

Quantifying the Relative Risk of Sex Offenders: Risk Ratios for Static-99R

Sexual Abuse: A Journal of
Research and Treatment
25(5) 482–515
© The Author(s) 2012
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1079063212469060
sax.sagepub.com



R. Karl Hanson¹, Kelly M. Babchishin^{1,2},
Leslie Helmus², and David Thornton³

Abstract

Given the widespread use of empirical actuarial risk tools in corrections and forensic mental health, it is important that evaluators and decision makers understand how scores relate to recidivism risk. In the current study, we found strong evidence for a relative risk interpretation of Static-99R scores using 8 samples from Canada, United Kingdom, and Western Europe ($N = 4,037$ sex offenders). Each increase in Static-99R score was associated with a stable and consistent increase in relative risk (as measured by an odds ratio or hazard ratio of approximately 1.4). Hazard ratios from Cox regression were used to calculate risk ratios that can be reported for Static-99R. We recommend that evaluators consider risk ratios as a useful, nonarbitrary metric for quantifying and communicating risk information. To avoid misinterpretation, however, risk ratios should be presented with recidivism base rates.

Keywords

risk ratio, relative risk, Static-99R, sex offenders

Actuarial risk tools are structured methods for combining risk factors into a total score (or risk categorization) and provide empirically derived estimates of recidivism probabilities (Dawes, Faust, & Meehl, 1989). Risk tools are widely used in forensic psychology (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Jackson & Hess, 2007) and corrections (McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). A

¹Public Safety Canada, Ottawa, ON, Canada

²Carleton University, Ottawa, ON, Canada

³Sand Ridge Secure Treatment Center, Mauston, WI, USA

Corresponding Author:

R. Karl Hanson, Corrections Research, Public Safety Canada, 340 Laurier Avenue West, Ottawa, ON, Canada, K1A 0P8, Canada.

Email: Karl.Hanson@ps.gc.ca

number of different risk tools are available for evaluating the risk for crime and violence (Otto & Douglas, 2010), and many are regularly updated through empirically based revisions and by expanding their normative samples (Andrews, Bonta, & Wormith, 2010; Harris & Hanson, 2010). No single risk tool has been identified as clearly superior in predictive accuracy (Hanson & Morton-Bourgon, 2009; Yang, Wong, & Coid, 2010); each risk tool has its own strengths and weaknesses, and they vary in terms of practical utility and relevance for specific referral questions. Static-99 (Hanson & Thornton, 2000) is the most commonly used actuarial scale for sex offenders, widely used in Canada and the United States for treatment planning (Jackson & Hess, 2007; McGrath et al., 2010), community supervision (Interstate Commission for Adult Offender Supervision, 2007), and civil commitment evaluations (Jackson & Hess, 2007).

Many decisions throughout an offender's progression in the criminal justice system will involve actuarial risk tools, including sentencing, security classification, parole decisions, and treatment and supervision intensity. However, when different risk tools are used (as is often the case; Jackson & Hess, 2007), it becomes important to interpret the information provided by these measures. Consider the following example. In 2006, the State of Virginia switched from using the RRASOR (Hanson, 1997) to Static-99 (Hanson & Thornton, 2000) for screening sex offenders for its Sexually Violent Predator civil commitment program (Joint Legislative Audit and Review Commission, 2011). Both the RRASOR and Static-99 were designed to assess the likelihood of sexual recidivism based on commonly available criminal history information. However, Virginia's policy change had the consequence of increasing the proportion of sex offenders requiring further review from 7% to 24%.

Although Static-99 is widely used in civil commitment evaluations (Jackson & Hess, 2007), Virginia may be unique in requiring a specific score on a specific test (in this case, a Static-99 score of 5, or, if the victim was 13 or younger, a score of 4). In 2008, the authors of Static-99 presented new norms with substantially lower recidivism rates than the original norms (Harris, Helmus, Hanson, & Thornton, 2008, see also www.static99.org). In 2009, Static-99 was further revised by adjusting the age weights, creating Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012). If Virginia now wants to switch to Static-99R, which score threshold should they use? More generally, how do we judge the equivalency of risk scores generated from different measures? We believe relative risk is one useful indicator that allows for the comparison of several risk tools.

Unlike many psychological tests, empirical actuarial risk tools are not intended to be internally consistent measures of a single latent construct. Namely, risk tools are criterion-referenced measures, and the primary consideration for selecting items is their relationship with the outcome of interest. New items are justified based on their incremental contribution to predicting recidivism. The predictive accuracy of risk tools increases as the number of different risk-relevant latent constructs increases. Even when the reasons for the association with recidivism are unknown to the test developers, items are retained because of their ability to predict recidivism (e.g., unrelated victim item of Static-2002; see Hanson & Thornton, 2003). Consequently,

actuarial risk scales are rarely homogeneous because the items of a good risk scale (as assessed by predictive accuracy) will target a diverse set of psychologically meaningful risk factors (Mann, Hanson, & Thornton, 2010). This intentional heterogeneity limits the applicability of both classical (Nunnally & Bernstein, 1994) and modern test theory (i.e., item response theory; Embretson & Reise, 2000). Other measurement models are needed.

Measurement Models

Actuarial risk scales provide several types of information that could be of value to decision makers. The most obvious is the recidivism rate estimates associated with specific scores (e.g., “40% of offenders with a score of 5 are reconvicted within 5 years”). Absolute recidivism rate estimates are important for certain decisions. They are useful, for example, in informing decisions involving thresholds such as civil commitment evaluations, which typically require some estimate or approximation of absolute risk (e.g., “more likely than not,” “significant probability”). However, absolute recidivism rate estimates may not be the most fundamental information provided by risk scale scores. In this article, we argue that a more fundamental attribute is relative risk. Relative risk provides information about a particular offender’s risk relative to other offenders and can be quantified by risk ratios. We then present methods of calculating risk ratios for one specific sex offender risk scale: Static-99R (Helmus, Thornton et al., 2012).

Like other criterion-referenced measures, the items and scoring procedures for Static-99R were developed by examining empirical relationships with the outcome of interest—in this case, recidivism. Unlike some criteria, however, recidivism is an inherently changing characteristic (albeit in one direction). At the beginning of the follow-up period, no one has reoffended yet, and individuals are expected to change from nonrecidivists to recidivists without changes in their initial scores. This feature of recidivism prediction requires quite different measurement models than those used for evaluating diagnostic accuracy, in which the outcome is either present or absent but estimated with uncertainty (e.g., positive predictive accuracy, false positives, likelihood ratios; e.g., Akobeng, 2006a, 2006b).

Typically, developers of risk assessment tools intend for the scores to rank offenders consistently across different follow-up periods (Harris, Phenix, Hanson, & Thornton, 2003; Thornton et al., 2003). Offenders with a score of 6 are expected to always be more likely to reoffend than offenders with a score of 2; unless there is new information, high risk offenders are expected to remain higher risk than low risk offenders throughout the follow-up period. Consequently, the fundamental attribute indexed by risk measures must be some form of relative risk (“this offender is riskier than that offender”), not absolute risk (“this offender has a 30% chance of reoffending”).

There are at least two ways in which relative risk could be quantified in applied assessments: percentile ranks and risk ratios. Percentile ranks are commonly used in psychology, sports, and performance evaluation, and measure the “unusualness” of

assessment findings (Crawford & Garthwaite, 2009). In the context of risk assessment, evaluators can state, for example, that an offender with a Static-99R score of “5” is in the top 15% in terms of risk for sexual recidivism (Hanson, Lloyd, Helmus, & Thornton, 2012). Although useful in certain contexts, percentile ranks have limitations as an indicator of relative risk. When more than one individual has the same risk score, there is no a priori expectation that a one-unit increase in risk score would be associated with an equal and consistent change in percentile ranks. More generally, the relationship between the distribution of scores and the likelihood of recidivism is often unknown, and is unlikely to be linear.

In comparison to percentile ranks, risk ratios are more closely aligned with the essential information provided by risk scales. Whereas percentiles measure “unusualness,” risk ratios measure changes in the likelihood of recidivism, which, if our analysis is correct, is a fundamental attribute of risk scales. The developers of Static-99R have specifically stated that each increase in Static-99R score is intended to represent a 30% increase in the relative likelihood of recidivism, as quantified by odds ratios or rate ratios of approximately 1.3 (Helmus, Thornton et al., 2012, p. 80). This working assumption was adopted in the development of Static-99 and Static-2002 to facilitate comparisons of the relative importance of specific risk factors across samples with diverse follow-up times (Hanson & Thornton, 2003, p. 13). The validity of this scaling heuristic has not been formally evaluated prior to the current study.

The measurement model for Static-99R also contains other assumptions. Given that there are multiple ways of arriving at the same total score (other than the minimum or maximum), each Static-99R item is assumed to be contributing stochastically, incrementally, and equally to recidivism risk. These assumptions imply that Static-99R relative risk ratios cannot decrease as scores increase and that the shape of the relationship between Static scores and risk should be relatively smooth, with no abrupt changes in relative risk from one score to the next. Furthermore, this model implies that the shape of the function linking Static-99R scores to recidivism should be approximately exponential (i.e., increase in scores are associated with the same proportional increase in recidivism rate) or logistic (i.e., an “S” shaped curve based on logits or log odds). There is, however, no a priori requirement that a one-unit increase provides strictly the same increase in relative risk across all values of the scale. It is quite plausible, for example, that increasing scores on a risk scale will eventually reach a point of diminishing returns. That is, once a certain number of risk factors have been identified, additional risk factors may not add anything new in the prediction of recidivism.

Evaluating the proposed measurement model for Static-99R requires methods for quantifying relative risk. We use the general term *risk ratios* to refer to statistics that quantify the extent to which one offender is riskier than another offender. There are three common forms of risk ratios: rate ratios, odds ratios, and hazard ratios. Each of these risk ratios are well-established metrics for describing relative risk, each with their own strengths and weaknesses (see Crowson, Therneau, Matteson, & Gabriel, 2007; Fleiss, 1994; Singer & Willett, 2003).

