Running Head: SEXUAL RECIDIVISM RATES

What Sexual Recidivism Rates Are Associated with Static-99R and Static-2000R Scores?

R. Karl Hanson[1], David Thornton[2], Leslie-Maaike Helmus[1,3], & Kelly M. Babchishin[1,3]

[1]Public Safety Canada

[2]Sand Ridge Secure Treatment Centre

[3]Carleton University

RDIMS:1110995-V8

In press, *Sexual Abuse: A Journal of Research and Treatment*, 15-JAN-2015

Author Note

The views expressed are those of the authors and not necessarily those of Public Safety Canada or the Wisconsin Department of Health Services. Funding for this project was provided in part by the Social Science and Humanities Research Council of Canada and the Ontario Graduate Scholarship Program.

We would like to thank Amy Phenix, Andrew Harris, and Dennis Doren for their contributions throughout this project. As well, we would like to thank the following researchers for granting us permission to use their data and for being patient with our ongoing questions: Alfred Allan, Tony Beech, Susanne Bengtson, Jacques Bigras, Sasha Boer, Jim Bonta, Sébastien Brouillette-Alarie, Jackie Craissati, Margretta Dwyer, Reinhard Eher, Doug Epperson, Tina Garby, Randolph Grace, Steve Gray, Andy Haag, Leigh Harkins, Steve Johansen, Niklas Långström, Robert Lehmann, Janet Neeley, Terry Nicholaichuk, Jean Proulx, Martin Rettenberger, Rebecca Swinburne Romine, Daryl Ternowski, Robin Wilson, and Annie Yessine. We would also like to thank Julie Blais for rater reliability.

Correspondence concerning this article should be addressed to R. Karl Hanson, Research Division, Community Safety and Countering Crime Branch, Public Safety Canada, 10th floor, 340 Laurier Avenue West, Ottawa, ON, Canada, K1A 0P8. Email: karl.hanson@ps-sp.gc.ca

Abstract

Empirical actuarial risk tools are routinely used to assess the recidivism risk of adult sexual offenders. Compared to other forms of risk assessment, one advantage of actuarial risk tools is that they provide recidivism rate estimates. Previous research, however, suggests that there is considerable variability in the recidivism rates associated with the most commonly used sexual offender risk assessment tools (Static-99/R, Static-2002/R). The current study examined the extent to which the variability in the recidivism rates across 21 Static-99R studies (*n* = 8,805) corresponded to the normative groups proposed by the STATIC development group (routine, treatment, high risk/high need). We found strong evidence that routine (i.e., complete) samples were, on average, less likely to reoffend with a sexual offence than offenders in the high risk/high need samples (i.e., those explicitly preselected on risk-relevant variables external to STATIC scales). The differences between routine/complete and high risk/high need samples, however, were only consistently observed for offenders with low or moderate scores; for offenders with high STATIC scores, the 5-year sexual recidivism rates for these two groups were not meaningfully different. There was only limited evidence to support treatment samples as a distinct sample type; consequently, the use of separate normative tables for treatment samples is not recommended. The current results reinforce the value of regularly updating the norms for empirical actuarial risk tools. Options are discussed on how STATIC scores could be used to inform recidivism rates estimates in applied assessments.

Keywords: sexual offenders, recidivism, prediction, Static-99R, Static-2002R

What Sexual Recidivism Rates Are Associated with Static-99R and Static-2000R Scores?

The assessment of recidivism risk is crucial to the effective management of sexual offenders. In recent years, empirical actuarial risk tools have become routine (Doyle, Ogloff, & Thomas, 2010; Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). These risk tools provide a structured method of combining empirically-derived risk factors (e.g., age, prior criminal history) into total scores, which are then linked to expected recidivism rates (e.g., 30% risk of reoffending after 10 years).

Static-99 (Hanson & Thornton, 2000) is by far the most commonly used risk assessment tool for sexual offenders in the US (Interstate Commission for Adult Offender Supervision, 2007; Jackson & Hess, 2007; McGrath et al., 2010), Canada (McGrath et al., 2010) and Australia (Doyle et al., 2010). It contains 10 items drawn from readily available demographic and criminal history information. The original version of Static-99 provided one - and only one - table linking scores to expected recidivism rates (Appendix 6 in Harris, Phenix, Hanson, & Thornton, 2003). A single table was justified by the authors because there were no meaningful differences in the recidivism base rates across the three samples used to construct that table (total sample size of 1,086). As well, these recidivism rates were relatively stable in seven early replication studies (Doren, 2004). With the exception of the expected value for a score of 4 (which was too high), Doren (2004) did not find significant differences between the expected and observed values for 5-year sexual recidivism rates.

The story got more complicated, however, when the authors re-normed Static-99 based on larger and more recent samples (Harris, Helmus, Hanson, & Thornton, 2008). Using a total of 6,406 sexual offenders from 17 replication samples, the sexual recidivism rates in the newer samples were systematically lower than those included in the original norms. Changing demographics motivated revisions to the original Static-99 age weights, resulting in Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012). A more significant

development, however, was that there was now meaningful variation in the observed recidivism rates across samples and settings (Helmus, Hanson, Thornton, Babchishin, & Harris, 2012). In a meta-analysis of 23 Static-99R recidivism studies (*n* = 8,106), Helmus, Hanson and colleagues found that Static-99R and Static-2002R provided stable estimates of relative risk (i.e., "this group of offenders are more likely to reoffend than this group"). In contrast, the absolute recidivism rates were meaningfully different across samples and settings. As an example, evaluators interested in a decision threshold of 15% after 5 years could use a cut-point that varied from a Static-99R score of 2 (48th percentile) to a cut-score score of 8 (99th percentile), depending on the sample.

The reasons for this variation in base rates are not fully understood. An important preliminary question is the extent to which the variation is systematic. Helmus's (2009) preliminary analyses suggested that relatively small amounts of this variability can be explained by design factors, such as recidivism criteria (charges/convictions), country, or year of release. Helmus's (2009) analyses, however, suggested that a substantial amount of the base rate variability was associated with the methods used by the researchers to select sexual offenders for the recidivism studies. Specifically, the lowest base rates were observed for samples with relatively little pre-selection (routine/complete samples), higher rates were observed for treatment samples, and the highest rates were observed for samples that were explicitly selected to be high risk, such as samples selected for detention because of concern about risk. The developers of the Static-99R were sufficiently persuaded by Helmus's analyses that they recommended against using a single prediction table, and, instead, presented separate recidivism rate tables for these three sample types (Phenix, Helmus, & Hanson, 2012; Thornton, Hanson, & Helmus, 2010).

The implication of this observed variability in base rates is that it became difficult to argue that there should be only one recidivism rate associated with a STATIC[1] score. Instead, evaluators interested in absolute recidivism rates had to decide how variables

[1] STATIC refers collectively to Static-99, Static-99R, Static-2002, and Static-2002R

external to the STATIC measure influence recidivism rates. The most recent norms posted on the Static-99 website (www.static99.org; Phenix et al., 2012) presume that there are meaningful risk clusters, and, consequently, include tables for routine samples, treatment samples, and those explicitly pre-selected on risk-relevant variables other than STATIC scores (the high risk/high need samples). Tables are also provided for a fourth sample ("non-routine"), which combines all the groups that do not meet the definition of the routine/complete sample type (i.e., treatment, high risk/high need, and other preselected samples).

The existence of multiple reference groups introduces a substantial amount of professional judgement into an otherwise highly structured actuarial tool. The history of risk assessment is replete with examples in which actuarial procedures outperform unstructured professional judgement (Ægisdóttir et al., 2006; Meehl, 1954), and sexual offender risk assessment is no exception (Hanson & Morton-Bourgon, 2009). Furthermore, when evaluators are allowed to adjust the results of empirical actuarial risk tools, the adjustments typically degrade predictive accuracy (e.g., Storey, Watt, Jackson, & Hart, 2012; Wormith, Hogg, & Guzzo, 2012; see also reviews by DeClue, 2013 and Hanson & Morton-Bourgon, 2009). Consequently, it is not surprising that there has been considerable debate concerning how evaluators should select reference groups, if at all (Abbott, 2009, 2011, 2013; Campbell & DeClue, 2010, Declue & Zavodny, 2013; Helmus et al., 2009; Sreenivasan, Weinberger, Frances, & Cusworth-Walker, 2010; Thornton et al., 2010; Wilson & Looman, 2010).

Abbott (2013) recommended selecting recidivism rate tables based on an *a priori* decision concerning the expected base rate of the sample to which the offender most closely belongs. Abbott identifies 5 references groups, the 4 identified by Phenix et al. (2012), as well as an aggregate sample (k = 23), having a base rate intermediate between Phenix et al.'s treatment and non-routine groups. In contrast, DeClue (2013; DeClue & Zavodny, 2013) argued that evaluators should only use the table for the routine samples because a) it

is most representative, and b) there is no evidence that choosing between reference groups improves predictive accuracy.

In recent years, the STATIC development group (Hanson & Thornton, 2012; Thornton et al., 2010) has recommended matching offenders to different reference groups and recommended that a primary consideration in the selection of the reference groups should be an assessment of the density of risk factors external to Static-99R/Static-2002R, in particular, the density of criminogenic needs. Their argument concedes that Static-99R and Static-2002R were never intended to measure all relevant risk factors (Hanson & Thornton, 2000; Hanson & Thornton, 2003) and that it is possible to improve the prediction of sexual recidivism by considering information external to Static-99/R (e.g., Allan, Grace, Rutherford, & Hudson, 2007; Lehmann, Goodwill, Gallasch-Nemitz, Biedermann, & Dahle, 2013; McGrath, Lasher, & Cumming, 2012; Olver, Wong, Nicholaichuk, & Gordon, 2007; Thornton, 2002; Thornton & Knight, 2013). This external information could address new content areas (e.g., negative attitudes toward women; crime scene behaviours) or provide better measures of the constructs already targeted by the existing items (e.g., lack of cooperation with supervision). Static-99R and Static-2002R even add incrementally to each other (Babchishin, Hanson, & Helmus, 2012). Consequently, base rates should change whenever offenders are preselected on these external risk factors. By inference, evaluators should use the high risk/high need norms whenever the density of external risk factors is equivalent to the levels found in other high risk/high need samples (Hanson & Thornton, 2012).

Although there is strong evidence that base rates vary, and that external risk factors can add incrementally to Static-99/R, none of the proposed procedures for selecting Static-99R reference groups have been directly evaluated in applied use (DeClue & Zavodny, 2013). Given the inherent risk associated with unstructured professional judgement, there are good reasons for caution. More fundamentally, there has been only limited research on the meaning and validity of the STATIC reference groups. Does it make sense to create

different reference groups at all? If most of the between-group variation is random, then evaluators would do well with a single table using some form of weighted average. If there are meaningful groupings of recidivism base rates, then it is necessary to choose between them. Consequently, there is a need for further research examining the extent to which the variability in the recidivism rates across Static-99R replication studies corresponds to the normative groups that have been proposed by the STATIC development group (Phenix et al., 2012).

Purpose of the Current Study

There are clear a priori reasons to expect differences between routine/complete samples and those preselected to have high risk/high need. There are also reasons to suspect that offenders who have been recommended for treatment have also been preselected on risk-relevant variables. Previous research, for example, has found that sexual offenders selected for treatment have greater histories of sexual criminality than sexual offenders who do not receive treatment (Grady, Edwards, Pettus-Davis, & Abramson, 2013; Duwe & Goldman, 2009; Hanson, Broom, & Stephenson, 2004; Jones, Pelissier, & Klein-Saffran, 2006). The findings concerning selection effects on general criminality have been mixed, with certain treatment samples showing decreased history of general crime (Duwe & Goldman, 2009; Jones, Pelissier, & Klein-Saffran Jones, 2006) and other samples showing an increased history of general crime (Hanson et al., 2004).

