Psychological Assessment

Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools

R. Karl Hanson, Kelly M. Babchishin, L. Maaike Helmus, David Thornton, and Amy Phenix

Online First Publication, September 12, 2016. http://dx.doi.org/10.1037/pas0000371

Psychological Assessment
1040-3590/16/$12.00 http://dx.doi.org/10.1037/pas0000371

Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools

R. Karl Hanson
Public Safety Canada, Ottawa, Ontario, Canada

Kelly M. Babchishin
University of Ottawa

L. Maaike Helmus
Wandering Vagabond

David Thornton
Sand Ridge Secure Treatment Centre, Madison, Wisconsin

Amy Phenix
Morro Bay, California

This article describes principles for developing risk category labels for criterion referenced prediction measures, and demonstrates their utility by creating new risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. Currently, risk assessments in corrections and forensic mental health are typically summarized in 1 of 3 words: low, moderate, or high. Although these risk labels have strong influence on decision makers, they are interpreted differently across settings, even among trained professionals. The current article provides a framework for standardizing risk communication by matching (a) the information contained in risk tools to (b) a broadly applicable classification of "riskiness" that is independent of any particular offender risk scale. We found that the new, common STATIC risk categories not only increase concordance of risk classification (from 51% to 72%)—they also allow evaluators to make the same inferences for offenders in the same category regardless of which instrument was used to assign category membership. More generally, we argue that the risk categories should be linked to the decisions at hand, and that risk communication can be improved by grounding these risk categories in evidence-based definitions.

Keywords: standards, criterion referenced tests, Static-99R, Static-2002R, risk communication

> We are greatly at a loss for a standard whereby to measure cold. The common instruments show us no more than the relative coldness of the air, but leave us in the dark as to the positive degree thereof; whence we cannot communicate the idea of any such degree to another person.
>
> —Robert Boyle (1665), quoted in Landsberg (1964, pp. 42–43)

Many of us involved with applied psychological assessment can empathize with Boyle's concerns. Boyle was writing at a time (17th century) when there were more than 35 different temperature scales in use, and whoever constructed a new type of thermometer simultaneously created a new scale to go along with it (Landsberg, 1964). Contemporary psychological assessment faces a similar challenge. Although there are a large number of measures that reliably rank individuals on constructs such as anxiety or antisocial traits, we have yet to establish consensus for communicating the results (Blanton & Jaccard, 2006).

The current study was motivated by our need, as test developers, to update the category labels for certain actuarial sexual offender risk assessment tools, specifically, Static-99R and Static-2002R (Hanson & Thornton, 2000; Helmus, Thornton, Hanson, & Babchishin, 2012). The primary purpose of these tools is to estimate the relative risk of sexual recidivism based on commonly available demographic and criminal history information. Like other empirically derived actuarial risk tools, norms for these tools are peri-

R. Karl Hanson, Public Safety Canada, Ottawa, Ontario, Canada; Kelly M. Babchishin, Royal's Institute of Mental Health Research, University of Ottawa; L. Maaike Helmus, Wandering Vagabond; David Thornton, Sand Ridge Secure Treatment Centre, Madison, Wisconsin; Amy Phenix, Morro Bay, California.

Funding for this project was provided in part by the Canadian Institute for Health Research (Banting Postdoctoral Fellowship). The views expressed are those of the authors and not necessarily those of Public Safety Canada or the Sand Ridge Secure Treatment Centre. R. Karl Hanson, Kelly Babchishin, L. Maaike Helmus, and David Thornton are coauthors of the Static-99R and Static-2002R risk tools. R. Karl Hanson, L. Maaike Helmus, David Thornton, and Amy Phenix are certified trainers for these instruments. The copyright for Static-99R and Static-2002R is held by the Government of Canada. We thank Anton Schweighofer and Ian Barsetti for helpful comments on previous versions of this article.

Correspondence concerning this article should be addressed to R. Karl Hanson, Research Division, Community Safety and Countering Crime Branch, Public Safety Canada, 340 Laurier Avenue, West, Ottawa, Ontario, Canada K1A 0P8. E-mail: karl.hanson@canada.ca

odically updated as new and better research becomes available. Our question was whether, then how, should we revise their risk category labels?