Risk Ratios

Rate ratios are the ratio of two probabilities (or rates). For example, if the recidivism rate for a score of 1 is $p_1 = .10$ and the recidivism rate for a score of 2 is $p_2 = .15$, the rate ratio for these adjacent scores would be $p_2/p_1 = .15/.10 = 1.50$ (i.e., offenders with a score of 2 have 1.5 times the recidivism rate of offenders with a score of 1). An odds is defined as $p/(1-p)$, and an odds ratio is a ratio of odds. Continuing with the current example, the odds of recidivism for a score of 1 would be $.10/(1-.10) = .111$, the odds of recidivism for a score of 2 would be $.15/(1-.15) = .176$, and the odds ratio for adjacent scores would be $.176/.111 = 1.59$ (i.e., the odds of recidivism for offenders with a score of 2 is 1.6 times the odds of recidivism for offenders with a score of 1). Odds are typically transformed into their natural log (log odds or *logits*) to normalize their variance, and odds ratios are commonly estimated through logistic regression (Hosmer & Lemeshow, 2000). Hazard rates are a quantity intrinsic to survival analysis where the basic data is time to a dichotomous outcome (Aalen, Borgan, & Gjessing, 2008). A hazard rate is the probability that if the event in question (e.g., sexual recidivism) has not already occurred, it will occur in the next time interval. A hazard ratio, of course, is the ratio of hazards. Hazard ratios are typically assumed to be constant over time and, thus, indicate the relative likelihood of recidivism in one group versus a comparison group. The hazard ratios can be best explained as being similar to a rate ratio, but whereas rate ratios are restricted to a fixed follow-up period (e.g., at 5-year), hazard ratios provide an indicator of relative risk at any given point in time.

When the base rates are low, all of these metrics provide similar results, as was the case in the current study. They increasingly diverge, however, when the expected probabilities increase. Rate ratios and hazard ratios assume exponential functions, which can result in impossible values (probabilities > 1) when the base rates are high. This is not the case for odds ratios because they are constructed to restrict the estimated probabilities between 0 and 1. In many prediction contexts, data are often well represented by exponential or logistic distributions. When the data do not fit the implied model (either exponential or logistic), however, it is possible to improve the fit by allowing one or more gradual curves in the function linking the scores to recidivism.

The current study examines the extent to which Static-99R's relative risk measurement model fit the available data. Specifically, each Static-99R point was expected to be associated with an approximately 30% increase in relative risk. Furthermore, relative risk was expected to be consistent across follow-up times (between 2 years and 10 years) and across samples. These three assumptions were examined using 8 samples comprising an aggregated sample of 4,037 sex offenders from Canada, United Kingdom, and Western Europe. The offenders in these samples had not been preselected on risk relevant characteristics, and, consequently, represent relatively routine samples from their respective jurisdictions. A secondary goal of this study was to develop risk ratios for Static-99R that can be used in applied evaluations.

Method

Measure

Static-99R. Static-99R is a 10-item actuarial scale that assesses recidivism risk of adult male sex offenders. The items and scoring rules are identical to Static-99 (Hanson & Thornton, 2000; Harris et al., 2003; see also www.static99.org) with the exception of updated age weights (Helmus, Thornton et al., 2012). The 10 items cover demographics (age at release; relationship history), sexual criminal history (prior sexual offences, any male victims, any unrelated victims, any stranger victims, any non-contact sexual offences), and general criminal history (prior sentencing dates, index nonsexual violence, prior nonsexual violence). Total scores range between -3 and 12. A recent meta-analysis found a moderate relationship between Static-99R and sexual recidivism (AUC = .69, 95% CI [.66, .72], $k = 22$, $n = 8,055$; Helmus, Hanson, Thornton, Babchishin, & Harris, 2012).

Samples

The 8 samples ($N = 4,037$) used in the current study were selected from a larger group of studies used for the renorming of Static-99 (Helmus, 2009). Of the 29 data sets available, 23 had the necessary information for calculating Static-99R risk ratios for sexual recidivism; however, only 8 approximated routine samples that had not been preselected on risk relevant characteristics or the need for treatment. These 8 samples were selected as most representative of the complete population of adjudicated sex offenders in their respective jurisdictions (see Table 1 for descriptive information from the samples).

Bartosh, Garby, Lewis, and Gray (2003). The study examined sex offenders released from the Arizona Department of Corrections and subject to registration and notification legislation.

Bigras (2007). The study examined 94% of all sex offenders receiving a federal sentence (2 or more years) in Quebec between 1995 and 2000 (6% refused participation in the research or were unable to provide consent).

Boer (2003). The study examined all male federal offenders serving a federal sentence for a sexual offence in British Columbia whose Warrant Expiry Date (WED; the end of their sentence) was between January 1990 and May 1994. Many offenders are granted conditional release prior to their WED; thus offenders in this sample were released as early as 1976.

Craissati, Bierer, and South (2011). The study examined all contact sex offenders on probation in two boroughs in South East London during the study period.

Eher, Rettenberger, Schilling, and Pfafflin (2008, 2009). The study examined sex offenders released from prison in Austria. The sample in this data set was approximately twice the size of the sample in an earlier report of this project (Eher et al., 2008).

Epperson (2003). The study examined sex offenders in North Dakota who were either incarcerated or on probation.

Table I. Descriptive Information.

Study	N	N _{5-year}	Static-99R M (SD)	Country	Recidivism criteria	Type of sample	Mostly treated	Release period	Mdn year release
Bartosh et al. (2003)	186	90	3.3 (2.9)	United States	Charges	Routine correctional	—	1996	1996
Bigras (2007)	480	206	2.1 (2.4)	Canada	Charges	Routine CSC	Mixed	1995-2004	1999
Boer (2003)	299	299	2.8 (2.8)	Canada	Conviction	Routine CSC	—	1976-1994	1990
Craissati et al. (2011)	209	200	2.2 (2.3)	United Kingdom	Conviction	Routine community supervision	Mixed	1992-2005	1998
Eher et al. (2008, 2009)	706	151	2.3 (2.3)	Austria	Conviction	Routine European prison	—	2000-2005	2003
Epperson (2003)	177	150	2.5 (2.6)	United States	Charges	Routine correctional	—	1989-1998	1995
Hanson et al. (2007)	702	—	2.4 (2.4)	Canada	Charges	Routine community supervision	—	2001-2005	2002
Långström (2004)	1,278	1,278	2.0 (2.4)	Sweden	Conviction	Routine European prison	No	1993-1997	1995
Total	4,037	2,374	2.3 (2.5)	—	—	—	—	1976-2005	1997

Note: CSC = Correctional Service Canada (administers all sentences of at least 2 years). Average Static-99R was computed using the largest sample size (N). Sample includes all cases available for Cox regression with sample as strata; three cases were deleted because the total follow-up time was less than the time to first observed recidivism event. Thirty-one cases from Hanson et al. (2007) excluded from all 5-year analyses because the 5-year recidivism rate was zero.

Hanson, Harris, Scott, and Helmus (2007). This prospective study followed offenders on community supervision between 2001 and 2005 in Canada, Alaska, and Iowa although only Canadian offenders were used in the current study.

Långström (2004). The study examined sex offenders released from prison in Sweden.

Recidivism

All samples used official criminal records to measure recidivism (four samples used charges; four used convictions). Although experts disagree on which definition is preferable for recidivism research, previous analyses with these samples did not find consistent differences in recidivism rates based on whether charges or convictions were used (Helmus, 2009). Either definition underestimates the true rate of recidivism. An additional concern with sexual recidivism research is that offenders can be charged with nonsexual offences for crimes that were actually sexual in motivation (e.g., plea bargains, sexually motivated abductions, homicide). In the current study, only two samples (22.6% of the total sample) obtained descriptions of the circumstances of the reoffence such that all sexually motivated offences were counted as sexual recidivism (Craissati, Bierer, & South, 2011; Hanson et al., 2007).

Plan of Analysis

Risk ratios. Relative risk was indexed using rate ratios, odds ratios, and hazard ratios. Rate ratios were calculated for adjacent Static-99R scores as follows:

$$\text{Rate ratio} = \frac{n_2/N_2}{n_1/N_1} = \frac{p_2}{p_1}$$

where N_i is the number of offenders with a specific score, and n_i is the number of these offenders known to have sexually reoffended. The p notation refers to probabilities or the number of recidivists divided by the number in that group.

Odds ratios are defined as $[p_2/(1-p_2)]/[p_1/(1-p_1)]$. Odds ratios were estimated through logistic regression or calculated by hand. Logistic regression is a form of regression in which the dichotomous dependent variable (recidivism) is transformed into log odds (Hosmer & Lemeshow, 2000). With one predictor variable (Static-99R), logistic regression estimates two regression coefficients (B_0 and B_1). B_1 is an estimate of relative risk; specifically, the average change in recidivism rate for each one-unit increase in risk scores is expressed as a log odds ratio. The exponent of B_1 (e^{B_1}) is an odds ratio. B_0 is an estimate of the recidivism base rate for offenders with a score of zero. For ease of presentation, Static-99R total scores were recoded such that zero represents the median value in routine samples (i.e., “2”; Hanson et al., 2012). As such, the B_0 is an estimate of the recidivism base rate for offenders with a score of 2 (see Helmus & Hanson, 2011, for application and review).

To link the logistic regression recidivism estimates to any specific time period, fixed follow-up periods are required (e.g., for a fixed 5-year recidivism outcome). Only offenders with at least 5 years of follow-up were included in the logistic regression analyses. Recidivism occurring after 5 years was excluded from these analyses to create an equal length of follow-up for each offender. Consequently, the sample size available for logistic regression is smaller than for statistics that allow varying follow-up (e.g., Cox regression).

In the context of the current study, hazard rates assess the risk, at a particular moment, that individuals will sexually reoffend if they have not already done so. The hazard rate is based on the probability of event occurrence during a particular time. Technically, it is defined as follows (Aalen et al., 2008, equation 1.2):

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

where the numerator is the conditional probability of the event happening in a particular time interval (Δt) provided that it has not already happened. This quantity is then divided by that same time interval (Δt) to create a rate. The hazard rate itself is defined as the value of this rate as the time interval becomes increasingly small (i.e., the limit as Δt approaches zero). A hazard ratio is the ratio of hazards for offenders with different scores.