The current study aimed to identify patterns in the recidivism rates across samples that could inform decisions concerning the expected recidivism rates that should be associated with STATIC scores. Specifically, we examined the extent to which the variability in the recidivism rates across 21 Static-99R studies (n = 8,805) corresponded to the normative groups proposed by the STATIC development group. The samples in the current study substantially overlapped with those previously used to construct STATIC norms (Helmus, 2009; Phenix et al., 2012). Sample preselection effects were quantified by two indicators. The first (categorical) indicator divided the samples into 3 groups: routine

(complete), treatment (some preselection), and high risk/high need. The second (quantitative) indicator was the average Static-99R or Static-2002R score per sample. If the samples had been preselected on risk-relevant variables, then the variation in average STATIC scores should be correlated with the average density of the latent variables responsible for recidivism risk. Consequently, an incremental effect of average Static-99R scores over individual Static-99R scores would provide evidence of preselection effects based on risk factors external to Static-99R. Such effects are likely given that none of the treatment or high risk/high needs samples were explicitly selected on the basis of the STATIC scales.

In summary, the specific hypotheses examined were as follows: The high risk/high need samples would have high Static-99R and Static-2002R scores, and their recidivism rates would be higher than expected, even after controlling for their high scores. The sexual recidivism base rates will vary based on the degree to which the samples have been preselected on risk-relevant characteristics, with the highest sexual recidivism rates found for the high risk/high needs groups, moderate for the treatment needs groups, and lowest for routine/complete samples (not selected on risk-relevant characteristics). As well, we expected the average STATIC scores for the samples (a proxy for pre-selection) to predict incrementally to the individual STATIC scores.

Method

Samples

This study used data from 8,805 sexual offenders from 21 samples (see Table 1 and Table 2). To be included in the current study, we required information on sexual recidivism and Static-99R total scores, and the sample needed to be one of three types: routine/complete, treatment, or high risk/high need. Samples that did not fit into one of these categories were excluded because the purpose of the current study was to examine the validity of the sample type classification. Furthermore, statistical outliers were excluded within each reference group if, following Hanson and Bussière's (1998) procedure, there was

significant total between-sample variability and the sample was an extreme value that counted for more than half of the variability. Of the 25 eligible datasets, four were excluded. Saum (2007; *n* = 168) was excluded because it had been identified as a statistical outlier (bias in the recidivism rates) in previous analyses (Helmus, Hanson et al., 2012). The Bridgewater Treatment Center sample (Knight & Thornton, 2007) was excluded because it was a much earlier cohort than the other studies (median release year of 1970), and it was an outlier in certain analyses. Previous STATIC norms used a combined Bridgewater sample that included a) offenders who were assessed but not committed and b) offenders who were assessed and committed. When these two subsamples were analyzed separately, the committed sample was identified as a statistical outlier within the high risk/high needs samples in the Static-99R analyses (specifically, it had unusually high base rates and unusually low predictive accuracy compared to the other studies).

Two studies were excluded because they did not obviously fit the classification criteria: one study examined preselected low risk offenders (Cortoni & Nunes, 2007), and another examined only sexual murderers released to the community (Hill et al., 2008). Additionally, 22 cases from Bonta and Yessine (2005) were deleted because the index sex offence charge/conviction was more than two years prior to the current offence. Static-99R and Static-2002R were developed on, and primarily intended for, individuals who are currently (or recently) serving a sentence for a sex offence. Two of the samples had not previously been used for STATIC norms (Hanson, Hanson, Lunetta, Phenix, Neeley, & Epperson, 2014; Lehmann et al., 2013) and one was substantially updated (Hanson, Helmus, & Harris, 2014). For additional information, refer to Helmus (2009) and the original studies. Because we used the raw datasets from these studies, it was possible to make some alterations in variable definitions (e.g., use a broader recidivism definition of charges instead of convictions), include unpublished updated information (where available), and conduct thorough data cleaning; consequently, description information for the studies in Tables 1 and 2 may deviate slightly from other reports of these samples.

Samples were from Canada (*k* = 9), the United States (*k* = 5), the United Kingdom (*k* = 2), and one each from Austria, Denmark, Germany, New Zealand, and Sweden. The average age at release was 40 years old (*SD* = 12; range from 18-86). Offenders were released between 1976 and 2009 (*Mdn* = 1997). All samples used official criminal records to measure recidivism, with half using charges as the recidivism criteria and half using convictions. Note that either definition underestimates the true rate of recidivism due to underreporting (e.g., Dobash & Dobash, 1995).

Table 2 presents average Static-99R/2002R scores per sample (for Static-99R, overall *M* = 2.6, *SD* = 2.6; for Static-2002R, overall *M* = 4.0, *SD* = 2.4). Offenders were followed up for an average of 8.5 years (*SD* = 4.8). The observed sexual recidivism rate for all cases was 11.6%, with a 5-year rate of 9.8%. The rate of any recidivism (including sexual and violent offences, as well as nonviolent offences) was 40.9% overall, with a 39.0% recidivism rate after 5 years. Note that these recidivism rates do not control for STATIC scores.

Sample Type

Each sample was classified according to the extent to which they were preselected on risk-relevant variables: samples with no obvious preselection (routine/complete), treatment samples, and high risk/high needs samples. The reliability of the three sample-type classification (developed by consensus among the authors of this study) was examined by providing study information and the classification criteria to a graduate student in forensic psychology who was otherwise unconnected with this project and who had no particular background in sexual offender research. The independent rater agreed on 19 of the 20 studies examined in the reliability study (Kappa = .92). Brief descriptions of the sample-type definitions are included below. The full classification guidelines provided to the independent rater are available upon request.

Routine/complete samples. Routine/complete correctional samples are a relatively random (i.e., unselected, complete) sample from a correctional system (not just

from one security level, institution, or treatment program). Some offenders in these samples would have been subsequently screened for treatment or other special measures (e.g., psychiatric admission or exceptional measures related to dangerousness), but these samples represent the full population of all offenders prior to any preselection processes. This group represents a hypothetical average of all sex offenders. Generally, the only preselection that can have occurred is based on sentence type (e.g., all offenders with a jail sentence) or methodological factors (e.g., minimum follow-up needed). Any preselection based on risk-relevant factors would make the sample non-routine.

Preselected for treatment needs. This group includes samples of offenders preselected on the basis of treatment needs. In other words, through some formal or informal process, some offenders are judged by somebody as requiring treatment interventions specific to sex offending. The quality of the treatment program and the quality of the offender's participation in and completion of the program is not a consideration in the definition of this group. Treatment samples in which all offenders in the correctional system are referred for sex offender treatment are not included (no preselection has occurred).

Preselected as high risk/needs. This category consists of samples of offenders preselected based on a perceived high level of risk and/or need. These samples typically include offenders considered for some infrequent measure/intervention/sanction typically reserved for the highest risk cases (e.g., detention until warrant expiry, indefinite sentence). This measure/intervention could involve treatment; samples from high-intensity treatment programs reserved for a small subset of offenders (ideally, <20%) and assigned on the basis of perceived high level of risk and/or needs are classified in this category.

Measures

Static-99R (Hanson & Thornton, 2000; Helmus, Thornton et al., 2012). Static-99R is an empirically derived actuarial risk assessment tool designed to predict sexual recidivism in adult male sex offenders (see also www.static99.org). It has ten items, and the total score (ranging from -3 to 12) can be used to place offenders in one of four relative

risk categories: low (-3 to 1), moderate-low (2 to 3), moderate-high (4 to 5), and high (6+). The Static-99R items are identical to Static-99 with the exception of updated age weights.

Static-2002R (Hanson & Thornton, 2003; Helmus, Thornton et al., 2012). Similar to Static-99R, Static-2002R is an empirical actuarial risk assessment tool for adult male sex offenders (see also www.static99.org). It has 14 items grouped into 5 main subscales: age at release, persistence of sex offending, sexual deviance, relationship to victims, and general criminality. The total score (ranging from -2 to 13) can be used to place offenders in one of five risk categories: low (-2 to 2), low-moderate (3 to 4), moderate (5 to 6), moderate-high (7 to 8), and high (9+). The items in Static-2002R are identical to Static-2002 with the exception of updated age weights.

Overview of Analyses

In order to summarize and compare results across samples, we conducted meta-analyses of means, proportions, and logistic regression coefficients, as well as moderator analyses (examining sample type). Incremental validity was tested using Cox regression analyses. Analyses were conducted using version 17 of SPSS, with the exception of the random-effects moderator analyses, which were conducted using Comprehensive Meta-Analysis. All data analyses were conducted independently by two of the authors. Disagreements typically involved copying and rounding errors and were easily resolved once identified.

Aggregation of findings. Both random-effects and fixed-effect meta-analyses were conducted using the procedures described by Borenstein, Hedges, Higgins, and Rothstein (2009). Conceptually, random-effects meta-analyses are preferred, as they allow generalization to the population of which the samples are presumed to be a part. In practice, however, random-effects analyses require estimating a between-study variability term (tau-square; τ^2), which is unstable when the number of studies is small (< 30, Schulze, 2007). When there is true variability, fixed-effects analyses are too liberal, and

when there is true homogeneity, random-effects analyses are too conservative (Overton, 1998).

The point estimates (means, proportions) for the sample-type comparisons were aggregated using random-effects meta-analyses. The parameters used to create recidivism rate estimates were aggregated using fixed-effect meta-analysis. Readers should note that this contrasts with the random-effects meta-analysis used to create the 2012 STATIC norms (Phenix et al., 2012). Our decision to privilege the fixed-effects analyses was primarily based on the difficulty of estimating τ^2 with small numbers of studies (*k* ranged from 2 to 10, with a median of 4 studies). Although random-effects aim to provide estimates for the population (of all possible STATIC recidivism studies), such a generalization cannot be done with confidence without a much larger sample of studies (30+).

Means. Means (STATIC total scores, age) were weighted by the inverse of the variance (i.e., the squared standard error of the mean, [SD^2/n]).

Proportions. Given the well-known problems with estimating variances of raw proportions (Cohen, 1988; Eisenhart, 1947), a variance stabilization transformation was used prior to weighting observed 5-year recidivism estimates by the inverse of the variance. Specifically, we used the most common variance stabilization transformation, the arcsine transformation (*Ă*) that is defined as $\breve{A} = 2\arcsin\sqrt{p}$, where *p* represents the proportion of sex offenders who reoffended within a 5-year follow-up. Unlike raw proportions, the variance of *Ă* depends only on the sample size (specifically, 1/*n*) and not the size of the proportion. All results were reported as proportions, however, because *Ă* in its original units (radians) is not easily interpreted.

Recidivism controlling for STATIC scores. The adjusted recidivism rates after controlling for STATIC scores were estimated using the intercept (B_0) from logistic regression. The intercept estimates the recidivism rate (as a logit) for offenders with a score of 0. Re-centering the scales produces B_0s that examine predicted base rates for any score.