There has been considerable discussion about how to communicate the results of norm referenced measures, for which scores can be interpreted as the position of an individual within a defined group (Crawford, Garthwaite, & Slick, 2009; Oosterhuis, van der Ark, & Sijtsma, 2016). Such an interpretation, however, poorly expresses the information contained in criterion referenced prediction measures, in which the goal of the assessment is to estimate the likelihood of a significant outcome, such as suicide (Berman & Silverman, 2014), major depression (King et al., 2008), success in law school (Thomas, 2003), or, in our case, recidivism by sexual offenders. Prediction measures are also different from diagnostic measures (e.g., x-rays for brain tumor), which can be evaluated in terms of diagnostic accuracy (e.g., false positives, false negatives, positive predictive value; Swets, 1988). For prognostic measures, in contrast, the outcome of interest is not present at the time of assessment and may never happen (e.g., risk of breast cancer; Moons, Royston, Vergouwe, Grobbee, & Altman, 2009; for review, see Helmus & Babchishin, in press).

There are no universal standards for labeling relative or absolute likelihoods of adverse events, nor do we expect there ever will be. A 10% chance of a hurricane is high risk (Monahan & Steadman, 1996); a 10% chance of rain is not. A 10% chance of your car's brakes failing is catastrophic (for reviews, see Hilton, Scurich, & Helmus, 2015; Visschers, Meertens, Passchier, & de Vries, 2009). We believe, however, that there are certain common principles worth considering when developing risk category labels within any specific domain. Although we focus on offender risk assessment, some of these principles may also be helpful when considering category labels in other areas of applied psychological assessment. Specifically, we argue that certain quantitative information (percentile ranks, risk ratios, estimates of the rates of outcomes) should inform the meanings ascribed to risk category labels.

The concept of risk is ubiquitous in applied decision making, and is a dominant concern of business and industry. For example, the International Organization for Standardization (ISO 31000) defines *risk* as the "effect of uncertainty on objectives" (Gjerdrum & Peter, 2011). A very similar definition has been adopted by proponents of the structured professional judgment (SPJ) approach to violence risk assessment (Douglas & Ogloff, 2003). In the user guide for the HCR-20^{V3}, for example, risk is defined as "a threat or hazard that is incompletely understood, and thus whose occurrence can be forecast only with uncertainty" (Douglas, Hart, Webster, & Belfrage, 2013, p. 4). From this perspective, it makes little sense to associate risk categories labels with precise, numeric estimates of recidivism risk. If risk is fundamentally uncertainty, then recidivism estimates that are not close to 0 or 1 are expressions of ignorance. Within the SPJ approach, the primary role of the risk assessor is not to estimate likelihoods; instead, evaluators are charged with developing a case formulation useful for guiding management and intervention strategies (Hart & Boer, 2010).

In contrast, the estimation of empirical probabilities for adverse outcomes has been an important theme in medical epidemiology since the 1990s (e.g., death by cancer; Aalen, Borgan, & Gjessing, 2008; Greenland, 1998; Rockhill, Byrne, Rosner, Louie, & Colditz, 2003). A guiding principle of much of this work is stochastic causality, meaning that outcomes have inherently probabilistic connections to initial conditions (Gillies, 2000; Popper, 1959): "The discovery that individual events are irreducibly random is probably one of the most significant findings of the twentieth century" (Zeilinger, quoted in Aalen et al., 2008, p. 347). From this perspective, when a weather forecaster states that "there is a 50% chance of rain tomorrow," it is not an expression of profound ignorance or doubt; instead, competent forecasters are communicating an informed, evidence-based, and accurate opinion about the likelihood of rain (Sanders, 1963).

Whereas weather forecasts are now routinely communicated to the public in numbers (temperature in degrees, percent probability of rain), offender risk assessments are typically communicated using words such as low, moderate, or high (Blais & Forth, 2014; Heilbrun, O'Neill, Strohman, Bowman, & Philipson, 2000; Heilbrun, Philipson, Berman, & Warren, 1999). For the Static-99R, in particular, evaluators almost always report its category labels in high-stakes evaluations (Chevalier, Boccaccini, Murrie, & Varela, 2015). Risk category labels also hold influence: Prospective jurors are far more influenced by the Static-99R category labels than by any of the numeric information associated with Static-99R scores (Varela, Boccaccini, Cuervo, Murrie, & Clark, 2014).