Cox regression is the most widely used method of estimating hazard ratios from survival data (Singer & Willett, 2003, p. 14). This versatile method estimates averaged hazard ratios that can control for differences in recidivism base rates and differences in the shape of the survival functions across samples. To allow for this in the current study, all Cox regression analyses included sample as strata. The hazard ratio (HR) for adjacent Static-99R scores was computed as $HR = e^{\beta \times \text{Score}}$. In this formula, β is the Cox regression coefficient, and “score” is the score on the measure. Given that β is a constant, this approach assumes that each Static-99R point increases the risk of recidivism by a fixed and constant amount (using an exponential scale). For example, if the hazard ratio for adjacent scores was 1.3, then the ratio of hazards for scores that were three points apart would be $(1.3)^3 = 1.3 \times 1.3 \times 1.3 = 2.2$.

Fit. Both logistic regression and Cox regression models start with the assumption that relative risk increases by a fixed and equal amount for each one-unit increase in the predictor. The validity of this assumption can be tested by examining whether a curve component improves the fit between observed and predicted values (change in deviance index χ^2 , defined as $-2[\log \text{likelihood}]$) when the square of the scores is added after entering the scores in the regression model.

The Hosmer and Lemeshow test (Bewick, Cheek, & Ball, 2005; Hosmer & Lemeshow, 2000) was used to assess the fit between the observed recidivism rates and the recidivism rates predicted by the logistic regression (logistic curve) and by Cox regression (exponential curve). A nonsignificant Hosmer and Lemeshow test suggests good fit between the observed and expected recidivism rates. For logistic regression,

the predicted values tested were those provided by the fitted logistic curve. Cox regression, however, does not directly produce recidivism rate estimates (only estimates of relative risk); consequently, testing the fit of the Cox regression model required selecting a base rate. In our analyses, we used the 5-year sexual recidivism rate for a score of 2 (which was estimated from logistic regression) as the base rate. This base rate was then multiplied by the hazard ratios in Appendix A to create the exponential function implied by the Cox regression analyses.

Centering Static-99R on different scores does not influence the substantive findings of either Cox regression or logistic regression. Cox regression coefficients are invariant to linear transformation of the predictors. Logistic regression coefficients change based on linear changes in scaling of the predictors, but the point estimates for particular scores (and their standard errors) are identical regardless of where the scale is centered (i.e., scaling from -5 to 10 provides the same results for each score as scaling from -3 to 12).

Stable across time. Schoenfeld residuals (\hat{S}) were used to formally verify that the hazard ratios are constant over time (i.e., the proportional hazard assumption in Cox regression; Singer & Willett, 2003). Schoenfeld residuals compared the observed and expected values of Static-99R scores:

$$\hat{S} = x_i - \text{Expected}[x]$$

Where x_i is the observed Static-99R score for person i who reoffended at t_i minus the expected Static-99R score for the risk set at t_i (Expected[x]). Expected[x] is a weighted average of expected Static-99R score by each individual's likelihood of reoffending at t_i . \hat{S} values are only computed for recidivists, but nonrecidivists contribute to the computation if they fall in a recidivists' risk set. If the proportionality assumption is violated, Schoenfeld residuals will be related with time to recidivism.

Meta-analysis. Univariate fixed-effect and random-effects meta-analysis was used to examine variability across samples (Borenstein, Hedges, Higgins, & Rothstein, 2009). In particular, Cochran's Q statistic was used to test whether the amount of between-sample variability was more than would be expected by chance, and the I^2 statistic was used to quantify the amount of variability, defined as $I^2 = [(Q - df)/Q] * 100$ for univariate meta-analysis. According to Higgins, Thompson, Deeks, and Altman (2003), I^2 values of 25%, 50%, and 75% are low, moderate, and high, respectively.

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

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

Whereas Cox regression models produce one parameter (i.e., a hazard ratio), logistic regression models produce two parameters (i.e., an odds ratio and a base rate). As such, multivariate meta-analysis was used to examine variability in the logistic regression model parameters. Specifically, multivariate meta-analysis allows for the analyses of multiple parameters while accounting for their correlation. Such an approach reduces the Type 1 (i.e., finding significance when none exist) error rate (Bagos, 2011). The MVMETA package for R program (Gasparrini, 2011) was used to conduct multivariate meta-analysis. This package can conduct both fixed-effect and random-effects analyses and produces the multivariate Q test while accounting for the relationship between parameters (r). For multivariate meta-analysis, the I^2 statistic is defined as $I^2 = [(Q - [1-r] - df)/Q]*100$ (Bagos, 2011).

Results

Assumption 1: Each Increase in Static-99R Score Is Associated With a Consistent Increase in Relative Risk of Approximately 30%

Table 2 presents the sexual recidivism rates for each Static-99R score at 5 years (fixed) follow-up ($n = 2,374$) along with the rate ratios and odds ratios for adjacent scores. To demonstrate how these statistics were calculated, consider the recidivism rate of offenders with a score of 1 ($p_1 = .029$) and the recidivism rate for offenders with a score of 2 ($p_2 = .040$). Using these rates, the rate ratio of these two adjacent scores was $p_2/p_1 = .040/.029 = 1.38$. In other words, offenders in our samples with a score of 2 had 1.38 times the recidivism rate of offenders with a score of 1 (see Table 2). Continuing with the current example, the odds of recidivism for offenders with a score of 1 was $p_1/(1-p_1) = .029/(1-.029) = .0299$, and the odds of recidivism for offenders with a score of 2 was $p_2/(1-p_2) = .040/(1-.040) = .0417$. Consequently, the odds ratio for these two adjacent scores was $.0417/.0299 = 1.39$, meaning that the odds of recidivism for offenders in our samples with a score of 2 was 1.39 times the odds of recidivism for offenders with a score of 1. In Table 2, rate ratios and odds ratios were not computed when one of the two observed recidivism rates was zero.

The shape of the function linking Static-99R and sexual recidivism appeared exponential for all scores up to and including 9 (see Figure 1). The observed recidivism rate for offenders with scores of 10 or 11 were lower than expected, but the sample sizes for these scores were too small for reliable estimates ($n = 7$ and $n = 1$, respectively). Overall, the rate ratios for adjacent scores ranged between 0.80 and 2.30, and odds ratios ranged between 0.72 and 2.52. The weighted means for both the rate ratios (1.31) and odds ratios (1.35) were extremely close to the expected value (1.3). Logistic regression was used to formally test how well the relationship between Static-99R and sexual recidivism fit a logistic function with a fixed 5-year follow-up ($n = 2,374$, without nesting). The Static-99R model had a B_1 of .325 ($SE = .035$) and a B_0 of -3.09

Table 2. Rate Ratios and Odds Ratios for Static-99R Predicted 5-Year Sexual Recidivism.

Static-99R score	Recidivism (%)	Number of recidivists/total	Rate ratio	Odds ratio
-3	0.0	0/37	—	—
-2	0.0	0/65	—	—
-1	2.7	7/260	—	—
0	2.4	7/290	0.90	0.89
1	2.9	10/342	1.21	1.22
2	4.0	14/348	1.38	1.39
3	5.9	20/338	1.47	1.50
4	6.2	17/274	1.05	1.05
5	14.3	27/189	2.30	2.52
6	14.0	15/107	0.98	0.98
7	16.0	12/75	1.14	1.17
8	29.6	8/27	1.85	2.21
9	35.7	5/14	1.20	1.32
10	28.6	2/7	0.80	0.72
11	0.0	0/1	—	—
12	—	—	—	—
	6.1	144/2,374	1.31 ^a	1.35 ^a

Note: The average odds ratio calculated from logistic regression was 1.38 ($B_1 = .325$; $SE = 0.035$).

^aAverages were weighted using sample size of the highest score.

($SE = .111$). These parameters translate to an average odds ratio of 1.38 (95% CI [1.29, 1.48]) and a base rate of 4.35% (95% CI [3.53, 5.35]) for a score of 2. The Hosmer and Lemeshow test was nonsignificant ($\chi^2 = 4.29$, $df = 6$, $p = .637$), indicating good fit between the data and the logistic function. Allowing a curve component did not significantly improve the model, χ^2 change = 0.01, $df = 1$, $p = .945$.

Cox regression was used to formally test how well the relationship between Static-99R and sexual recidivism fit an exponential function. The Static-99R model based on all cases (varying follow-up; $n = 4,037$, sample as strata) had a β of .332 ($SE = .022$), which translates to an average hazard ratio of 1.39 (95% CI [1.33, 1.46]). Allowing a curve component did not significantly improve the model, χ^2 change = 0.01, $df = 1$, $p = .906$. Restricting the Cox regression analysis to the same fixed 5-year period as the regression analysis produced similar results: the Static-99R model had a β of .308 ($SE = .032$), which translates to an average hazard ratio of 1.36 (95% CI [1.28, 1.45], $n = 2,373$). Once again, allowing a curve component did not significantly improve the model, χ^2 change = 0.27, $df = 1$, $p = .600$. The Hosmer and Lemeshow test for the hazard rate exponential function (excluding the curve component) was nonsignificant ($\chi^2 = 5.02$, $df = 6$, $p = .541$), indicating good fit between the data and the exponential function.

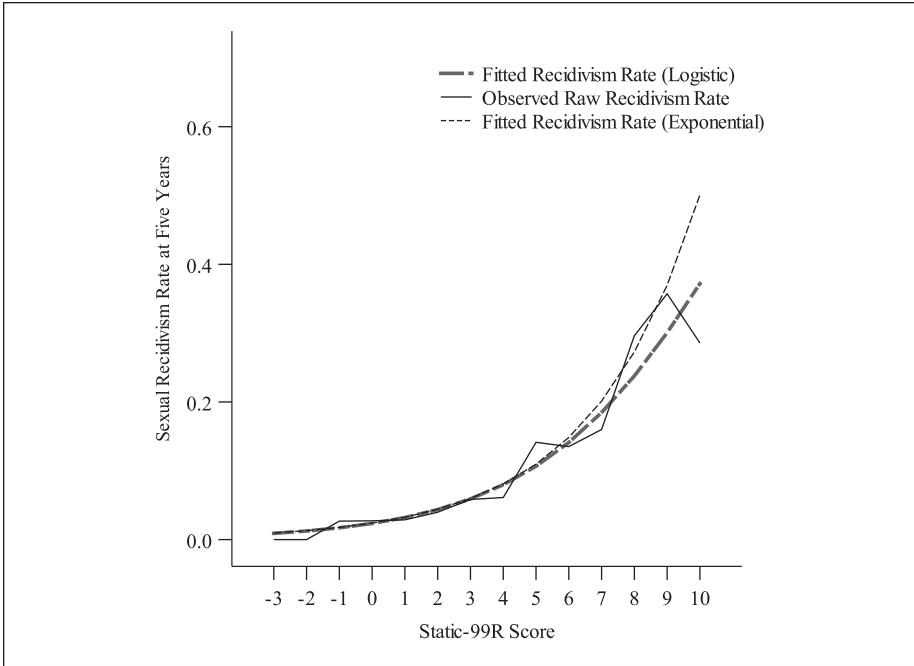


Figure 1. Observed 5-year sexual recidivism rates plotted along with the 5-year sexual recidivism rates derived from logistic and Cox regression analyses.