We used B_0 coefficients centered on 2 for the Static-99R and centered on 3 for the Static-2002R, as both these scores represent the median value in the population of adjudicated sexual offenders (Hanson, Lloyd, Helmus, & Thornton, 2012) and, therefore, can be considered to describe offenders in the middle of the risk distribution. For ease of interpretation, the B_0 logits and their confidence intervals were transformed back into probabilities (p), where $p = 1/(1 + e^{-LOGIT})$. Logistic regression was also used to estimate sexual recidivism rates after fixed follow-up times of 5 years and 10 years. The between-group analyses, however, focused on the 5-year rates because there were only two studies with sufficient data to compute 10-year recidivism rates for Static-99R (Bengston, 2008; Nicholaichuk, 2001). Sufficient data were defined as having at least 10 cases with fixed 10-year follow-up and the logistic regression estimates not mis-specified (lack of algorithm convergence).

Moderator analyses. Tests for differences between the sample types used the $Q_{Between}$ test from random-effects analyses (Borenstein et al., 2009). This test partitions the overall variability across studies into variability within each level of the moderator (i.e., within each sample type) and variability between levels (i.e., variability due to the moderator).

Incremental effects. To test the incremental effects of sample type and mean STATIC score over the original STATIC scales, Cox regression analyses were used (Allison, 1984). Cox regression calculates hazard ratios, or the extent to which the probability of failure (i.e., recidivism) varies as a function of predictor variables. From these analyses, we reported the exponent of the regression coefficient (*Exp(B)*), which is a hazard ratio indicating the variable's relationship with recidivism. For example, if the hazard ratio for Static-99R was 1.20, then each one-score increase in the scale is associated with an increase in recidivism of 1.20, or 20%, averaged across the follow-up period. The Wald statistic tests the significance of the hazard ratio. When multiple predictors are entered into a model, the *Exp(B)* and the Wald test refer to the statistics after controlling for the effect of

the other predictors. To assess differences in the model when one or more predictors are added in a new step, the chi-square test of the change in the model was examined.

Results

A meta-analysis of the 21 studies found significant variability in the base rate associated with the median Static-99R score (i.e., a score of 2; $Q = 62.2$, $p < .001$, $I^2 = 67.8$, see Table 3). Similarly, in the seven Static-2002R studies, there was significant variability in the base rate associated with the median Static-2002R score (i.e., a score of 3; $Q = 29.9$, $p < .001$, $I^2 = 79.9$). There was also significant variability in the relative risk parameter (*B1*) for Static-2002R ($p = .012$). The between sample variability in the relative risk parameter for Static-99R was not statistically significant, but close ($Q = 28.8$, $p = .092$). Using Cox regression, the relative risk parameters were significantly different between the routine/complete and high risk/high needs sample for both Static-99R (χ^2 [1] = 21.26, $p < .001$) and Static-2002R (χ^2 [1] = 11.89, $p < .001$).

Compared to the routine/complete samples and treatment needs samples, the high risk/high need samples had higher Static-99R and Static-2002R scores (see Table 4). There were, however, no overall differences in the Static-99R or Static-2002R scores between the routine/complete samples and the treatment needs samples.

As expected, the sexual recidivism rates were significantly higher in the high risk/high need samples than in the other groups, and they remained significantly higher than the routine/complete samples after controlling for Static-99R total scores (by a factor of 2.2). The same pattern was observed for Static-2002R (2.4 times higher for high risk/high need compared to routine/complete samples), but the difference from the other types of samples was not statistically significant. Note that the adjusted rates presented in Table 4 are centered on the median values of the scales (i.e., a score of "2" for Static-99R and a score of "3" for Static-2002R; see Hanson, Lloyd, et al., 2012). Contrary to expectation, there were no significant differences in the raw or adjusted recidivism rates

between the routine/complete and the treatment samples for either Static-99R or Static-2002R.

In terms of actual distributions of scores, for Static-99R, both the routine/complete and treatment needs samples had scores ranging from -3 to 11, with 11% of the routine/complete samples and 9% of the treatment needs samples having scores of 6+ (i.e., in the high risk category). In contrast, the Static-99R scores for the high risk/needs samples ranged from -3 to 12, with 35% of offenders scoring 6 or above. For Static-2002R, routine/complete and treatment need samples both had scores ranging from -2 to 11, with 12% and 18% of scores being 7+, respectively (7+ represents the closest percentile match to a Static-99R score of 6+; Hanson, Lloyd et al., 2012). In the high risk/need samples, Static-2002R scores ranged from -1 to 12, with 27% of offenders scoring 7 or above. Conversely, the proportion of offenders with low scores on the Static-99R (1 or less) was 36%, 42%, and 13% for routine, treatment, and high risk/high need samples, respectively. A similar pattern was observed for Static-2002R, with 28%, 40% and 15% of routine, treatment, and high risk/high need samples, respectively, having scores of 2 or less.

Preselection effects were also examined using Cox regression survival analysis in the aggregated dataset (n = 8,805, with 1,021 sexual recidivists). The average Static-99R scores of the samples had an incremental effect over individual Static-99R scores (see Table 5, Step 2a), with a hazard ratio of 1.12 (95% CI of 1.05 to 1.19). The treatment needs and high risk/high needs groups had significantly higher sexual recidivism rates than the routine/complete samples after controlling for Static-99R scores (see Table 5; Step 2b). Specifically, the hazard ratio of 1.25 for the treatment group indicated that the sexual recidivism rates were approximately 25% higher in the treatment compared to the routine/complete samples after controlling for Static-99R; for the high risk/high needs group, the hazard ratio was 1.58.

When both sample type (routine/complete, treatment, high risk/high needs) and average Static-99R scores were considered simultaneously, the sample type effects

remained significant, but the effect for average Static-99R scores was no longer significant (see Table 5, Step 3). In other words, sample type was a better predictor of the recidivism rate variability than average Static-99R score, after accounting for individual Static-99R score. The same pattern of results was observed for Static-2002R (see Table 6); however, the Static-2002R analyses included only one treatment sample, and this group was not significantly different from the routine/complete samples.

The relationship between Static-99R adjusted base rates for a score of 2 and average Static-99R scores is presented in Figure 1. The high risk/high needs samples were clustered in the upper right corner, with high Static-99R scores as well as high sexual recidivism rates after adjusting for Static-99R scores. In contrast, there was no visible separation between the routine/complete and treatment samples.

The expected recidivism rates associated with Static-99R and Static-2002R scores were calculated using averaged logistic regression parameters. Specifically, logistic regression equations were calculated for each sample based on sexual recidivism after 5-year and 10-year follow-up periods, with the predictor scores centered on their median value (2 for Static-99R; 3 for Static-2002R). In order to calculate the logistic regression parameter, each sample must have at least 10 cases, and not be mis-specified in logistic regression analyses. This resulted in eliminating the following studies from the Static-99R 10-year recidivism estimates: Wilson et al. (2007A & B; 1 recidivist – mis-specified model); Bonta and Yessine (2005; n = 3), and Hanson et al. (2012, n = 3). The intercept (*B0*) and discrimination parameters (*B1*) were then aggregated using fixed effect meta-analysis. Following the recommendations of Vergouwe, Steyerberg, Eijkemans, and Habbeman (2005) for producing stable logistic regression models, estimates were presented only for analyses with roughly 100 recidivists or more.

Each parameter was aggregated separately for the high risk/high needs samples (k = 5), and for the routine/complete samples (k = 10; see Table 6). Unlike previously released STATIC norms (Phenix et al., 2012), the regression parameters in Table 6 allowed

for different base rates and different discrimination (slope) parameters for the two reference groups. Once the samples were divided in routine/complete and high risk/high need samples, the between-study variability was substantially reduced and was nonsignificant for 8 of the 10 parameters. However, there was still significant variability in the base rates (*B0*) for the routine/complete samples used to create the 5-year recidivism rate estimates for both Static-99R (Q = 34.2, df = 9, I^2 = 73.7%, p < .001) and Static-2002R (Q = 12.3, df = 3, I^2 = 75.6%, p = .006).

As expected, the 5-year base rates were higher in the high risk/high need samples compared to the routine/complete samples for both Static-99R (11.3 versus 5.6) and Static-2002R (13.3 versus 6.8). These recidivism rates are the parameters in Table 7 transformed from logits into percentages using the formula described in Appendix A. There were also meaningful differences in discrimination (*B1*s) between the high risk/high need and routine/complete samples, with the strongest discrimination being found for the routine/complete samples. For the 5-year estimates, the odds ratio associated with a one point increase in Static-99R score was 1.45 for the routine/complete samples and 1.28 for the high risk/high needs samples ($e^{[0.368494]}$ = 1.45; $e^{[0.25091]}$ = 1.28; from Table 7). The same pattern was observed for Static-2002R (1.48 versus 1.24). Using the procedures described in Appendix A, the logistic regression coefficients were used to compute the expected 5-year sexual recidivism rates displayed in Figure 2 (and provided in table form in Appendix B). The expected recidivism rates for the high risk/high needs groups were noticeably higher than the rates for the routine/complete samples throughout most of the range of scores; however, the differences between the groups disappeared for the very highest scores (the lines crossed at a score of 8 for both measures). Recidivism rate tables were not constructed for the treatment groups samples because the rates were similar to the rates for the routine/complete samples, and the population from which these samples would be drawn was difficult to define.

Discussion

We found strong evidence that the base rate variation across samples was not random and that some of the variation can be explained by systematic differences in sample characteristics. Specifically, samples that were explicitly selected as high risk/high need had the highest risk scores and the highest recidivism rates. Furthermore, the average STATIC scores for each sample added incrementally to the prediction of sexual recidivism after controlling for the STATIC scores of the individual offenders. Such a pattern of results is parsimoniously explained by assuming that the high risk/high needs samples were selected on variables correlated with Static-99R and Static-2002R scores and on risk-relevant variables external to these particular risk measures.

Assuming that each increase in Static-99R score corresponds to a 40% increase in relative risk (risk ratio of 1.40, or 0.336 in logit units; Hanson, Babchishin, Helmus, & Thornton, 2013), the difference in the base rates between the routine/complete samples and high risk/high needs samples was approximately 2 Static-99R units at the middle of the risk distribution (-2.06 – [-2.83] = 0.77/[0.336] = 2.3 for Static-99R; -1.88 – [-2.62] = 0.74/[0.336] = 2.2 for Static-2002R, from Table 6). In other words, a sex offender with a Static-99R score of 2 from a high risk/high needs sample would have an expected recidivism rate similar to the recidivism rate expected for a sex offender from a routine/complete sample with a score of 4. Note, however, that the difference in risk between the routine/complete and high risk/high needs groups decreased as the scores increased, such that the difference was small for a Static-99R score of 6 (less than 1 STATIC unit) and absent (no difference) for a score of 8.

Contrary to expectation, we did not find substantial differences between treatment and routine/complete samples. The average risk scale scores were not significantly different across these groups. In the Cox regression analyses, offenders in the treatment samples reoffended more quickly than the routine/complete samples, but the absolute difference was small (9.2% versus 7.6% after 5 years), and the difference was not significant in the more conservative random-effects meta-analysis. Consequently, there is some evidence that

treatments samples are not routine/complete, but it is not clear that treatment samples are sufficiently distinct to justify their own category. Furthermore, it is not clear how treatment samples should be defined, as policy decisions could substantially influence who is referred for treatment. In jurisdictions that require all sexual offenders to receive treatment, treatment samples will resemble routine/complete samples. In jurisdictions that provide treatment only to the neediest offenders, however, treatment samples will resemble high risk/high need samples. In contrast, it is possible to set *a priori* definitions for routine/complete (all offenders) and high risk/high need samples (e.g., riskiest 20%) that are independent of shifting policy decisions.

