002R. *Sexual Abuse: A Journal of Research and Treatment*, 24(1), 64-101. doi: 10.1177/1079063211409951

Knight, R. A. & Thornton, D. (2007). *Evaluating and improving risk assessment schemes for sexual recidivism: A long-term follow-up of convicted sexual offenders*. Washington, D.C.: U.S. Department of Justice. Retrieved July 1, 2008, from: https://www.ncjrs.gov/pdffiles1/nij/grants/217618.pdf.

Langan, P. A., Schmitt, E. L., & Durose, M. R. (2003). Recidivism of sex offenders released from prison in 1994. Washington, D.C.: U.S. Dept of Justice. Available at: http://bjs.gov/content/pub/pdf/rsorp94.pdf.

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

Marshall, W. L. (2006). Diagnostic problems with sexual offenders. In, W.L. Marshall, Y.M. Fernandez, L.E. Marshall, G.A. Serran (Eds.), *Sexual Offender Treatment: Controversial Issues* (pp. 33-43), West Sussex, UK: John Wiley and Sons.

Murrie, D. C., Boccaccini, M. T., Caperton, J., & Rufino, K. (2012). Field Validity of the Psychopathy Checklist–Revised in Sex Offender Risk Assessment. *Psychological Assessment*. 24(2), 524-549. doi: 10.1037/a0026015

Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009) Rater (dis)agreement on risk assessment measures in sexually violent predator

proceedings: evidence of adversarial allegiance in forensic evaluation? Psychology, Public Policy, and the Law, *15,* 19–53. doi: 10.1037/a0014897.

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

Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2013). *VRS-SO norms*. Unpublished document.

Padilla, J. (2006, October). *Communication to Jim McEntee, Public Defender*. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7.

Phenix, A., Helmus, L., & Hanson, R. K. (2012, July). *Static-99R and Static-2002R evaluators' workbook*. Retrieved July 30, 2012, from: http://www.static99.org/pdfdocs/Static-99RandStatic-2002R_EvaluatorsWorkbook2012-07-26.pdf.

Phenix, A., Helmus, L., & Hanson, R. K. (2009, September). *Static-99R Evaluator's Workbook*. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7.

Sachsenmaier, S., Thornton, D., & Olson, G. (2011, November). *Structured risk assessment forensic version (SRA-FV): Score distribution, inter-rater reliability, and margin of error in an SVP population.* Presentation at the 30th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Toronto, Canada. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90-100.

Thornton, D. (2012, May). *Scoring guidelines for need assessment: Structured risk assessment- forensic version – 1.52*. Unpublished paper, available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7

Thornton, D. (2011, December). *Notes for using SRA-FV level of need index*. Unpublished paper, available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7

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*. *The Forum. 22,* Summer.

Thornton, D. & Knight, R. (2009, October). *Using SRA Need domains based on structured judgment to revise relative risk assessments based on Static-2002 and Risk Matrix 2000.* Presentation at the 28th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Dallas, Texas.

Wollert, R. (2010, March). *The use of probability mathematics in sexual violent predator (SVP) evaluations*. Workshop presented at California Department of Mental Health, Seaside, CA, March 9, 2010. Available at: https://www.dropbox.com/sh/oss0b9ut7hx69me/RH1Uzz2qb7

Wollert, R. & Waggoner, J. (2009). Bayesian computations protect sexually violent predator evaluations from the degrading effects of confirmatory bias and illusions of certainty: A reply to Doren and Levenson. *Sexual Offender Treatment*, 1, 1-23.

Wong, S. C. P., Olver, M., Nicholaichuk, T., & Gordon, A. (2003). *Violence risk scale: Sexual Offender Version*. Unpublished scoring manual. Available at: http://shop.psynergy.ca/3-VRS-SO-Manual-VRSSO01.htm.

Zavodny, D. L., DeClue, G., & Cohen, J. (2012, August). *Sexual recidivism base rates relevant to SVP assessments*. Paper presented at the 120th Annual Convention of the American Psychological Association, Orlando, FL.

Appendix A: References for Studies Listed in Table 2

Stable 2007

Routine Corrections

Gillespie, L., & Anderson, D. (2008). Using validated assessments to guide supervision: Partnerships for Sustainability. Department of Public Safety, Hawaii.

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* (Corrections Research User Report No. 2007-05). Ottawa, ON: Public Safety Canada.

Preselected Treatment Need

Bélair, S. (2008). *La prédiction de la récidive sexuelle, violente et criminelle chez les agresseurs de femmes adultes et d'enfants*. Ph.D. thesis, University of Montréal.

Coles, R. L. (2011). *Dynamic risk assessment, personality disorder, and key developmental variables in sexual offenders.* Ph.D. thesis, University of Birmingham.

Cookman, C. L. (2010). The utility of the Assessment of Risk Management of Intellectually Disabled Individuals who Offend – Sexually (ARMIDILO-S) for a community-based service. Ph.D. dissertation. Massachusetts School of Professional Psychology.

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

Nunes, K. L., & Babchishin, K. M. (2012). Construct validity of Stable-2000 and Stable-2007 scores. Sexual Abuse: A Journal of Research and Treatment, 24, 29-45. doi:10.1177/1079063211404921

Willis, G. M., & Johnston, L. (2012). Planning helps: The impact of release planning on subsequent re-entry experiences of child sex offenders. *Journal of Sexual Aggression*, 18, 194-208. doi:10.1080/13552600.2010.506576

Preselected High Risk Need

Sowden, J., & Olver, M. (2012, June). *Examining the relationship between treatment readiness, therapeutic change, and recidivism among sex offenders.* CPA Conference, Halifax, NS, Canada.

SRA-FV

Preselected Treatment Need

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

Trapold, J. (2012). *Identifying sexually violent interests in various types of sex offenders using items on the Structured Risk Assessment: Forensic Version Light.* Ph.D. thesis. Alliant International University.

Preselected High Risk Need

D'Orazio, D., & Thornton, D. (2011). *Levels of criminogenic need in outpatient and SVP populations*. ATSA presentation, Toronto.

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

VRS-SO

Preselected Treatment Need

Beggs, S. M., & Grace, R. C. (2010). Assessment of dynamic risk factors: An independent validation study of the Violence Risk Scale: Sexual Offender Version. *Sexual Abuse: A Journal of Research and Treatment, 22,* 234-251. doi:10.1177/1079063210369014

Kingston, D. A. (2010). The offence progression in sexual offenders: An examination of the self-regulation model of the offence process. Ph.D. thesis. University of Ottawa.

Nunes, K. L., & Cortoni, F. (2008). Assessing treatment change in sexual offenders. Correctional Services Canada.

Preselected High Risk Need

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

Sowden, J., & Olver, M. (2012, June). *Examining the relationship between treatment readiness, therapeutic change, and recidivism among sex offenders.* CPA Conference, Halifax, NS, Canada.