Figure 1 presents a graph of the observed 5-year recidivism rate and the probabilities estimated from the logistic regression and Cox regression analyses (using a 5-year period). The predicted recidivism rates from logistic regression were calculated by adding the intercept (B_0 , the predicted value for a score of 2) to the product of the slope (B_1) and the independent variable (x , the Static-99R score): $y = B_0 + x(B_1)$. Of note, the calculations are in log odds, or *logit* units. The following equation can be used to transform logits into probabilities: $p = 1/(1 + e^{-LOGIT})$. For example, the logistic regression estimate of the 5-year recidivism rate for a score of 5 was $-3.09 + (5 - 2) * (.325) = -2.115$ in logit units, or $10.8\% = 1/(1 + e^{2.115})$. The logistic regression coefficients used to create Figure 1 were based on 5-year recidivism data without nesting.

To compute the recidivism rates based on the Cox regression analyses, a base rate and the hazard ratios are required. For the base rate, we used the 5-year sexual recidivism rate estimated from logistic regression for a score of 2 ($B_0 = -3.09$, or 4.35%). This base rate was then multiplied by the hazard ratios in Appendix A to produce the expected recidivism rates implied by the Cox regression analyses (e.g., a score of 5: $4.35\% \times 2.70 = 11.7\%$). The one case with a score of 11 was not plotted in Figure 1 but was included in all analyses.

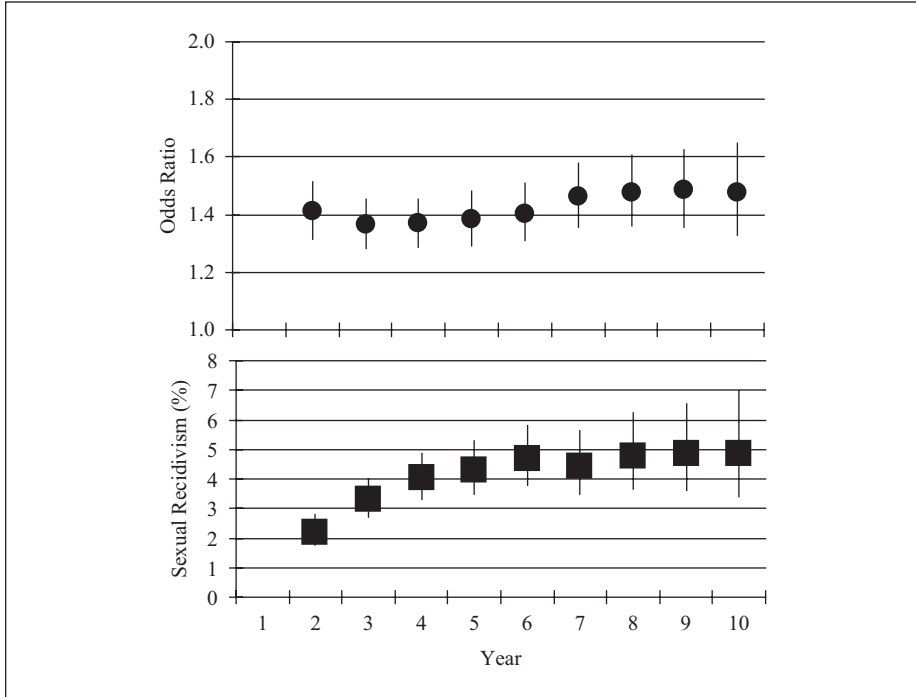


Figure 2. Stability of odds ratios and base rates of sexual recidivism derived from logistic regression across follow-up time.

The logistic and exponential fitted curves were virtually identical between Static-99R scores of -3 and 6 , which is expected given the similarity in regression coefficients (0.325 and 0.332) and the relatively low base rates. For the very highest scores (7 or more), the exponential curve estimated slightly higher values than did the logistic curve.

Assumption 2: Stable Over Time

To examine the stability of relative risk over time, the relative risk and base rates derived from logistic regression were calculated in 1-year increments for fixed follow-up periods from 2 to 10 years (see Figure 2). The coefficients were transformed into odds ratios and percentages for ease of interpretation. As expected, the base rate of sexual recidivism increases as the follow-up time increased. In contrast, the odds ratios did not show meaningful variation based on the length of follow-up times.

Schoenfeld residuals (\hat{S}) were used to formally test the proportional hazard assumption in Cox regression (i.e., that the hazard ratios are constant over time; Singer & Willett, 2003). As can be seen in Figure 3, no observable patterns between \hat{S} and time

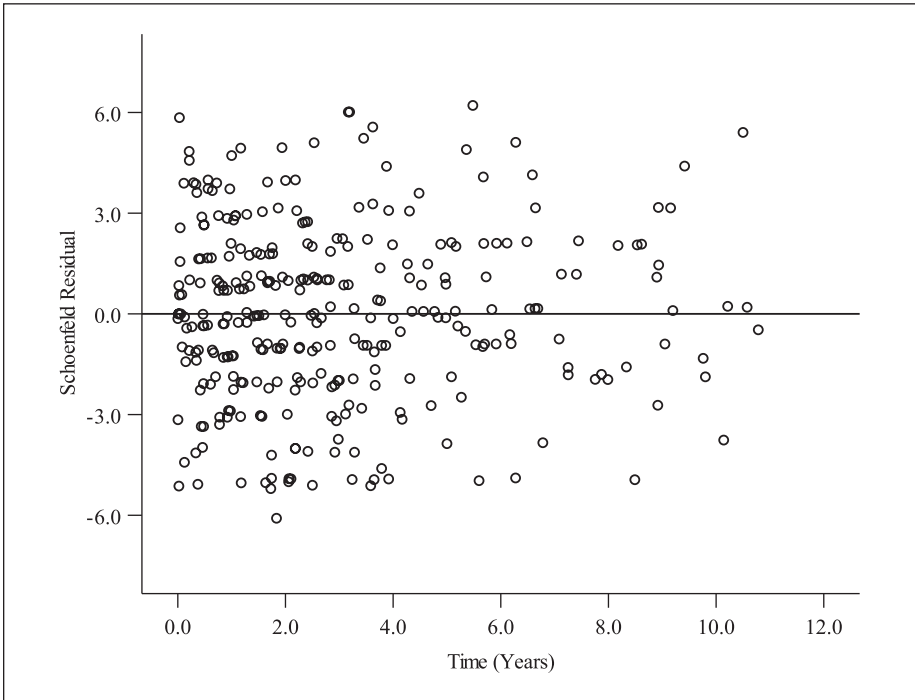


Figure 3. Schoenfeld residuals plotted as a function of time.

to sexual recidivism emerged, and the correlation was not statistically significant ($r = .026$; $n = 308$, $p = .652$). As such, the formal statistical test from Cox regression confirmed the consistency of relative risk across time.

Assumption 3: Stable Across Samples

To assess the stability across samples of relative risk indicators, risk ratios were calculated for each sample based on both hazard ratios from Cox regression (with complete follow-up times) and odds ratios from logistic regression (for fixed 5-year follow-up; see Appendix B). For Cox regression, the fixed-effect average β was 0.332 (95% CI [0.288, 0.375]) and random-effects average was 0.330 (95% CI [0.282, 0.379]), $k = 8$, $N = 4,037$. These translate into hazard ratios of 1.394 and 1.391, respectively. The Cochran Q statistic was nonsignificant ($Q = 8.19$, $df = 7$, $p = .316$) indicating no more variability than would be expected by chance and the amount of between-study variance was small ($I^2 = 14.5\%$).

Table 3 presents the meta-analytic summary of the logistic regression analyses predicting 5-year sexual recidivism for both the B_0 (base rate) and B_1 (odds ratio). As with the hazard ratios, the estimated odds ratio was 1.38 and 1.39 (fixed effect and

Table 3. Meta-Analysis of Logistic Regression Parameters From Static-99R Predicting 5-Year Recidivism Rates.

	Fixed-effect			Random-effects			K	N	Q	I ²
	95% CI			95% CI						
	M	LL	UL	Mn	LL	UL				
Odds ratio	1.38	1.29	1.48	1.39	1.28	1.52	7	2,374	7.39	18.84%
Base rate (%)	4.96	4.02	6.09	5.02	3.21	7.77	7	2,374	17.92**	66.52%

Note: Parameters transformed into odds ratio and percentages for ease of interpretation. Base rates represent average recidivism rate for a Static-99R score of 2.

***p* < .01.

random-effects estimates, respectively) and there was no significant variability in B_1 across samples ($Q=7.39, df=6, p=.286; I^2=18.8\%$). There was, however, significant variability in the base rates of recidivism, which ranged from 0.3% to 11.4% for a score of 2 ($Q=17.92, df=6, p=.006; I^2=66.5\%$). The multivariate meta-analysis similarly showed significant variability in the two-parameter regression equation ($B_0 + B_1$) across samples ($Q=25.18, df=12, p=.014; I^2=44.6\%$).

Risk Ratios for Static-99R Scores

To use risk ratios in applied risk communication, there needs to be an explicit method of connecting them to Static-99R scores. Of the various options available, we used the hazard ratio for sexual recidivism calculated by Cox regression using continuous Static-99R scores ($\beta = 0.3318; SE = .022$), with sample as strata ($k = 8, n = 4,037$). The Cox regression estimate for the full data set was privileged over the other possible estimates because it was based on the largest sample size. All the other estimates (e.g., random effects-meta analysis, logistic regression), however, provided essentially equivalent values.