The current results reinforce the value of regularly updating empirically-derived recidivism estimates. Even though the current samples substantially overlapped with the samples used to create the 2012 STATIC norms, the results were meaningfully different in certain respects. Specifically, the empirical justification for the treatment group largely disappeared, and the shape of the recidivism curve was now different for the high risk/high needs samples compared to the routine/complete samples. The sensitivity of the results to relatively small changes in sample composition was unexpected and should remind us all (researchers, evaluators, and decision makers) of the need to interpret empirically-derived recidivism rates estimates with caution. Although we believe the current results are credible, they should not be considered final. As more and better recidivism studies are conducted, the estimates should improve, as should their scientific credentials.

In the 2012 STATIC norms, a common (i.e., pooled) slope parameter was used to estimate recidivism rates for all reference groups because the variability in the slopes was no more than would have been expected by chance ($\chi^2 = 20.35$, $p = .44$; *B1* for Static-99R at 5 years; Helmus, Hanson et al., 2012). In the current dataset, the between-study variability for Static-99R slope parameters was similarly not significant in the meta-analysis of the logistic regression parameters; consequently, the current findings did not compel the need for different slope parameters for Static-99R. In the more powerful cox regression

analyses, however, the variability in the slopes was significantly different between the routine/complete samples and the high risk/high needs samples for both Static-99R and Static-2002R. This finding supports Donaldson, Abbott, and Michie's (2012) reanalysis of the previous logistic regression estimates that suggested that the 2012 estimates were poorly centered for the high risk/high needs samples. Consistent with the pattern identified by Donaldson et al. (2012), Helmus and Thornton's (2014) meta-analysis of the individual STATIC items found that they all predicted sexual recidivism but also found that certain items worked better (had better predictive accuracy) in routine/complete samples than in high risk/high needs samples.

Another reason to allow the discrimination parameters to vary was that the resulting recidivism rate curves (Figure 2) had a compelling interpretation. Specifically, for offenders with high STATIC scores, high levels of external risk factors would be expected. It is unlikely that an individual would accumulate such extensive criminal histories without also having substantial risk-relevant needs. Consequently, it would make little difference whether the samples were explicitly selected on risk-relevant characteristics: they would all be expected to be high risk. In contrast, selecting high risk/high need offenders from those with low or moderate STATIC scores should identify individuals who have *unusually* high levels of life problems that increase their likelihood of offending (i.e., criminogenic needs) compared to other offenders with the same STATIC scores.

The consistency in recidivism rates for the high scores returns Static-99R users to a more direct link between a specific score and a single STATIC recidivism rate estimate. However, for the vast majority of offenders (all those with scores less than 7), base rate variability persists. Furthermore, in certain circumstances, the variability could be sufficient that it would influence the ultimate decisions concerning whether offenders exceed thresholds defined in policy or law. The current findings help interpret STATIC base rate variability, but they do not provide an empirically-validated method for addressing it. The following section discusses the pros and cons of possible options for applied practice.

The meaningful, nonrandom variation in recidivism base rates argues against combining all samples into a single normative table. By combining all available samples, the estimated recidivism rates would be biased upward based on the proportion of high risk/high needs samples that happen to be available to researchers. It is important to remember that the routine/complete samples include offenders at all risk levels, including the naturally occurring proportion of high risk/high need offenders. In routine/complete samples, the naturally occurring proportion of high risk/high need offenders would be relatively small (< 20%). When preselected high risk/high need samples are combined with routine samples, the overall proportion of high risk/high need offenders in the combined sample would be artificially increased, thereby creating recidivism rate estimates that are too high. Consequently, we recommend against using norms based on mixed samples that are expected *a priori* to have different recidivism rates.

Following DeClue (2013), another option would be to restrict interpretation to Static-99R and Static-2002R norms for routine/complete samples. Pegging STATIC recidivism rate estimates to routine/complete samples has the conceptual advantage of providing a stable waypoint for evidence-based debate concerning the recidivism rates that *ought* to be associated with specific scores. Using only routine/complete norms in practice also minimizes the problems associated with selecting comparison groups. Furthermore, the routine/complete sample recidivism rates should be plausible estimates for most cases. The problem, of course, is that routine/complete sample norms would underestimate the risk for offenders who have low or moderate STATIC scores but are high risk for other reasons (e.g., first conviction for a sexual offence but multiple paraphilias and frontal lobe damage). The use of the routine/complete sample norms also does not solve the problem of base rate variability. Even for routine/complete samples, the between-study variability was more than expected by chance, suggesting that there is meaningful variation in routine/complete samples that has yet to be explained.

Another option for addressing base rate variability would be to report two rates: the rate for the routine/complete samples and the rate for the high risk/high needs samples. Such an approach avoids the problems associated with selecting reference groups, and, in many cases, the resulting range will completely fall on either side of the decision threshold. Problems arise, however, whenever they do not (i.e., the rates for the different reference groups crosses the decision threshold). In such cases, answers to the referral question require a judgement of which norms (if either) best suit the case at hand. Notice that this decision cannot be avoided by using any particular norms as the default. If evaluators follow DeClue and Zavodny's (2013) recommendation to use routine/complete sample norms in all cases, evaluators are still making a (contestable) decision concerning the appropriate reference group.

Even though plausible decision rules for choosing reference groups could be examined in future studies, it is not clear that this approach is the optimal method for considering the impact of external risk factors. First, using only two tables reduces what is likely to be a continuous distribution of risk into a dichotomy. Furthermore, referencing only routine/complete samples and high risk/high need samples does not provide a conceptual handle for evaluators to identify offenders who are significantly less risky than expected.

If the impact of the external risk factors is known, then it is possible to empirically combine STATIC with external factors into an overall evaluation of risk. Currently, recidivism rate tables are available that combine Static-99R with SOTIPS (McGrath et al., 2012) and that combine STABLE-2007 with Static-99R, Static-2002R, or Risk Matrix-2000 (Helmus & Hanson, 2013; Helmus, Hanson, Babchishin, & Thornton, 2014). Very similar tables are presented for the Static + Dynamic sections of the VRS-SO, which have the added benefit of empirical adjustments based on positive response to treatment (Olver, Beggs Christopherson, Grace, & Wong, 2013). A related approach is to model the incremental effects of external risk factors using logistic regression. For example, Thornton and Knight

(2013) provide the necessary parameters to calculate recidivism rate estimates based on any unique combination of Static-99R and SRA-FV scores.

The use of empirical recidivism tables that combine STATIC measures with external risk measures limits the need for speculation concerning unmeasured risk factors and should improve predictive accuracy. However, empirically combining STATIC scores with other measures has the effect of creating a new actuarial measure, which needs to be evaluated on its own merits. There has been much less research on these new combined measures than there has been for Static-99/R. For both STABLE-2007 and the VRS-SO, the calibration of the risk estimates has been acceptable in new samples (Eher, Matthes, Schilling, Haubner-MacLean, & Rettenberger, 2012; Olver et al., 2013). With only a handful of studies, however, the stability of these estimates remains largely unknown. Just as STATIC base rates vary across samples and settings, it is likely that there is more than a trivial amount of variability in the recidivism rates associated with STATIC + OTHER combinations. This potential variability, however, should be smaller than recidivism estimates produced using only STATIC scores, as more relevant information has been incorporated.

An additional concern is how information derived from group data should be applied to individuals. There are ongoing debates in the philosophical (Courgeau, 2004; Gillies, 2000; Hájek, 2007) and professional literature (Greenland, 1998; Hanson & Howard, 2010; Hart, Michie, & Cook, 2007) concerning how empirically derived probabilities should be related to individuals, if at all. The concept of empirical probabilities gained popularity in the mid-20th century in the fields of epidemiology and medicine and contrasts with the much earlier concept of logical probabilities (e.g., games of chance). The basic approach of empirical probability research is to assign an individual to a reference class based on individual characteristics (e.g., smoker) and life circumstances (e.g., city with high levels of air pollution). Based on such features, researchers empirically estimate the proportion of that class who will have the outcome of interest (e.g., lung cancer). There is a problem, of

course, in defining which of many potential classes the individual belongs (e.g., smoker, good diet, regular exercise, no family history; Hájek, 2007). There is general agreement that the class should consider all risk-relevant factors (both individual and situational), but, by creating increasingly specific classes, the sample size decreases to uninformative levels. The developers of the STATIC scales have addressed this problem by adopting a cumulative stochastic model, in which each point adds equally and incrementally to the overall risk (Hanson et al., 2013), an approach that makes the (testable) assumption that there is relatively little information provided by interactions between predictors.

Once group-based empirical probabilities have been identified, there are several different ways in which they can be interpreted (Gilles, 2000; Greenland, 1998). We will mention three: the frequentist, the subjectivist, and the propensity interpretations. For frequentists, empirical probabilities are defined as mathematical abstractions (i.e., limits of an infinite series) and have no substantive meaning when applied to an individual case (Greenland, 1998). Although the limits of infinite series can never be observed, they can be estimated by observing the *same* experiment repeated a large number of times. For subjectivists, probabilities are not characteristics of the world; instead, they index the certainty of belief in certain states of affairs (e.g., I am 50% confident that Mr. X will reoffend; Greenland, 1998). Within the propensity interpretation, in contrast, the likelihood of actualizing certain outcomes is a property of the individual (Gillies, 2000; Popper, 1959) not an abstract limit or a subjective belief. Of these three approaches, we believe that the propensity interpretation provides the meaning most closely aligned with applied risk assessment and legal decision making.

Although not the focus of the current study, our results have important implications for the evaluation of sexual offender treatment programs. The most common design for sexual offender treatment outcome studies are cohort and incidental assignment designs, which crucially depend on the assumption that the expected recidivism rates are stable across samples and settings. This assumption is untenable. Consequently, researchers need

to be particularly vigilant to the bias introduced by comparing treated samples to samples with different base rates. Even after controlling for many well-established risk factors (i.e., Static-99R scores), equivalence in risk cannot be assumed. For the record, we do not interpret the marginally higher recidivism rates among treatment samples compared to routine/complete samples as evidence that treatment makes offenders more likely to reoffend. Comparison studies typically find that sexual offenders who complete treatment are lower risk than equivalent groups that do not receive treatment (Hanson, Bourgon, Helmus, & Hodgson, 2009; Lösel & Schmucker, 2005). Instead, we interpret the current results as supporting the need for strong research designs (Långström et al., 2013).

Limitations

Meta-analyses are limited by both the total sample size and the number of distinct samples. In particular, it is difficult to obtain stable estimates of between-study variability and to detect moderator effects given the modest number of samples in the current study. The final recidivism estimates were based on fixed-effect analyses, which is conceptually limited in its generalizability (i.e., it describes only the studies included in the analyses). Random-effects analyses would have been preferable (allowing generalization to the population of sex offender samples), but there was an insufficient number of studies to reliably calculate the between-study variability necessary for random-effects analyses. Notably, however, when between-study variability is small, the differences between fixed-effect and random-effects results are very similar (which was the case for most parameters used in developing the Static-99R estimates).

Another significant limitation was that the original sample-type classification was assigned *post hoc* on existing samples, which means that the classification was not blind to the between-sample differences in recidivism rates (Helmus, 2009). Additionally, there was little information concerning the actual procedures used to create the samples examined in this meta-analysis, so considerable extrapolation was required. Although STATIC scores were not explicitly used to select any of the high risk/high need samples (Static-99 was not

yet invented at that time), it is likely that variables very similar to those included in the STATIC measures were used to select participants in these samples (e.g., young age, prior sexual offences, prior history of violence). These data limitations make it difficult to derive precise or consistent definitions of preselected high risk/need samples for applied practice. Other limitations pertain to restrictions in the comprehensiveness of the data included. There was insufficient data to produce reliable 10-year recidivism estimates for routine/complete samples. The samples included offenders who varied in age at release from 18 to 86; less is known about offenders who are older than 86 at release. The preselected high risk/need samples did not include any data from the United States (however, previous analyses have not demonstrated compelling jurisdictional differences; Helmus, 2009). Additionally, logistic regression estimates may be unstable when extrapolating to Static-99R/2002R scores with no or few cases; consequently, recidivism estimates were not reported for scores where there were less than 10 cases available. For example, there are no estimates provided for Static-99R scores of 11 or 12 in routine/complete samples. Based on our data, such scores are likely to occur in less than 0.07% of cases in routine/complete samples. When such situations do arise, however, it is possible that the closest available recidivism estimate (for a score of 10) may underrepresent the offender's risk. Such underrepresentation is likely to be small, given that the logistic distribution is such that relative increases in recidivism per Static-99R score decrease for the highest scores (e.g., Hanson et al., 2013).

Recommendations for Practice

When Static-99R or Static-2002R is used as a stand-alone measure, our general recommendation is to report only the recidivism rates for routine/complete samples (see Appendix B), as these are the most representative of the population of all convicted sex offenders, and these samples are easiest to define conceptually. Evaluators could acknowledge base rate variability by including a caution that the risk could be higher or lower based on factors not measured by Static-99R/Static-2002R. It is important to

remember that routine/complete samples include all offenders, including those who meet the criteria for high risk/high need samples and treatment samples. In the routine/complete samples, however, special groups are represented in the proportion that they naturally occur and are not overrepresented as they are in the preselected samples.

Although the recidivism rates for the routine/complete should be the default choice, we believe that the recidivism rates for the high risk/high need samples should be used when there is a strong, case-specific justification. A primary consideration in this justification should be density of risk factors external to the STATIC measure, such as scores on the Stable-2007. The current results indicate that certain classes of offenders reoffend at rates higher than expected for routine/complete samples. Previous research indicates that there are a number of factors that add incrementally over STATIC scores (Lehmann et al., 2013; McGrath et al., 2012; Thornton, 2002; Thornton & Knight, 2013) and that these factors should be considered in comprehensive risk assessments. Although we recommend that comprehensive evaluations consider both reference groups, the ability of evaluators to improve accuracy by choosing reference groups has yet to be empirically tested.

Contrary to previous user guidance (Phenix et al., 2012), we recommend against using recidivism rate estimates based on samples preselected for treatment. The current study found only limited empirical support for treatment samples as a distinct sample type. More fundamentally, it is difficult to define the population that treatment samples represent because of the wide variation in the policies for sexual offender treatment in different jurisdictions.

In some earlier presentations and training materials, members of the STATIC development group recommended choosing reference groups based on the administrative procedures that preceded the offender being assessed (Hanson & Thornton, 2008). Although de-emphasised, this is still an option in recent user guidance (Phenix et al., 2012). It has, however, proved difficult in practice. Previous case-management decisions are

clearly relevant for comprehensive risk assessments, but evaluators need to critically examine prior assessments when making their own decisions about offenders' likelihood of recidivism.

We recommend that local STATIC norms be used only when they have greater scientific credibility than the available aggregated norms. Local norms can account for the unique cultural and social features of a specific jurisdiction, but they are difficult to produce with confidence. In agreement with Vergouwe and colleagues (2005), we recommend 100 recidivists for stable logistic regression estimates. As such, local norms should be based on approximately 1,000 cases, assuming overall base rates of 10%. In comparison, the routine/complete sample norms presented in Appendix B are based on 358 sexual recidivists (total n = 4,325) for Static-99R, and 217 sexual recidivists (total n = 1,964) for Static-2002R.

Given the instability of absolute recidivism rates, evaluators should also consider whether it is necessary to report absolute recidivism rates at all. For many decisions, relative risk ranking is sufficient (e.g., provide high intensity supervision to the top 20% of sex offenders). It would also be possible to report an overall recidivism base rate and relative risk parameters (e.g., the 5-year sexual recidivism base is expected to be between 5% and 15% [or use a credible local base rate] and based on Mr. X's Static-99R score, his expected recidivism rate is half/twice/the same as offenders in the middle of the risk distribution; see Hanson et al., 2013).

Use of relative risk metrics (percentiles, rate ratios) does not, however, solve the problem of base rate variability. Even when STATIC scores consistently rank offenders across different samples, offenders' risk levels can still be higher or lower than the group averages based on risk factors not measured by the STATIC scale. The observed consistency in relative risk metrics across samples (Hanson et al., 2012; Hanson et al., 2013; Helmus, Hanson et al., 2012) simply indicates that the items included in the STATIC scales are risk-relevant for diverse samples of sexual offenders. Nevertheless, percentiles

and risk ratios are plausible metrics for communicating the risk information provided by these particular sets of items. In these forms of risk communication, the focus is not on obtaining precise estimates of recidivism rates; instead, the focus is on positioning the offender relative to others.

Given that consumers of risk reports greatly prefer nominal risk categories (low, moderate, high) to numbers of any form (Grann & Pallvik, 2002; Heilbrun et al., 2004), it is important that numeric risk indicators (relative risk, absolute recidivism rates) are translated into terms that are understood and accepted in the practice setting.

Concluding Comment

The use of any norms for any scale requires a professional judgement concerning their validity. There is widespread agreement that inferences about individual recidivism rates should be based on sound scientific procedures and that the recidivism rate estimates should be based on samples that most closely resemble the case at hand. In routine correctional practice, decisions concerning the credibility of risk assessment procedures are typically the responsibility of central administrators who set standardized policy concerning the use and interpretation of risk tools. In forensic psychology and psychiatry, however, the expert witness typically has the dual responsibility of providing opinions concerning the case at hand and defending the broad scientific background on which these opinions are based (Faigman, Monahan, & Slobogin, 2013). Consequently, even if an expert witness uses a mechanical risk assessment tool, the professional opinions proffered by the expert should not be mechanical. Instead, expert opinion should be based on a carefully reasoned judgement concerning the appropriateness of this specific risk assessment procedure, for this specific offender, for this specific purpose.

References

Studies with an asterisk were included in the meta-analysis.

Abbott, B. R. (2009). Applicability of the new Static-99 experience tables in sexually violent predator risk assessments. *Sexual Offender Treatment, 4*(1), 1-24.

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

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.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341-382. doi:10.1177/0011000005285875

*Allan, M., Grace, R. C., Rutherford, B., & Hudson, S. M. (2007). Psychometric assessment of dynamic risk factors for child molesters. *Sexual Abuse: A Journal of Research and Treatment, 19,* 347-367. doi:10.1007/s11194-007-9052-5

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data* (No. 46). Thousand Oaks, CA: Sage.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012). Even highly correlated measures can add incrementally to predicting recidivism among sex offenders. *Assessment, 19*, 442-461. doi:10.1177/1073191112458312

*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 & Comparative Criminology, 47, 422-438. doi:10.1177/0306624X03253850

*Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime & Law, 14,* 85-106. doi:10.1080/10683160701483104

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

*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.

*Bonta, J., & Yessine, A. K. (2005). [Recidivism data for 124 released sexual offenders from the offenders identified in *The National Flagging System: Identifying and responding to high-risk, violent offenders* (User Report 2005-04). Ottawa: Public Safety and Emergency Preparedness Canada]. Unpublished raw data.

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

*Brouillette-Alarie, S., & Proulx, J. (2008, October). *Predictive and convergent validity of phallometric assessment in relation to sexual recidivism risk.* Poster presented at the annual conference for the Association for the Treatment of Sexual Abusers, Atlanta, GA.

Campbell, T. W., & DeClue, G. (2010). Maximizing predictive accuracy in sexually violent predator evaluations. *Open Access Journal of Forensic Psychology, 2*, 148-232.

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

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

Courgeau, D. (2004). Probabilités, demographie et sciences sociales [Probabilities, demography and the social sciences]. *Mathematics and Social Sciences, 42*(167), 27-50.

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

DeClue, G. (2013). Years of predicting dangerously. *Open Access Journal of Forensic Psychology, 5,* 16-28.

DeClue, G., & Zavodny, D. L. (2013). Forensic use of the Static-99R: Part 3. Choosing a comparison group. *Open Access Journal of Forensic Psychology, 5,* 151-182.

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

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

Doren, D. M. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse: A Journal of Research and Treatment, 16,* 25-36. doi:10.1023/B:SEBU.0000006282.36584.a0

Doyle, D. J., Ogloff, J. R. P., & Thomas, S. D. M. (2010). An analysis of dangerous sexual offender assessment reports: Recommendations for best practice. *Psychiatry, Psychology and Law, 18*, 537-556. doi:10.1080/13218719.2010.499159

Duwe, G., & Goldman, R.A. (2009). The impact of prison-based treatment on sex offender recidivism. *Sexual Abuse: A Journal of Research and Treatment, 21,* 279-307. doi:10.1177/1079063209338490

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

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

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

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

Faigman, D. L., Monahan, J., & Slobogin, C. (2013). Group to individual (G2i) inference in scientific expert testimony. *University of Chicago Law Review*, 81, No. 2, 2014; Virginia Public Law and Legal Theory Research Paper No. 2013-34; Vanderbilt Public Law Research Paper No. 13-47. doi:10.2139/ssrn.2298909

Gillies, D. (2000). Varieties of propensity. *British Journal of the Philosophy of Science, 51,* 807-835.

Grady, M.D., Edwards, D., Pettus-Davis, C., & Abramson, J. (2013). Does volunteering for sex offender treatment matter? Using propensity scores analysis to understand the effects of volunteerism and treatment on recidivism. *Sexual Abuse: A Journal of Research and Treatment, 25,* 319-346. doi:10.1177/1079063212459085

Grann, M., & Pallvik, A. (2002). An empirical investigation of written risk communication in forensic psychiatric evaluations. *Psychology, Crime & Law, 8,* 113-130.

Greenland, S. (1998). Probability logic and probabilistic induction. *Epidemiology, 9,* 322-332.

*Haag, A. M. (2005). [Recidivism data from 198 offenders detained until their warrant expiry date. From: Do psychological interventions impact on actuarial measures: An analysis of the predictive validity of the Static-99 and Static-2002 on a re-conviction measure of sexual recidivism. *Dissertations Abstracts International, 66* (08), 4531B. (UMI No. NR05662)]. Unpublished raw data.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese, 156,* 563-585. doi:10.1007/s11229-006-9138-5

Hanson, R. K., Babchishin, K. M. , Helmus, L., & Thornton, D. (2013). Quantifying the relative risk of sex offenders: Risk ratios for Static-99R. *Sexual Abuse: A Journal of Research and Treatment, 25*, 482-515. doi:10.1177/1079063212469060

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior, 36*, 865-891. doi:10.1177/0093854809338545

Hanson, R. K., Broom, I., & Stephenson, M. (2004). Evaluating community sex offender treatment programs: A 12-year follow-up of 724 offenders. *Canadian Journal of Behavioural Science, 36,* 87-96.

Hanson, RK, & Bussière, MT. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*, 348-362. doi: 10.1037/0022-006X.66.2.348

*Hanson, R. K., Helmus, L., & Harris, A. J. R. (2014). *Assessing the risk and needs of supervised sexual offenders: A prospective study.* Unpublished manuscript.

Hanson, R. K., & Howard, P. D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior, 4,* 275 – 281. doi:10.1007/s10979-010-9227-3.

*Hanson, R. K., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of the Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management, 1,* 102-117.