The next decision concerned the choice of the reference category. Given that risk ratios are indicators of relative risk, they only have meaning when comparing scores. We selected the median value (Static-99R score of 2; Hanson et al., 2012) as the standard reference category; consequently, the relative risk values for specific scores can be interpreted as the extent to which the offender’s risk for sexual recidivism is higher or lower than that of sexual offenders with a score in the middle (midpoint) of the risk distribution. Based on the above assumptions, a table with the risk ratios (hazard ratios) for each Static-99R score is presented in Appendix A. Specifically, the hazard ratio for the reference category (Static-99R score of 2) was set at 1. Then, for each unit increase from the reference category, the hazard ratio was multiplied by 1.393. For example, the score of 3 was one unit higher than a score of 2, so $1*(1.393) \approx 1.39$. For a score of 4, which is two units above the reference category, the hazard ratio was calculated as $1*(1.393)*(1.393) = 1.393^2 \approx 1.94$. When scores were lower than the

reference category, then hazard ratios were calculated by dividing 1 by 1.393 for each unit below the reference category. For example, the hazard ratio for a score of -1 (3 units below the reference category) was $1/[(1.393)^*(1.393)^*(1.393)]$ or, more simply, $e^{(-3)\ln(1.393)} \approx 0.37$.

Discussion

The current study examined the extent to which Static-99R scores can be used to represent differences in relative risk among sex offenders. Based on 8 routine samples ($N = 4,037$), the observed recidivism rates could be well represented by a statistical model in which a one-unit increase in Static-99R score indicated an increase in recidivism risk of approximately 1.4 for sexual recidivism (using either an exponential model or a logistic model). This value is very close to the expected value of 1.3, which was the heuristic used in developing Static-99R. The rate of increase in relative risk was consistent across the time periods (2 to 10 years) as well as across the samples examined in the current study.

In general, there has been little attention to measurement models for actuarial risk tools. Some researchers have argued for a cumulative stochastic model (Aalen et al., 2008; Hammond & O'Rourke, 2004). Specifically, stochastic models assume that the relationship between the items and the outcome is probabilistic, such that the relationship will always contain some amount of inherent randomness (i.e., the estimated recidivism rates will never be zero or one). The model is cumulative in that each item is assumed to contribute incrementally to increased risk. Our results provided strong support that the cumulative stochastic model fits the relationship between Static-99R scores and recidivism risk. The main contenders to the cumulative stochastic model are models that specify meaningful *if-then* interactions between key variables, such as decision tree models (Steadman et al., 2000), or interactions between risk factors and age (Lussier & Davies, 2011). Although plausible, statistical interactions are notoriously difficult to empirically identify and to replicate (Busemeyer & Jones, 1983; Russell & Bobko, 1992).

The stability in relative risk contrasts with the variability in absolute recidivism rates observed in the current study. This pattern (consistent relative risk; variable absolute risk) has been observed in a larger set of 23 Static-99R recidivism studies (Helmus, Hanson, et al., 2012) as well as in other applied prediction problems, including predicting car accidents (Ingre et al., 2006; Rosén, Stigson, & Sander, 2011) and suicide (Paterson et al., 2008). It is likely a feature of all functional, but imperfect, risk tools. Base rate variability would be expected to decline as the factors that influence the base rate are identified and incorporated into risk tools.

Communicating "Risk" Accurately

This study was motivated by the need to find nonarbitrary metrics for risk communication. Forensic evaluators and decision makers prefer to communicate risk using

nominal categories (Grann & Pallvik, 2002). However, we commonly disagree with what “low,” “moderate,” and “high” risk actually represent (e.g., Hilton, Carter, Harris, & Sharpe, 2008; Monahan & Silver, 2003). Research has found that the addition of numeric indicators lead to fewer interpretation errors compared to the use of nominal risk categories alone (e.g., Karelitz & Budescu, 2004). But what numbers should we use? Absolute recidivism rates are clearly important; however, they are difficult to estimate with accuracy. Base rates vary across samples and settings for reasons that are not fully understood (Helmus, Hanson, et al., 2012). Another important consideration is that the observed recidivism rates underestimate the actual recidivism rates because many sexual offences are never detected. There is no scientific consensus on the extent of the underestimation. Based on these concerns with absolute recidivism rates, it is worth considering other numeric indicators for risk communication. The results of the current study suggest that risk ratios are a fundamental feature of actuarial risk scores and could provide a stable reference point for communicating the results of actuarial risk assessment.

The utility of risk ratios for applied risk assessments, however, has yet to be established. To our knowledge, this study is the first to present risk ratios for any of the commonly used forensic risk assessment tools. Consequently, evaluators using risk ratios need to exercise special care because the concepts and terminology are likely to be unfamiliar to their audiences. Fortunately, this is not a new problem. Risk ratios are rare in forensic assessments, but they are standard practice in medical risk communication (e.g., “The risk of developing lung cancer is about 23 times higher among men who smoke cigarettes . . . compared with never smokers,” Centers for Disease Control and Prevention, 2012).

Risk communication between physicians and patients has been extensively studied, and established guidelines are available (Fischhoff, Brewer, & Downs, 2011). Patients are likely to have the clearest and least biased understanding of risk when (a) a variety of numeric indicators are used (e.g., absolute risk, relative risk; Karelitz & Budescu, 2004), (b) the information is presented in different formats (e.g., categories, numbers, and pictures; Burkiewicz, Vesta, & Hume, 2008; Dansereau & Simpson, 2009; Kurz-Milcke, Gigerenzer, & Martignon, 2008; Visschers, Meetens, Passchier, & de Vries, 2009), and (c) clear instructions are provided on how to interpret the risk information (Evans, Handley, Perham, Over, & Thompson, 2000; Gigerenzer & Hoffrage, 1995; Kahneman, 2003; Nisbett, Krantz, Jepson, & Kunda, 1983). It is quite likely that the same principles apply to risk communication between forensic evaluators and decision makers (see Babchishin & Hanson, 2009, for review).

The current article examined risk ratios, but they are not the only metric for quantifying relative risk. Percentile ranks are another plausible metric and are commonly used for quantifying individual differences in psychology (e.g., IQ, *T*-scores; Crawford & Garthwaite, 2009). Evaluators and decision makers should be aware, however, that percentile ranks and risk ratios provide different information; consequently, an offender identified as high risk based on a risk ratio threshold (e.g., twice as likely to

reoffend as the midpoint) may or may not be identified as high risk based on a threshold defined by percentile ranks (e.g., $Z > 1$; greater than one standard deviation above the mean). In general, the relationship between percentile ranks and risk ratios would not be expected to be linear, and is likely to be different for different risk scales.

The relative usefulness of these metrics depends on the context of the assessment. Percentile ranks are well suited to decisions concerning the distribution of finite resources (e.g., police prioritizing offenders for registration compliance, determining which of two offenders should receive a home visit today from a probation officer). In contrast, risk ratios are well suited to decisions concerning matching the intensity of interventions to the degree of riskiness (e.g., matching the length of treatment to initial risk levels).

Risk ratios are also well suited to quantifying and comparing the results of different risk scales (Babchishin & Hanson, 2009). Evaluators often use different risk tools (Jackson & Hess, 2007) and risk tools do not always provide the same results. For example, Barbaree, Langton, and Peacock (2006) found that less than 8% ($n = 20$) of sex offenders sampled ($N = 262$) were consistently identified as high risk or as low risk by five commonly used actuarial risk tools. This is perhaps not surprising considering that the authors of risk scales rarely specify the basis for their nominal risk categories. If risk ratios are, as we have argued, a fundamental property of risk scales, then risk ratios are a better metric than percentile ranks for combining the information contained in different risk tools. Percentile ranks measure unusualness and are directly influenced by the distribution of scores. In contrast, risk ratios quantify how much riskier an offender is compared to a reference category.

The choice of the reference category is an important decision for the interpretation of relative risk models, as it can substantially influence the risk ratios associated with specific scores. Although evaluators are free to choose any reference category appropriate to their communication needs, the consistent use of the same reference category (whatever it is) facilitates communication across settings and jurisdictions. Of the various options, the median has certain advantages. First, it has a substantive meaning, namely, it is the middle value of a routine distribution of Static-99R scores. Second, in contrast to the mean ($M = 2.1$; Hanson et al., 2012), it is a real (observable) value of the risk scale. Furthermore, the median values have been found to be stable (the same) in international comparisons of routine samples in Canada, Sweden, and California (Hanson et al., 2012). One disadvantage of the median, however, is that it is easy to confuse (a) the recidivism rate of offenders with the median score, with (b) the average recidivism rate of the sample. Given that risk is distributed exponentially, the recidivism rate associated with the median value would be expected to be consistently lower than the average (overall) recidivism rate (and, more precisely, by a factor of 0.71, see Appendix C). Evaluators should also be aware that the recidivism rate expected for the average score in a sample should be lower than the overall recidivism rate in that sample assuming that the distribution of Static-99R scores is continuous and the distribution of risk associated with these scores is exponential.

Special Considerations When Using Risk Ratios

To use risk ratios to compare risk tools, two types of information are required: (a) the rate ratios for that scale, and (b) a nonarbitrary reference category that can be defined for all the measures being compared. In the current study, the distribution of risk was exponential, such that it could be modeled with a single parameter (i.e., an increase of 39% per score). A smooth exponential function, however, is not a requirement for using risk ratios to quantify risk. Risk ratios can provide useful information regardless of the shape of the relationship between scores and risk, as long as that shape is reasonably stable across samples, settings, and time. What is required to compare risk ratios is a consistent, nonarbitrary reference category. As previously discussed, a plausible reference category would be the median score for unselected, routine populations. For example, a Static-99R score of 5 indicates a 2.7 increase in relative risk compared to sexual offenders at the midpoint of the risk distribution (median score of 2; Appendix A), whereas a Static-2002R score of 6 would indicate an equivalent increase in relative risk (a risk ratio of 2.6) from the middle (i.e., median) Static-2002R score of 3 (Babchishin, 2011, see also www.static99.org). Administrators and evaluators may want to consider such risk ratios when judging the equivalence of thresholds for different risk tools. Risk ratios, however, are only one indicator that should be considered. In general, judgments concerning equivalence of risk scales should be linked to the decisions at hand. If the decision is based on absolute risk (e.g., civil commitment), then equivalent test scores would be scores that provide equivalent estimates of absolute recidivism risk.