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 scales. *International Journal of Forensic Mental Health, 11*, 9-23. doi:10.1080/14999013.2012.667511

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1-21. doi:10.1037/a0014421

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24(*1), 119-136. doi:10.1023/A:1005482921333

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of the Static-2002* (User Report 2003-01). Ottawa, ON: Solicitor General Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/2003-01-not-sttc-eng.pdf

Hanson, R. K., & Thornton, D. (2008, October). *Recommendations for interpreting multiple norms for the Static-99.* Presentation at the 27th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Atlanta, GA. Available from www.static99.org.

Hanson, R. K., & Thornton, D. (2012, October). *Preselection effects can explain variability in sexual recidivism base rates in Static-99R and Static-2002R validation studies.* Presentation at the 31st Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Denver, CO.

*Harkins, L., & Beech, A. R. (2007). *Examining the effectiveness of sexual offender treatment using risk band analysis*. Unpublished manuscript.

Harris, A. J. R., Helmus, L., Hanson, R. K., & Thornton, D. (2008, October). *Are new norms needed for Static-99*? Paper presented at the 27th Annual Conference of the Association for the Treatment of Sexual Abusers, Atlanta, GA.

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

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments Evaluating the margins of error of group v. individual predictions of violence. *The British Journal of Psychiatry*, *190*(49), s60-s65. doi:10.1192/bjp.190.5.s60

Heilbrun, K., O'Neill, M., Stevens, T. N., Strohman, L. K., Bowman, Q., & Lo, Y. W. (2004). Assessing normative approaches to communicating violence risk: A national survey of psychologists. *Behavioral Sciences and the Law, 22,* 187-196. doi:10.1002/bsl.570

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. (2011). More fun with statistics! How to use logistic regression to predict criminal recidivism risk. *Crime Scene, 18*(2), 7-12.

Helmus, L., & Hanson, R. K. (2013). STABLE-2007: Updated recidivism rates. Unpublished document. Ottawa: Public Safety Canada.

Helmus, L., Hanson, R.K., Babchishin, K.M., & Thornton, D. (2014). Sex offender risk assessment with the Risk Matrix 2000: Validation and guidelines for combining with the STABLE-2007. *Journal of Sexual Aggression.* Advance online publication. doi:10.1080/13552600.2013.870241

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*(9), 1148-1171. doi:10.1177/0093854812443648

Helmus, L., Hanson, R. K., & Thornton, D. (2009). Reporting Static-99 in light of new research on recidivism norms. *The Forum* (ATSA Newsletter), *19*(4). Available at http://newsmanager.commpartners.com/atsa/issues/2009-01-21/5.html

Helmus, L., & Thornton, D. (2014). *Stability, predictive, and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism: A meta-analysis*. Unpublished manuscript.

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

Hill, A., Habermann, N., Klusmann, D., Berner, W., & Briken, P. (2008). Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy and Comparative Criminology, 52,* 5-20. doi:10.1177/0306624X07307450

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

Interstate Commission for Adult Offender Supervision. (2007). *Sex offender assessment information survey* (ICAOS Documents No. 4-2007). Lexington, KY: Author.

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts**.** *Sexual Abuse: A Journal of Research and Treatment, 19,* 409-448. doi:10.1007/s11194-007-9062-3

*Johansen, S. H. (2007). Accuracy of predictions of sexual offense recidivism: A comparison of actuarial and clinical methods. *Dissertations Abstracts International, 68* (03), B. (UMI No. 3255527).

Jones, N., Pelissier, B., & Klein-Saffran, J. (2006). Predicting sex offender treatment entry among individuals convicted of sexual offense crimes. *Sexual Abuse: A Journal of Research and Treatment, 18,* 83-98. doi:10.1007/s11194-006-9005-4

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.

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

Långström, N., Enebrink, P., Laurén, E.-M., Lindblom, J., Werkö, S., & Hanson, R.K. (2013). Preventing sexual violence against children: Systematic review of medical and psychological interventions. *BMJ, 347*:f4630. doi:10.1136/bmj.f4630.

Lehmann, R. J., Goodwill, A. M., Gallasch-Nemitz, F., Biedermann, J., & Dahle, K. P. (2013). Applying crime scene analysis to the prediction of sexual recidivism in stranger rapes. *Law and Human Behavior*, *37*(4), 241-254. doi: 10.1037/lhb0000015

*Lehmann, R. J. B., Hanson, R. K., Babchishin, K. M., Gallasch-Nemitz, F., Biedermann, J., & Dahle, K.-P. (2013). Interpreting multiple risk scales for sex offenders: Evidence for averaging. *Psychological Assessment, 25,* 1019-1024. doi:10.1037/a0033098

Lösel, F., & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology, 1,* 117-146. doi:10.1007/s11292-004-6466-7

McGrath, R. J., Cumming, G. F., & Burchard, B. L., Zeoli, S., & Ellerby, E. (2010). *Current practices and emerging trends in sexual abuser management: The Safer Society 2009 North American Survey* (ISBN: 978-1-884444-85-2). Brandon, VT: Safer Society Press.

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse: A Journal of Research and Treatment.* Advance online publication. doi:10.1177/1079063211432475

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

*Nicholaichuk, T. (2001, November). *The comparison of two standardized risk assessment instruments in a sample of Canadian Aboriginal sexual offenders.* Paper presented at

the annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, San Antonio, TX.

Olver, M. E., Beggs Christofferson, S. M., Grace, R. C., & Wong, S. C. P. (2013). Incorporating change information into sexual offender risk assessments using the Violence Risk Scale-Sexual Offender Version. *Sexual Abuse: A Journal of Research and Treatment.* Advance online publication. doi:10.1177/1079063213502679

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

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354-379. doi:10.1037//1082-989X.3.3.354

Phenix, A., Helmus, H., & Hanson, R.K. (2012). *Static-99R and Static-2002R Evaluators' Workbook.* Downloaded from www.static99.org on April 19, 2013.

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science, 10*(37), 25-42.

Saum, S. (2007). A comparison of an actuarial risk prediction measure (Static-99) and a stable dynamic risk prediction measure (Stable-2000) in making risk predictions for a group of sexual offenders. *Dissertations Abstracts International, 68* (03), B. (UMI No. 3255539).

Schulze, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Zeitschrift für Psychologie/Journal of Psychology, 215*, 90-103. doi:10.1027/0044-3409.215.2.90

Sreenivasan, S., Weinberger, L. E., Frances, A., & Cusworth-Walker, S. (2010). Alice in actuarial-land: Through the looking glass of changing Static-99 norms. *Journal of the American Academy of Psychiatry and the Law, 38,* 400-406.

Storey, J. E., Watt, K. A., Jackson, K. J., & Hart, S. D. (2012). Utilization and implications of the Static-99 in practice. *Sexual Abuse: A Journal of Research and Treatment*, *24*, 289-302. doi:10.1177/1079063211423943

*Swinburne Romine, R., Dwyer, S. M., Mathiowetz, C., & Thomas, M. (2008, October). *Thirty years of sex offender specific treatment: A follow-up Study*. Poster presented at the conference for the Association for the Treatment of Sexual Abusers, Atlanta, GA.

*Ternowski, D. R. (2004). Sex offender treatment: An evaluation of the Stave Lake Correctional Centre Program. *Dissertations Abstracts International, 66* (06), 3428B. (UMI No. NR03201).

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment, 14,* 139-153. doi:10.1177/107906320201400205

Thornton, D., Hanson, R. K., & Helmus, L. (2010). Moving beyond the standard model for actuarial assessment for sexual offenders. *California Coalition on Sexual Offending, CCOSO Quarterly Newsletter, Perspectives,* Issue: Spring 2010. Available at http://ccoso.org/newsletter.php.

Thornton, D., & Knight, R. A. (2013). Construction and Validation of SRA-FV Need Assessment. *Sexual Abuse: A Journal of Research and Treatment*, Advance online publication. doi:10.1177/1079063210384274

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbeman, J. D. F. (2005). Substantial effective samples sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology, 58,* 475-483. doi:10.1016/j.jclinepi.2004.06.017

*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., & Looman, J. (2010). What can we reasonably expect to accomplish in conducting actuarial risk assessments with sexual offenders in civil commitment settings? A response to Campbell and DeClue: "Maximizing predictive accuracy in sexually violent predator evaluations." *Open Access Journal of Forensic Psychology, 2*, 306-321.

*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. doi:10.1111/j.1468-2311.2007.00480.x

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, *39*, 1511-1538. doi:10.1177/0093854812455741

Table 1

Descriptive Information

Study	*n*	Age at release *M* (SD)	Age range	Country	Recidivism Criteria	Type of Sample	Release Period
Routine/Complete							
Bartosh et al. (2003)	186	38 (12)	18-75	U.S.	Charges	Corrections	1996
Bigras (2007)	483	43 (12)	20-77	Canada	Charges	CSC Reception Centre	1995-2004
Boer (2003)	299	41 (12)	20-80	Canada	Conviction	CSC Release	1976-1994
Craissati et al. (2011)	209	38 (12)	18-72	U.K.	Conviction	Community supervision	1992-2005
Eher et al. (2009)	706	41 (12)	18-77	Austria	Conviction	European prison	2000-2005
Epperson (2003)	177	37 (13)	19-77	U.S.	Charges	Prison and Probation	1989-1998
Hanson, Helmus et al. (2014)	764	41 (14)	18-84	Canada	Charges	Community supervision	2001-2009
Hanson, Lunetta et al. (2014)	495	42 (11)	20-86	U.S.	Charges	California Prison Release	2006-2007
Långström (2004)	1,278	41 (12)	18-77	Sweden	Conviction	National Prison Release	1993-1997
Lehmann, Hanson et al. (2013)	936	38 (12)	18-78	Germany	Conviction	Berlin Police Registry	1994-2009
Preselected Treatment							
Allan et al. (2007)	492	42 (12)	19-77	New Zealand	Charges	Prison treatment	1990-2000
Brouillette-Alarie & Proulx (2008)	228	36 (10)	18-67	Canada	Conviction	Prison & community treatment	1979-2006
Harkins & Beech (2007)	197	43 (12)	18-76	U.K.	Conviction	Prison & community treatment	1994-1998
Johansen (2007)	273	38 (11)	20-75	U.S.	Charges	Prison treatment	1994-2000
Swinburne Romine et al. (2008)	680	38 (12)	18-82	U.S.	Conviction	Community treatment	1977-2007
Ternowski (2004)	247	44 (13)	20-79	Canada	Charges	Prison treatment	1994-1998
High Risk/High Need							
Bengtson (2008)	311	33 (10)	18-66	Denmark	Charges	Forensic psychiatric evaluations	1978-1995
Bonta & Yessine (2005)	133	40 (10)	20-66	Canada	Conviction	Preselected high risk	1992-2004
Haag (2005)	198	37 (10)	21-73	Canada	Conviction	Detained until end of sentence	1995
Nicholaichuk (2001)	281	35 (9)	19-69	Canada	Conviction	High intensity treatment	1983-1998
Wilson et al. (2007a & b)	232	42 (11)	20-78	Canada	Charges	Preselected high risk	1994 -2007
Total	8,805	40 (12)	18-86	-	-	-	1976-2009

Note. CSC = Correctional Service Canada (administers all sentences of at least two years). For range of ages at release, age was not rounded up (e.g., an offender who was 50.9 years old was still considered age 50).

Table 2

Recidivism Information

				Sexual Recidivism				General (Any) Recidivism			
				Overall		5 Years		Overall		5 Years	
	Static-99R *M* (*SD*)	Static-2002R *M* (*SD*)	Years Follow-Up *M* (*SD*)	*n*	Recid (%)	*n*	Recid (%)	*n*	Recid (%)	*n*	Recid (%)
Routine/Complete											
Bartosh et al. (2003)	3.3 (2.9)	-	5.0 (0.2)	186	11.8	90	13.3	186	55.4	90	58.9
Bigras (2007)	2.1 (2.4)	3.5 (2.5)	4.6 (1.9)	483	6.2	206	8.7	483	23.8	204	34.3
Boer (2003)	2.8 (2.8)	3.9 (2.7)	13.3 (2.1)	299	8.7	299	3.7	299	48.5	299	33.4
Craissati et al. (2011)	2.2 (2.3)	-	9.1 (2.7)	209	11.5	200	7.5	209	41.1	200	23.0
Eher et al. (2008)	2.3 (2.3)	-	3.9 (1.1)	706	4.0	151	2.0	706	26.2	151	27.8
Epperson (2003)	2.5 (2.6)	-	7.9 (2.5)	177	14.1	150	10.7	-	-	-	-
Hanson, Helmus et al. (2014)	2.4 (2.4)	3.5 (2.5)	7.2 (2.3)	764	10.9	629	12.2	764	31.5	629	32.1
Hanson, Lunetta et al. (2014)	2.1 (2.3)	-	5.3 (0.7)	495	5.5	473	4.9	495	47.3	466	46.4
Långström (2004)	2.0 (2.4)	-	8.9 (1.4)	1,278	7.5	1,278	5.4	-	-	-	-
Lehmann et al. (2013)	3.5 (2.2)	4.1 (1.9)	9.6 (3.3)	936	16.2	849	13.4	936	61.6	849	56.7
Preselected Treatment											
Allan et al. (2007)	1.8 (2.3)	-	5.7 (2.9)	492	9.6	298	11.7	492	25.2	298	26.2
Brouillette-Alarie & Proulx (2008)	3.9 (2.3)	-	9.9 (4.5)	228	20.2	199	14.6	-	-	-	-
Harkins & Beech (2007)	2.2 (2.6)	3.7 (2.8)	10.4 (1.1)	197	14.2	197	9.6	197	36.0	197	27.9
Johansen (2007)	2.9 (2.3)	-	9.1 (1.1)	273	7.7	272	5.9	273	53.5	272	42.6
Swinburne Romine et al. (2008)	1.7 (2.2)	-	16.8 (7.8)	680	13.8	569	8.4	-	-	-	-
Ternowski (2004)	1.6 (2.5)	-	7.5 (1.0)	247	8.1	247	6.5	247	19.8	247	17.8
High Risk/High Need											
Bengtson (2008)	3.8 (2.4)	4.6 (2.4)	16.2 (4.2)	311	33.8	310	19.7	311	64.6	310	42.9
Bonta & Yessine (2005)	5.0 (2.1)	-	5.5 (2.4)	133	15.8	81	17.3	133	48.9	81	49.4
Haag (2005)	3.9 (2.3)	5.7 (2.3)	7.0 (0.0)	198	25.3	198	19.7	-	-	-	-
Nicholaichuk (2001)	4.8 (2.4)	-	6.4 (4.0)	281	18.5	168	22.6	-	-	-	-
Wilson et al. (2007a & b)	5.1 (2.3)	-	5.2 (3.0)	232	10.3	103	11.7	232	35.8	103	41.7
Overall	2.6 (2.6)	4.0 (2.4)	8.5 (4.8)	8,805	11.6	6,967	9.8	5,731	40.9	4,293	39.0

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

Table 3

Meta-analysis of logistic regression coefficients for Static-99R and Static-2002R predicting sexual recidivism after 5 years of follow-up

	Fixed Effect		Random Effects		*Q*	I^2	*n/N*	*k*
	M	95% CI	*M*	95% CI				
Static-99R								
Base Rate ($B0_2$ as %)	6.7	6.0 – 7.5	6.5	5.3 – 8.0	62.19***	67.8	685/6,967	21
Relative Risk (e^{B1})	1.39	1.34 – 1.44	1.39	1.33 – 1.46	28.80	30.5	685/6,967	21
Static-2002R								
Base Rate ($B0_3$ as %)	8.0	6.8 – 9.4	6.6	4.4 – 10.0	29.92***	79.9	332/2,651	7
Relative Risk (e^{B1})	1.40	1.32 – 1.48	1.42	1.29 – 1.57	16.35*	63.3	332/2,651	7

Note: n/N is the number of recidivists and the total number of offenders. The base rate is the expected recidivism rate for offenders with median scores on Static-99R (2) or Static-2002R (3).

*$p < .05$, ***$p < .001$.

Table 4

Differences in risk and recidivism between routine, preselected, and high risk/high need samples

	Routine/Complete				Preselected Treatment				High Risk/ High Need			
	N (k)	*M*	95% CI		*N (k)*	*M*	95% CI		*N (k)*	*M*	95% CI	
			LL	*UL*			*LL*	UL			*LL*	UL
Age (years)	5,529 (10)	**40.3**	39.1	41.5	2,117 (6)	**40.2**	37.8	42.7	1,155 (5)	**37.2**	34.0	40.3
Total Static-99R	5,533 (10)	**2.5**[a]	2.1	2.9	2,117 (6)	**2.3**[a]	1.6	3.0	1,155 (5)	**4.5**[b]	4.0	5.1
Total Static-2002R	2,394 (4)	**3.8**[a]	3.4	4.1	190 (1)	**3.7**[a]	3.3	4.1	498 (2)	**5.2**[b]	4.5	6.4
Sexual Recidivism at 5 years												
Raw Rate	4,325 (10)	**7.6%**[a]	5.2%	10.5%	1,782 (6)	**9.2%**[a]	6.9%	11.7%	860 (5)	**18.7%**[b]	15.5%	22.1%
Controlling for Static-99R	4,325 (10)	**4.8%**[a]	3.1%	7.4%	1,782 (6)	**6.9%**	4.7%	10.0%	860 (5)	**10.3%**[b]	6.3%	16.4%
Controlling for Static-2002R	1,964 (4)	**5.1%**	2.6%	9.8%	190 (1)	**5.2%**	1.6%	15.4%	497 (2)	**12.3%**	6.3%	22.0%

Note. LL= Lower limit; UL = Upper limit. Random-effects meta-analysis was used to compute point estimates, calculate confidence intervals, and compare groups. The meta-analysis of the raw percentages used the variance stabilization transformations. Logistic regression intercepts (centered on the median) were used when controlling for Static-99R (2) and Static-2002R (3).

[a,b] Groups with different superscripts were significantly different from one another ($p < .05$).

Table 5

Incremental effects of Sample Type, individual Static-99R scores and mean (group) Static-99R scores

	Exp(*B*)	95% CI	Wald	*df*	*p*
Step 1: Static-99R	1.36	1.32 – 1.39	624.62	1	<.001
Step 2a: Static-99R	1.34	1.30 – 1.37	499.15	1	<.001
Mean Static-99R	1.12	1.05 – 1.19	11.59	1	<.001
Step 2b: Static-99R	1.34	1.30 – 1.37	522.01	1	<.001
Treatment	1.25	1.07 – 1.45	8.18	1	.004
High Risk/Need	1.58	1.36 – 1.85	33.24	1	<.001
Step 3: Static-99R	1.34	1.30 – 1.37	498.71	1	<.001
Mean Static-99R	0.99	0.89 - 1.09	0.08	1	.777
Treatment	1.24	1.07 – 1.45	7.45	1	.005
High Risk/Need	1.62	1.28 – 2.06	16.34	1	<.001

Note. N = 8,805 with 1,021 sexual recidivists. Mean (group) computed as the average Static-99R score per sample. Sample was not use as strata given that each sample had a unique mean score. For Step 2a, $\Delta\chi^2$ was 11.38, df = 1, p < .001. For Step 2b, $\Delta\chi^2$ was 33.39, df = 2, p < .001.

Table 6

Incremental effects of Sample Type, individual Static-2002R scores and mean (group) Static-2002R scores

	Exp(*B*)	95% CI	Wald	*df*	*p*
Step 1: Static-2002R	1.34	1.29 – 1.39	239.52	1	<.001
Step 2a: Static-2002R	1.32	1.27 – 1.37	203.09	1	<.001
Mean Static-2002R	1.34	1.16 – 1.54	16.60	1	<.001
Step 2b: Static-2002R	1.31	1.26 – 1.36	198.67	1	<.001
Treatment	0.98	0.66 – 1.45	0.01	1	.913
High Risk/Need	1.68	1.37 – 2.07	24.72	1	<.001
Step 3: Static-2002R	1.31	1.26 – 1.36	198.03	1	<.001
Mean Static-2002R	0.98	0.76 – 1.26	0.03	1	.869
Treatment	0.98	0.66 – 1.45	0.02	1	.902
High Risk/Need	1.73	1.21 – 2.47	8.95	1	.003

Note. *N* = 3,083 with 461 sexual recidivists. Mean (group) computed as the average Static-2002R score per sample. Sample was not use as strata given that each sample had a unique mean score. For Step 2a, $\Delta\chi^2$ was 15.56, *df* = 1, *p* < .001. For Step 2b, $\Delta\chi^2$ was 24.22, *df* = 2, *p* < .001.

Table 7

Fixed-effect meta-analysis of logistic regression parameters for routine and high risk/high need samples

	Parameter	*SE*	*Q*	I^2	*k*	*N*	*N* recidivists
Static-99R							
5-year estimates							
Base Rate ($B0_2$)							
Routine/complete Samples	-2.826701	.079267	34.23***	73.71	10	4,325	358
High Risk/High Need	-2.063626	.153444	3.14	0.00	5	860	164
Relative Risk (*B1*)							
Routine/complete Samples	.368494	.025118	13.61	33.88	10	4,325	358
High Risk/High Need	.250091	.042447	4.85	17.52	5	860	164
10-year estimates							
Base Rate ($B0_2$)							
High Risk/High Need	-1.442304	.186037	0.60	0.00	2	350	98
Relative Risk (*B1*)							
High Risk/High Need	.230673	.056260	0.0	0.00	2	350	98
Static-2002R							
5-year estimates							
Base Rate ($B0_3$)							
Routine/complete Samples	-2.623861	.109268	12.31**	75.62	4	1,964	217
High Risk/High Need	-1.877664	.173236	2.69	62.81	2	497	97
Relative Risk (*B1*)							
Routine/complete Samples	.395056	.036062	6.61	54.60	4	1,964	217
High Risk/High Need	.217648	.050133	1.41	29.32	2	497	97

Note: The median values of the correlation of estimates (from combining routine/complete and HRN) were as follows: Static-99R 5-year estimates, r = -.737911, Static-99R 10-year estimates, r = -.728513, Static-02R 5-year estimates, r = -.744130.