The best methods for estimating absolute recidivism rates from actuarial risk scores remains an active topic of debate (Babchishin, Hanson, & Helmus, 2012; Seto, 2005; Thornton, Hanson, & Helmus, 2010; Vrieze & Grove, 2010). The basic dilemma is that, as research progresses, variables are found that are incremental to any particular scale, and these incremental variables are known to evaluators and decision makers well before they are formally included in risk assessment schemes (Hanson, 1998). Static-99R, in particular, was never intended to measure all relevant risk factors, and it is easy to find studies in which external variables predict incrementally to Static-99/R scores (e.g., McGrath, Lasher, & Cumming, 2012; Olver, Wong, Nicholaichuk, & Gordon, 2007; Thornton, 2002). Consequently, evaluators must distinguish between the risk presented by Static-99R scores alone and their global evaluation of risk based on all sources of information.

When considering risk attributable to Static-99R scores, the best source of information concerning absolute recidivism rates would be large validation studies (>100 recidivists) that are both local and recent (ideally updated annually). Such studies would provide credible estimates of both the local base rate and the sample-specific risk ratios. When such studies are not available, then evaluators can choose the norms created from other samples that most closely approximate the case-at-hand (see www.static99.org for details).

There are some circumstances, however, in which it would be possible and desirable to calculate absolute recidivism rates using the risk ratios presented in the current study and independent, local base rate estimates. Such estimates should be considered when (a) there is doubt about the applicability of norms based on other settings, (b) a credible, local validation study has not been conducted, (c) there is a reliable local estimate of the overall sexual recidivism rate, (d) the sample can be assumed to be routine (i.e., not preselected on risk relevant variables), and (e) only a small proportion of the sample (<10%) would be expected to have recidivism rates greater than 50%. In such circumstances, it is not unreasonable to multiply the risk ratio (in whatever form) with an estimate of the recidivism rate to get a plausible estimate of absolute risk (see Appendix C).

The above procedure requires, of course, knowing the recidivism rate of a reference category. Fortunately, this is computationally simple when the distribution of Static-99R scores is routine and the overall recidivism rate is known. In such situations, the recidivism rate for a score of 2 can be estimated by dividing the overall recidivism rate by 1.41 (see Appendix C, for details). Note that the recidivism rate for the median value is expected to be lower than the overall (average) recidivism rate because risk is distributed exponentially. In the current set of routine samples, this procedure resulted in good calibration between the observed and expected values (using either the exponential or logistic models).

The present results also suggest that, in the absence of local norms, it is hard to anticipate base rates. The current samples were selected because they had not been deliberately preselected on risk-relevant variables. Nevertheless, there was significant and meaningful variation in the recidivism base rates in these "routine" samples. The reasons for this variation are not fully known. Preliminary analyses of a larger set of Static-99 studies suggests that relatively little of the between-study variation can be attributed to jurisdictions (e.g., United States vs. Canada) or to experimenter-controlled design variables (e.g., charges vs. convictions; Helmus, 2009). Samples that have been preselected on risk-relevant variables (e.g., Bengtson, 2008; Knight & Thornton, 2007) have higher base rates than routine samples (Helmus, 2009), but it is not obvious how preselection effects would influence the current set of samples, all of which could be considered routine.

The current study only examined risk ratios for sexual recidivism in routine samples of sexual offenders. The generalizability of these findings to other outcome criteria (e.g., any violent recidivism) or to nonroutine samples has not been established. In particular, both theory and preliminary analyses suggest that a consistent exponential distribution is unlikely to well-represent the recidivism data when a substantial portion (>10%) of the sample have expected recidivism rates of greater than 50% (e.g., predicting violent recidivism among offenders preselected as high risk). For example, assuming a risk ratio of 1.4 and base rate of 55%, a 1-point increase would lead to a predicted value of 77% and a 2-point increase would lead to an (impossible) predicted value of 108%. Evaluators should also not assume that the risk ratios observed in the

current study are identical for all types of recidivism. To the extent that different risk factors predict different types of outcomes, then variation in the Static-99R risk ratios would be expected for different recidivism criteria (e.g., violent recidivism, any recidivism).

The current study only examined officially recorded recidivism, which underestimates the true recidivism rate. Confidential interviews with sexual offenders often find 5 to 10 undetected sexual offences for every offence resulting in a conviction (Abel & Harlow, 2001; Hindman & Peters, 2001; Weinrott & Saylor, 1991). We do not, however, believe that the observed base rates should be multiplied by a factor of 5 or 10. Even when the detection rate per offence is low, the detection rate per offenders could be much higher because those who commit many offences are likely to eventually get caught. It is also worth noting that the most serious offences are those most likely to be reported to the police (Fisher, Daigle, Cullen, & Turner, 2003). Consequently, officially recorded recidivism should be considered a measure of the severity and frequency of sexual offending.

The current study found orderly, exponential relationships between Static-99R risk scores and officially recorded sexual recidivism, which is the criterion Static-99R was designed to predict. Further research is needed to examine the utility of relative risk ratios for modeling unofficial sexual recidivism. Treatment programs of self-referred pedophiles provide one opportunity for conducting such research. For example, preliminary results of Dunkenfeld Project in Germany (Beier et al., 2009) found that despite child pornography use declining with treatment, the rates of self-reported child pornography use continued to be high (Kuhle et al., 2012). In contrast, no recidivism was identified by official records.

Implications for Applied Risk Assessment

Evaluators interpreting Static-99R should be aware that relative risk is an intrinsic and stable property of the risk scores. Although the utility of risk ratios in forensic risk assessment has yet to be established, it is a promising, nonarbitrary metric for quantifying the information provided by actuarial risk tools. Using Appendix A, for example, an offender with a Static-99R score of 0 could be described as being a member of a group whose expected rate of sexual recidivism is approximately half that of sex offenders at the midpoint of the risk distribution (defined as a median score of 2). Similarly, an offender with a score of 7 could be described as a member of a group whose expected rate of sexual recidivism is approximately 5 times that of sex offenders at the midpoint of the risk distribution. Given that the use of risk ratios is uncommon in forensic risk communication, evaluators should exercise special care that they are properly understood. In particular, relative risk information should always be presented with reference to base rates because risk ratios result in overestimation of risk when base rates are ignored (Elmore & Gigerenzer, 2005). For example, when interpreting a risk ratio of 2.0, it makes considerable difference whether the base rate is 4.9% or 49%.

Our results also suggest that using Static-99R to estimate absolute recidivism rates should be conceptualized as requiring two separate estimates: an estimate of the offender's relative risk and an estimate of the recidivism base rate for the sample to which the offender should be compared. Of these two, the estimate of the base rate requires more justification than the relative risk estimate because base rates are less stable across samples and settings.

Given that risk communication improves when information is provided in multiple forms, we encourage evaluators using the STATIC risk instruments to present three different quantitative indicators: absolute recidivism rates, percentiles, and risk ratios (see templates available at www.static99.org). Specific reports, however, should emphasize the indicator most closely aligned with the referral question. If the question concerns allocation of finite resources, then focus on percentiles. Risk ratios, on the other hand, are the preferred metric for describing the "riskiness" of an individual in comparison to other sexual offenders. Risk ratios are also particularly useful when comparing the results of multiple risk scales. If the question concerns absolute recidivism rates, then neither percentiles nor risk ratios may be relevant (except as intermediate steps to the final result). When the evaluation question is not specified, or the report will be used for multiple purposes, we recommended that all three indicators are reported and interpreted.

Risk ratios are new to forensic psychology, and, consequently their use requires special efforts to educate decision makers. If the arguments in this article are correct, then these efforts are more than rewarded by the possibility of increased precision of risk communication. Percentile ranks do not inform decision makers about how well a measure discriminates between recidivists and nonrecidivists. A score in the top 10% of a valid risk scale could look the same as score in the top 10% of a untested (or invalid) measure. In contrast, decision makers would quickly recognize the difference between a risk ratio that is 5 times above the midpoint, and a risk ratio that is not meaningfully different from the midpoint (1).

Implications for Research

Given our penchant for nominal risk categories (e.g., "low," "moderate," "high"), it is important that these labels be clearly defined. Linking nominal labels to nonarbitrary numerical definitions, such as risk ratios, is one method to minimize misunderstandings. However, further research is needed on generating numerical indicators before they can be widely used in applied risk communication.

As a prerequisite for numeric indicators, test developers need to articulate the measurement models associated with actuarial risk scales. In general, these models would specify the expected relationship between scores and the risk-relevant information used by decision makers. Understanding these measurement models would help evaluators identify what is being assessed by risk tools, and could provide nonarbitrary metrics for risk communication. In particular, we recommend that researchers test the extent to which the cumulative stochastic/exponential models demonstrated in the current

study apply to other risk measures. Research is also needed to identify nonarbitrary reference categories (e.g., medians), which can then be the basis for comparing risk ratios, or other indicators of relative risk (e.g., percentiles), across different assessment tools.

There is considerable research on risk communication to the public, but less is known about risk communication among professionals. Consequently, there is a need for further research directed toward optimizing offender risk communication among evaluators and decision makers (judges, parole board, case managers). Evaluating risk communication requires specifying the superordinate goals of the assessment. Although the specific goals vary across context, there are also some common goals. For example, effective risk communication should result in decision makers understanding and remembering the information provided and should improve the quality of their subsequent decisions (e.g., increasing consistency and fairness, optimizing public safety).

Researchers interested in the psychometric properties of Static-99R may want to conduct additional studies examining the incremental contribution of individual items. The current analyses only focused on total scores, not items. Although the total scores were consistently associated with a risk ratio of 1.4, it is possible that certain items will show greater discrimination than others. Given that there are many different ways of increasing a score by 1 point, a one-unit increase in the total score represents the average contribution of different items, all of which are drawn from the same pool of 10 items. Consequently, the observed consistency in the risk associated with each point should not be surprising.

We doubt, however, that major improvements in predictive accuracy will be obtained by tweaking the item weights or by examining distinct combinations of individual items (Babchishin et al., 2012). Instead, we believe that the greatest potential for improving risk assessment is in developing reliable, valid, and nonarbitrary assessments of the constructs responsible for recidivism risk (Babchishin et al., 2012; Hanson, 2009; Mann et al., 2010). The empirical, actuarial risk assessment of the future should be using empirically derived weights for combining psychologically meaningful risk factors (Mann et al., 2010).