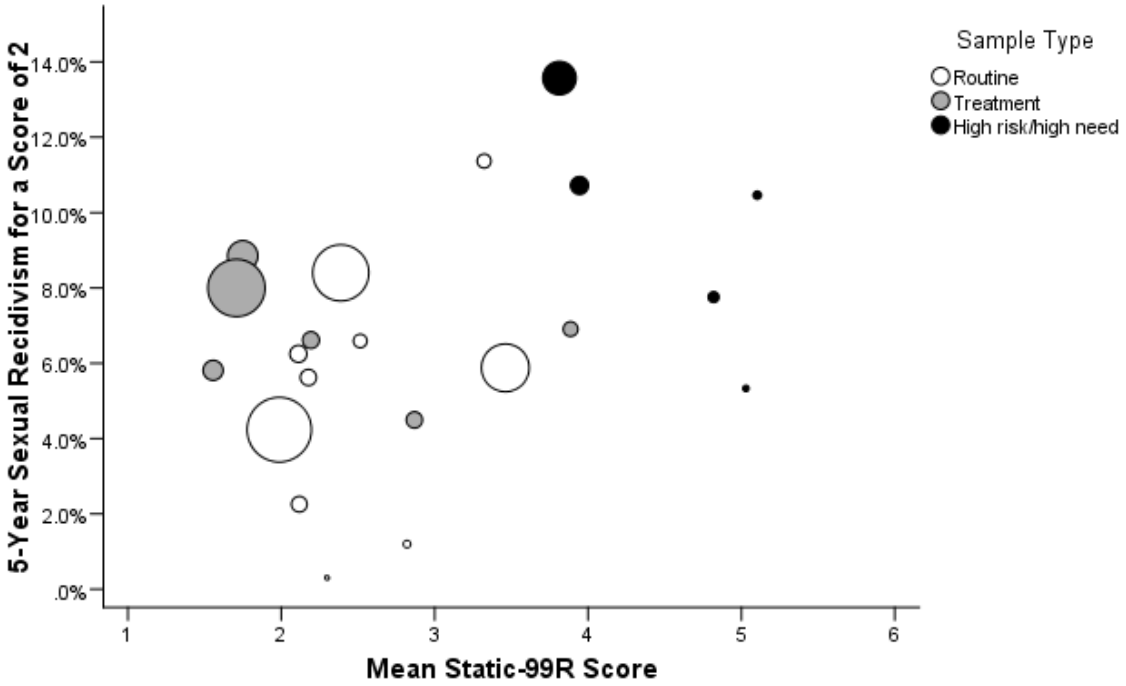
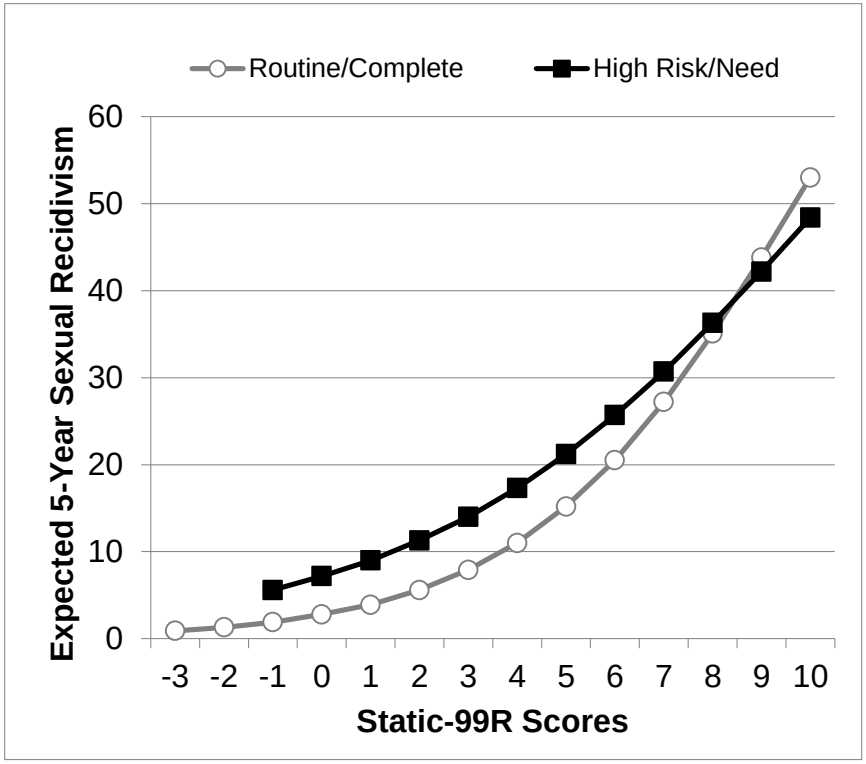


*Figure 1.*The relationship between Static-99R adjusted sexual recidivism rates and the average Static-99R score per sample. Size of bubble based on weight of the sample in the meta-analysis of Static-99R recidivism rates.



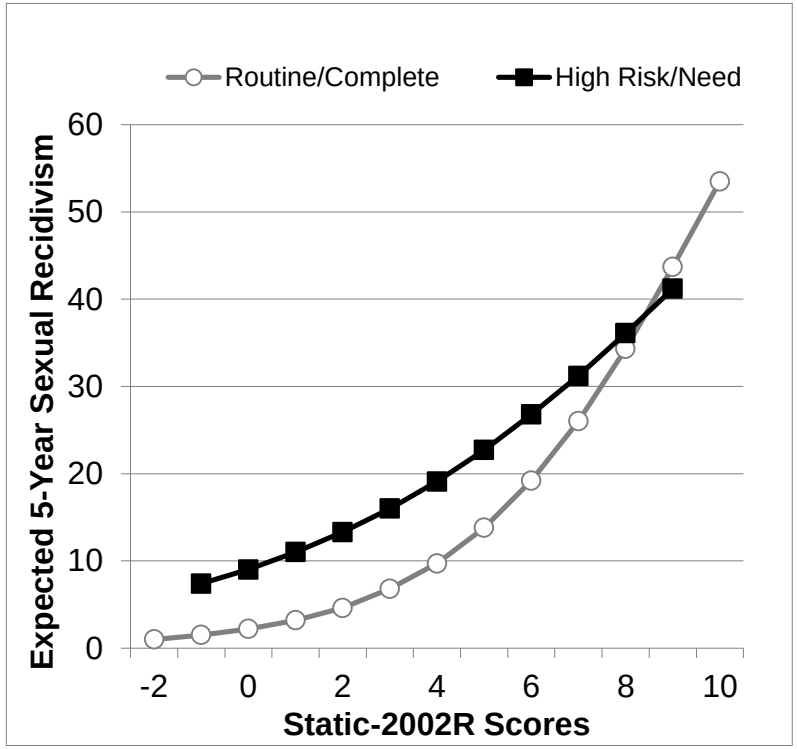


Figure 2. Five-Year sexual recidivism rates for Static-99R and Static-2002R normative groups

.

Appendix A

Calculating Recidivism Rate Estimates

The following procedures were used to calculate the recidivism rate estimates shown in Appendix B Table 1B and Table 2B from the logistic regression parameters in Table 7. In logistic regression, the predicted value, y, is a logit, or log odds, defined as ln(*p*/(1-*p*)), and

$$y = B0 + (B1)*(\text{STATIC SCORE})$$

For example, the logit of the expected 5-year recidivism rate (*p*) for offenders with a Static-99R score of 3 from routine/complete samples would be -2.458207 = (-2.826701) + (0.368494)*(+1). Notice that the Static-99R score is quantified as +1, not +3, because the *B0*s in Table 7 were centered on the median values (2 for Static-99R). To transform a logit into a probability, use either of the following formula:

$$p = \frac{e^{LOGIT}}{1+e^{LOGIT}} \quad \text{or, equivalently,} \quad p = \frac{1}{1+e^{-LOGIT}}$$

Continuing with the current example, a logit of -2.458207 is equivalent to a recidivism rate (*p*) of 7.88%.

To compute the confidence intervals of the predicted values, first compute the standard errors of the logits. Because the predicted value is a composite score, its standard error is based on the standard error of *B0 (*referred to as SE_{B0}*)*, the standard error of the *B1 (*referred to as SE_{B1}*)*, and the correlation of the estimates (*r*; see Hosmer & Lemeshow, 2000):

$$SE_{logit} = \sqrt{\left(SE_{B0}\right)^2 + 2 \cdot x \cdot r \cdot SE_{B0} \cdot SE_{B1} + x^2\left(SE_{B1}\right)^2}$$

In this formula, *x* refers the risk score for which we are predicting recidivism. The correlation of estimates is a scaling factor that ensures that the standard errors do not change based on arbitrary changes to the raw scores (e.g., centering on medians). Note that the standard errors are different for each score.

From the standard error, confidence intervals for the estimated recidivism rates (as logits) can be obtained in the usual way:

$$95\%\ CI = logit \pm (1.96 \cdot (SE_{logit}))$$

Once you have the confidence intervals in logit units, you can transform them into probabilities using the formula reported above.

Continuing with our example (Static-99R score of 3; logit of -2.458), from Table 6 we obtain SE_{B0} = 0.079267, SE_{B1} = .025118, and the correlation of estimates of -.737911. With these values, the standard error of our estimated recidivism rate would be

$SE_{logit} = [(0.079267)^2 + (2*1(-.737911)(0.079267)(0.025118)) + (1)^2(0.025118)^2]^{1/2}$

$= \sqrt{.006283 - .002938 + .0006309} = 0.063.$

Using 0.063 as the *SE* for a Static-99R score of 3, the 95% confidence interval around our logit would be $-2.458 \pm (1.96 \cdot (0.063)) = -2.582$ to -2.334, or 7.0% to 8.8%.

Substituting the appropriate values from Table 7, the above procedures could be used to calculate recidivism rate estimates (and their standard errors) for other sample types (high risk/high need) and other follow-up periods (i.e., 10 years). These tables are available at www.static99.org.

A basic tutorial on logistic regression in the context of recidivism prediction is provided by Helmus and Hanson (2011).

Appendix B

Table 1B

Observed and estimated 5-year sexual recidivism rates for Static-99R: Routine/complete Samples

Score	Fixed Follow-up		Logistic Regression Estimates		
	Recidivists/ total	Observed Recidivism Rate (%)	Predicted Recidivism Rate	95% CI	
-3	0/61	0.0	0.9	0.6	1.3
-2	1/90	1.1	1.3	1.0	1.8
-1	10/357	2.8	1.9	1.4	2.5
0	13/468	2.8	2.8	2.2	3.5
1	23/590	3.9	3.9	3.3	4.7
2	24/661	3.6	5.6	4.8	6.5
3	48/675	7.1	7.9	7.0	8.8
4	58/576	10.1	11.0	10.0	12.1
5	52/365	14.2	15.2	13.8	16.6
6	47/231	20.3	20.5	18.4	22.8
7	36/133	27.1	27.2	24.0	30.7
8	29/79	36.7	35.1	30.5	40.0
9	10/26	38.5	43.8	37.8	50.1
10	5/10	50.0	53.0	45.6	60.3
11	2/3	66.7			
Total	358/4,325	8.3			

Table 2B

Observed and estimated 5-year sexual recidivism rates for Static-2002R: Routine/complete Samples

Score	Fixed Follow-up		Logistic Regression Estimates		
	Recidivists/ total	Observed Recidivism Rate (%)	Predicted Recidivism Rate	95% CI	
-2	0/24	0.0	1.0	0.6	1.7
-1	0/35	0.0	1.5	0.9	2.3
0	4/83	4.8	2.2	1.5	3.2
1	5/137	3.6	3.2	2.3	4.4
2	7/245	2.9	4.6	3.6	6.0
3	18/306	5.9	6.8	5.5	8.2
4	34/399	8.5	9.7	8.3	11.3
5	46/323	14.2	13.8	12.2	15.6
6	34/190	17.9	19.2	16.9	21.6
7	30/103	29.1	26.0	22.6	29.8
8	19/60	31.7	34.3	29.1	40.0
9	12/42	28.6	43.7	36.5	51.2
10	5/11	45.5	53.5	44.4	62.4
11	3/6	50.0			
12	-	-			
Total	217/1,964	11.0			