Conclusions

The current study found substantial support for a cumulative stochastic model to describe the relationship between Static-99R scores and sexual recidivism. Each increase in Static-99R score was associated with a stable and consistent increase in relative risk of approximately 40%. This model suggests that using Static-99R to estimate absolute recidivism rates necessarily involves separate estimates of (a) relative risk and (b) the base rate for the appropriate reference sample. We recommend that evaluators report risk ratios in their applied assessments; however, given that risk ratios may be unfamiliar to forensic decision makers, special care is required to ensure that information concerning relative risk clarifies rather than confuses risk communication. In particular, risk ratios should always be presented with base rate information.

Appendix A

Table A1. Risk Ratios (defined as Hazard Ratios) for Static-99R Scores (centered on 2) Routine Samples.

Sexual recidivism		
Static-99R score	Frequency (n)	Hazard ratio
-3	73	0.19
-2	105	0.26
-1	384	0.37
0	473	0.52
1	565	0.72
2	599	1.00
3	598	1.39
4	491	1.94
5	333	2.70
6	209	3.77
7	120	5.25
8+	87	7.32

Note: For sexual recidivism, hazard ratios based on Cox regression coefficients derived from entering the continuous (i.e., unclumped) Static-99R scores ($\beta = .3318$; $SE = .022$), with sample as strata ($k = 8$, $n = 4,037$). Due to small sample size, risk ratios are not presented for Static-99R scores greater than 8.

Appendix B

Table B1. Relative Risk Estimates per Sample.

Sample	Cox regression	Logistic regression (5-year sexual recidivism)	
	β (SE)	B_1 (SE)	B_0 (SE)
Bartosh et al. (2003)	0.163 (0.072)	0.137 (0.115)	-2.054 (0.367)
Bigras (2007)	0.295 (0.077)	0.399 (0.117)	-2.708 (0.321)
Boer (2003)	0.341 (0.078)	0.478 (0.139)	-4.412 (0.605)
Craissati et al. (2011)	0.312 (0.078)	0.341 (0.112)	-2.820 (0.332)
Eher et al. (2008, 2009)	0.434 (0.082)	1.013 (0.401)	-5.801 (1.522)
Epperson (2003)	0.367 (0.069)	0.347 (0.107)	-2.651 (0.371)
Hanson et al. (2007)	0.367 (0.054)	—	—
Långström (2004)	0.341 (0.041)	0.308 (0.050)	-3.118 (0.151)
Overall (random-effects meta-analytic average)	0.330 (0.025)	0.331 (0.044)	-2.941 (0.238)

Note: Cox regression used the complete follow-up ($N = 4,037$) whereas logistic regression used the 5-year follow-up ($n = 2,374$). Logistic regression B_0 centered on a score of 2. Logistic regression coefficients were not estimated for Hanson et al. (2007) because the 5-year recidivism rate was zero (0/31).

Appendix C

Estimating Absolute Recidivism Rates Associated with Static-99R Scores Using Only Risk Ratios and Observed Base Rates

The following procedure for estimating absolute recidivism rates associated with Static-99R scores would be useful when a local Static-99R validity study has not been conducted but there is a reliable estimate of the overall sexual recidivism rate for a routine sample of sexual offenders. This procedure is not recommended for samples that have been preselected on risk relevant variables, or when a substantial portion (>10%) of the estimated recidivism rates are greater than 50%.

By definition, the overall (average) recidivism rate, rr_T , is a function of the recidivism rate associated with each score, rr_s , and the frequency of each score, f_s :

$$rr_T = \sum_{s=-3}^{8+} f_s(rr_s) \tag{1}$$

In the present example, the Static-99R scores (s) range from -3 to $8+$, which was the range for which risk ratios were estimated in this study. With routine samples, the frequency of each score can be estimated from previous calculations of percentiles for Canadian samples, which have been relatively stable in international comparisons (Hanson et al., 2012). When recidivism rates follow a known exponential distribution, it is only necessary to estimate the recidivism rate associated with any single score (rr_s) to estimate the recidivism rates associated with all the other scores. The following calculations were based on anchoring the calculations on the recidivism rate associated with a Static-99R score of 2 (rr_2) as the reference category.

$$rr_s = rr_2 \left(HR^{[s-2]} \right) \tag{2}$$

The risk ratio is estimated by a hazard ratio (HR) of 1.393, which is the change in hazard rates for a one unit increase in Static-99R scores (see Appendix A: $e^{[.3318]} = 1.393$, based on using the complete sample [$n = 4,037$] to predict sexual recidivism). For example, the recidivism rate for a scores of 3 would be $rr_3 = rr_2(1.393)^{(3-2)}$ or 1.393 times the recidivism rate of a score of 2 (see Appendix A). A Static-99R score of 0 would have a recidivism rate that is $rr_0(1.393)^{(0-2)} = rr_2(1.393)^{-2} = rr_2(1/[1.393]^2) = .515$, or approximately half of the recidivism rate for a score of 2. Note that negative exponents are the equivalent of division, i.e., $HR^{-2} = (1/HR^2)$.

Substituting Equation 2 in Equation 1 provides:

$$rr_T = \sum_{s=-3}^{8+} f_s \left(rr_2 \left[HR^{(s-2)} \right] \right) \tag{3}$$

Given that rr_2 is a constant, Equation 3 can be reorganized to solve for rr_2 :

$$rr_2 = \frac{rr_T}{\sum_{s=-3}^{8+} \left[f_s \left(HR^{[s-2]} \right) \right]} \tag{4}$$

Using the distribution of Static-99R scores in Table 3 from Hanson et al. (2012) and the hazard ratio of 1.393, the right-hand denominator equals 1.407. Consequently, when the distribution and discrimination of Static-99R scores can be assumed to be “routine,” it is possible to estimate the absolute recidivism rate for a score of 2 by dividing the total recidivism rate by 1.407. This value can then serve as the anchor for calculating the absolute recidivism rates associated with all the other scores (see Equation 2, or Appendix A). Note that the overall (average) recidivism rate is higher than the recidivism rate for the median value because risk follows an exponential (not a linear) distribution.

Worked Example

As a demonstration, imagine that a single jurisdiction, Meehlia, wants to start using Static-99R to estimate absolute recidivism rates. Static-99R has not previously been used in Meehlia, and officials have concerns about the applicability of any of the available norms. Fortunately, they have good information concerning sexual recidivism rates in their jurisdiction, namely, it is 6.066% after 5 years, the same as the base rate for the data presented in Table 2 ($rr_T = 144/2,374 = 0.06066$). This base rate is the only new information that the Meehlian researchers need in order to calculate recidivism rates for specific scores. Using the estimation procedure described above, $rr_2 = rr_T / (1.407) = 0.06066 / 1.407 = 0.04311$; in other words, based on an overall average sexual recidivism of 6.066%, a recidivism rate of 4.31% would be expected for a score of 2 in routine samples. This estimate is very close to the logistic regression estimate of 4.35% (95% CI [3.53%, 5.35%]) calculated using the complete data (without nesting) in Table 2 ($B_0 = -3.09$; $SE = .111$). Note that this logistic regression estimate was calculated assuming that the data in Table 2 was selected from a single routine sample. The logistic regression estimates from meta-analysis (allowing for base rate variability across samples) are presented in Table 3.

To estimate the recidivism rate associated with a different score (e.g., 5), evaluators could use Equation 2: $rr_5 = rr_2 (HR^{[5-2]}) = 4.31(1.393^3) = 11.7\%$. Equivalently, the same value can be found by multiplying the recidivism rate for the reference category (4.31%) by the hazard ratio for a score of 5 found in Appendix A: $4.31\% \times 2.70 = 11.6\%$ (slight differences due to rounding).

Continuing with our hypothetical example, consider that over the next few years the Meehlians collect Static-99R scores on a large number (>500) of routine cases and notice that their distribution of scores is significantly different from the distributions presented in Hanson et al. (2012). The Meehlian researchers can then replace the Hanson et al. (2012) estimates of f_s in Equation 4 with their local distribution of Static-99R scores, thereby increasing the precision of local recidivism rate estimates.

Acknowledgments

Thank you to the following researchers for granting us permission to use their data and for being patient with our ongoing questions: Jacques Bigras, Sasha Boer, Jackie Craissati, Reinhard Eher, Doug Epperson, Tina Garby, Steve Gray, and Niklas Långström.

Authors' note

The views expressed are those of the authors and not necessarily those of Public Safety Canada.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

(References marked with an asterisk indicate studies included in the analyses.)

- Aalen, O. O., Borgan, O., & Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. New York, NY: Springer.
- Abel, G. G., & Harlow, N. (2001). *The stop child molestation book*. Bloomington, IN: Xlibris
- Akobeng, A. K. (2006a). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Paediatrica*, *96*, 338-341. doi:10.1111/j.1651-2227.2006.00180.x
- Akobeng, A. K. (2006b). Understanding diagnostic tests 2: Likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatrica*, *96*, 487-491. doi:10.1111/j.1651-2227.2006.00179.x
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 199-225). New York, NY: Routledge.
- Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 84-94. doi:10.1207/s15327752jpa8701_07
- Babchishin, K. M. (2011, November). *Risk ratios for Static-99R and Static-2002R*. Paper presented at the 36th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Atlanta, GA.
- Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Moving beyond the "low," "moderate," and "high" typology of risk communication. *Crime Scene*, *16*(1), 11-14. Retrieved from [http://www.cpa.ca/cpasite/userfiles/Documents/Criminal%20Justice/Crime%20Scene%202009-05\(1\).pdf](http://www.cpa.ca/cpasite/userfiles/Documents/Criminal%20Justice/Crime%20Scene%202009-05(1).pdf)
- Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012). Even highly correlated measure can add incrementally to predicting recidivism among sex offenders. *Assessment*, *19*, 442-461. doi:10.1177/1073191112458312.
- Bagos, P. G. (2011). Meta-analysis of haplotype-association studies: Comparison of methods and empirical evaluation of the literature. *BMC Genetics*, *12*(8), 1-16. doi:10.1186/1471-2156-12-8
- Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006). Different actuarial risk measures produce different risk rankings for sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 423-440. doi:10.1007/s11194-006-9029-9

- *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
- Beier, K. M., Neutze, J., Mundt, I. A., Ahlers, C. J., Goecker, D., Konrad, A., & Schaefer, G. A. (2009). Encouraging self-identified pedophiles and hebephiles to seek professional help: First results of the Prevention Project Dunkelfeld (PPD). *Child Abuse and Neglect*, 33, 545-549. doi:10.1016/j.chiabu.2009.04.002
- 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
- Bewick, V., Cheek, L., & Ball, J. (2005). Statistical review 14: Logistic regression. *Critical Care*, 9, 112-118. doi:10.1186/cc3045
- *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.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.
- Burkiewicz, J. S., Vesta, K. S., & Hume, A. L. (2008). Improving effectiveness in communicating risk to patients. *Consultant Pharmacist*, 23, 37-43. doi:10.4140/TCP.n.2008.37
- Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549-562. doi:10.1037/0033-2909.93.3.549
- Centers for Disease Control and Prevention. (2012). *Smoking and tobacco use: Cancer*. Retrieved from http://www.cdc.gov/tobacco/basic_information/health_effects/cancer/index.htm
- *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
- Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *Clinical Neuropsychologist*, 23, 193-204. doi:10.1080/13854040801968450
- Crowson, C. S., Therneau, T. M., Matteson, E. L., & Gabriel, S. E. (2007). Primer: Demystifying risk—Understanding and communicating medical risks. *Nature Clinical Practice Rheumatology*, 3, 181-187. doi:10.1038/ncprheum0397
- Dansereau, D. F., & Simpson, D. D. (2009). A picture is worth a thousand words : The case for graphic representations. *Professional Psychology: Research and Practice*, 40, 104-110. doi:10.1037/a0011827
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668-1674. doi:10.1126/science.2648573
- Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2008). Failure of Static-99 and SORAG to predict relevant reoffense categories in relevant sexual offender subtypes: A prospective study. *Sexual Offender Treatment*, 8(1), 1-20. Retrieved from <http://www.sexual-offender-treatment.org/sot-1-2008.html>

- *Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2009). [Data from sex offenders released from prison in Austria]. Unpublished raw data.
- Elmore, J. G., & Gigerenzer, G. (2005). Benign breast disease: The risk of communicating risk. *New England Journal of Medicine*, *353*, 297-299. doi:10.1056/NEJMe058111
- Embretson, S., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Lawrence Erlbaum.
- *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.
- Evans, J. St. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000) Frequency versus probability formats in statistical word problems. *Cognition*, *77*, 197-213. doi:10.1016/S0010-0277(00)00098-6
- Fischhoff, B., Brewer, N. T., & Downs, J. S. (2011). *Communicating risks and benefits: An evidence-based user's guide*. Silver Spring, MD: Food and Drug Administration (FDA), US Department of Health and Human Services.
- Fisher, B.S., Daigle, L.E., Cullen, F.T., & Turner, M.G. (2003). Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Criminal Justice and Behavior*, *30*, 6-38. doi:10.1177/0093854802239161
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York, NY: Russell Sage.
- Gasparrini, A. (2011). *MVMETA: Multivariate meta-analysis and meta-regression*. R package version 0.2.3.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704. doi:10.1037/0033-295X.102.4.684
- Grann, M., & Pallvik, A. (2002). An empirical investigation of written risk communication in forensic psychiatric evaluations. *Psychology, Crime & Law*, *8*, 113-130. doi:10.1080/10683160208401812
- Hammond, S. A., & O'Rourke, M. M. (2004). Developing a psychometric model for risk assessment. *Issues in Forensic Psychology*, *5*, 100-113.
- Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism* (Corrections Research User Report 97-04). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.
- Hanson, R. K. (1998). What do we know about sexual offender risk assessment? *Psychology, Public Policy and Law*, *4*, 50-72.
- Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology/Psychologie canadienne*, *20*, 172-182. doi:10.1037/a0015726
- *Hanson, R. K., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (Corrections Research User Report No. 2007-05). Ottawa, Ontario, Canada: Public Safety Canada.
- Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. *International Journal of Forensic Mental Health*, *11*, 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*, 119-136. doi:10.1023/A:1005482921333
- Hanson, R. K., & Thornton, D. (2003). *Notes on the development of Static-2002*. (Corrections Research User Report No. 2003-01). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.
- Harris, A. J. R., & Hanson, R. K. (2010). Clinical, actuarial, and dynamic risk assessment of sexual offenders: Why do things keep changing? *Journal of Sexual Aggression, 16*, 296-310. doi:10.1080/13552600.2010.494772
- 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 Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Atlanta, Georgia.
- Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504. doi:10.1037/1082-989X.3.4.486
- Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MR58443). Also available from www.static99.org
- Helmus, L., & Hanson, R. K. (2011). More fun with statistics! How to use logistic regression to predict criminal recidivism risk. *Crime Scene, 18*(2), 8-12. Retrieved from <http://www.cpa.ca/docs/file/Sections/Criminal%20Justice%20Psychology/Crime%20Scene%202011-10.pdf>
- 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, 33*, 1148-1171. doi:10.1177/0093854812443648
- Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: Journal of Research and Treatment, 24*, 64-101. doi:10.1177/1079063211409951.
- Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560.
- Hilton, N. Z., Carter, A., Harris, G. T., & Sharpe, A. J. B. (2008). Does using nonnumerical terms to describe risk aid violence risk communication? *Journal of Interpersonal Violence, 23*, 171-188. doi:10.1177/0886260507309337
- Hindman, J., & Peters, J. M. (2001). Polygraph testing leads to better understanding of adult and juvenile sex offenders. *Federal Probation, 65*, 8-15
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.
- Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, G., & Pickles, A. (2006). Subjective sleepiness and accident risk: Avoiding the ecological fallacy. *Journal of Sleep Research, 15*, 142-148. doi:10.1111/j.1365-2869.2006.00517.x

- Interstate Commission for Adult Offender Supervision. (2007). *Sex offender assessment information survey* (ICAOS 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
- Joint Legislative Audit and Review Commission. (2011). *Review of the civil commitment of sexually violent predators* (Report # 423 to the Governor and the General Assembly of Virginia). Richmond: Commonwealth of Virginia. Available from <http://jlarc.virginia.gov/reports/rpt423.pdf>
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697-720. doi:10.1037/0003-066X.58.9.697
- Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology*, *10*, 25-41. doi:10.1037/1076-898X.10.1.25
- 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.
- Kuhle, L. F., Neutze, J., Amelung, T., Grundmann, D., Scherner, G., Konrad, A., & Beier, K. M. (2012, September). *Treatment change in child pornography offending in pedophiles and hebephiles in the Prevention Project Dunkelfeld*. Paper presentation at the 12th Conference of the International Association for the Treatment of Sexual Offenders, Berlin, Germany.
- Kurz-Milcke, E., Gigerenzer, G., & Martignon, L. (2008). Transparency in communication: Graphical and analog tools. *Annals of the New York Academy of Sciences*, *1128*, 18-28. doi:10.1196/annals.1399.004
- *Langström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 107-120. doi:10.1177/107906320401600202
- Lussier, P., & Davies, G. (2011). A person-oriented perspective on sexual offenders, offending trajectories, and risk of recidivism: A new challenge for policymakers, risk assessors, and actuarial prediction? *Psychology, Public Policy, and Law*, *17*, 530-561. doi:10.1037/a0024388
- Mann, R. E., Hanson, R. K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. *Sexual Abuse: A Journal of Research and Treatment*, *22*, 191-217. doi:10.1177/1079063210366039
- McGrath, R. J., Cumming, G. F., Burchard, B. L., Zeoli, S., & Ellerby, L. (2010). *Current practices and emerging trends in sexual abuser management: The Safer Society 2009 North American survey*. Brandon, VT: Safer Society Foundation.
- McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse: A Journal of Research and Treatment*, *24*, 431-458. doi:10.1177/1079063211432475
- Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal for Forensic Mental Health*, *2*, 1-6. doi:10.1080/14999013.2003.10471174

- Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*, 339-363. doi:10.1037//0033-295X.90.4.339
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York, NY: McGraw-Hill.
- 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
- Otto, R. K., & Douglas, K. S. (Eds.). (2010). *Handbook of violence risk assessment*. New York, NY: Routledge.
- Paterson, B., Dowding, D., Harries, C., Cassells, C., Morrison, R., & Niven, C. (2008). Managing the risk of suicide in acute psychiatric inpatients: A clinical judgement analysis of staff predictions of imminent suicide risk. *Journal of Mental Health*, *17*, 410-423. doi:10.1080/09638230701530234
- Rosén, E., Stigson, H., & Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis and Prevention*, *43*, 25-33. doi:10.1016/j.aap.2010.04.003
- Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and likert scales: Too coarse for comfort. *Journal of Applied Psychology*, *77*, 336-342. doi:10.1037/0021-9010.77.3.336
- Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment*, *17*, 156-167. doi:10.1037/1040-3590.17.2.156
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and even occurrence*. New York, NY: Oxford University Press.
- Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P., Robbins, P. C., Mulvey, E. P., & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, *24*, 83-100. doi:0147-7307/00/0200-0083
- 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, Spring). Moving beyond the Standard Model for Actuarial Assessment for Sexual Offenders. *California Coalition on Sexual Offending, CCOSO Quarterly Newsletter, Perspectives*. Retrieved from <http://ccoso.org/newsletter.php>
- Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risks for sexual and violent recidivism. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.). *Sexually coercive behavior: Understanding and management* (pp. 225-235). Annals of the New York Academy of Sciences, Volume 989. New York: New York Academy of Sciences.
- Visschers, V. H. M., Meetens, R. M., Passchier, W. W. F., & de Vries, N. N. K. (2009). Probability information in risk communication: A review of the research literature. *Risk Analysis*, *29*, 267-287. doi:10.1111/j.1539-6924.2008.01137.x

- Vrieze, S. I., & Grove, W. M. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological Assessment, 22*, 382-395. doi:10.1037/a0019228
- Weinrott, M. R., & Saylor, M. (1991). Self-report of crimes committed by sex offenders. *Journal of Interpersonal Violence, 6*, 286-300. doi:10.1177/088626091006003002
- Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*, 740-767. doi:10.1037/a0020473