Although risk labels are both influential and widely used, it is often not clear what they mean. There is only a loose association in natural language between verbal labels for likelihood (e.g., rare) and numeric probabilities (e.g., 5%; Beyth-Marom, 1982; Lichtenstein & Newman, 1967). Similarly, risk categories labels (e.g., low risk) are associated with widely varying recidivism rates within groups of mental health practitioners (e.g., Hilton, Carter, Harris, & Sharpe, 2008) and judges (e.g., Monahan & Silver, 2003). Disagreements in interpretation persist irrespective of experience with forensic risk assessment (Slovic, Monahan, & MacGregor, 2000).

Such disagreements are understandable. As we were updating the risk category labels for Static-99R and Static-2002R, we found no accepted standards that connect risk category labels to specific meanings, such as recidivism rates, psychological features, or expected treatment needs. Risk category labels are interpretations of risk assessment results. For SPJ measures, assignment of these labels is the responsibility of the evaluator (e.g., Douglas et al., 2013). In contrast, many actuarial measures have labels preassigned by the test's developers (e.g., a Static-99R score of 6 and above is "high"). However, different evaluators and different test developers do not use risk category labels in the same way. Consequently, it should not be surprising that there is substantial variation in the observed recidivism rates for offenders ascribed the same label (e.g., "high risk") by different risk tools (Singh, Fazel, Gueorguieva, & Buchanan, 2013, 2014).

If we are going to advance the professional lexicon for risk communication, we need to have standardized metrics to represent the information contained in risk assessments (Babchishin & Hanson, 2009; Blanton, & Jaccard, 2006). In this article, we review principles for creating risk category labels, and demonstrate the utility of these principles by creating new, standardized risk category labels for the Static-99R and Static-2002R sexual offender risk assessment tools. Just as the results of norm referenced tests can be quantified by percentiles and related metrics, several metrics are available for quantifying the information contained in criterion referenced prediction measures (e.g., risk ratios, probabilities). Using these quantitative indicators to guide the construction of risk categories has the promise of increasing concordance across measures, and increasing the consensus of their interpretation, over that achieved using only natural language (Budescu, Por, & Broomell, 2012; Karelitz & Budescu, 2004).

Static-99R and Static-2002R

The STATIC risk scales (Static-99, Static-99R, Static-2002, and Static-2002R; see Appendix for items) are the most commonly used tools in the world to assess the recidivism risk posed by sexual offenders (Neal & Grisso, 2014). Although not all sexual offenses are equally serious, the base rate of sexual recidivism is sufficiently low (Helmus, Hanson, Thornton, Babchishin, & Harris, 2012) and public concern about sexual victimization is sufficiently high that any new sexual offense is problematic. Consequently, the STATIC tools focus on any sexual recidivism, not imminence or severity. Static-99 consists of 10 items assessing criminal history, demographic information, and victim characteristics, with total scores ranging between 0 and 12. It is widely used in Canada, the United States, and Australia for treatment planning (Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010), community supervision (Interstate Commission for Adult Offender Supervision, 2007), and for preventative detention hearings (Blais & Forth, 2014; Doyle, Ogloff, & Thomas, 2011; Jackson & Hess, 2007; Neal & Grisso, 2014). Although the predictive accuracy of Static-99 is not notably better than other actuarial tools designed for sex offenders, it is the most researched (Hanson & Morton-Bourgon, 2009), and can be scored by diverse professionals using commonly available information.

Static-2002 was created to provide a similar scale (i.e., simple and easy to score) but with increased coherence and conceptual clarity (see Hanson & Thornton, 2003). In a multisite study across eight diverse samples, Static-2002 had significantly greater accuracy than Static-99 in predicting sexual, violent, and any recidivism, although the difference for sexual recidivism was small (Hanson, Helmus, & Thornton, 2010). Static-2002 is also widely used, particularly in Canada (McGrath et al., 2010).

Informed by research suggesting that Static-99 did not adequately account for the relationship between age at release and recidivism (Barbaree, Langton, & Blanchard, 2007; Barbaree, Langton, Blanchard, & Cantor, 2009; Hanson, 2006; Thornton, 2006), new age weights were developed (Helmus, Thornton, et al., 2012). Although Static-2002 accounted for age better than Static-99, the revised age item was applied to both scales because of the expectation that the optimal age item should be the same for both scales. After this revision, the scales were called Static-99R and Static-2002R. With the improved incorporation of age, both scales demonstrated similar predictive accuracy, although they did add incrementally to each other in the prediction of recidivism (Babchishin, Hanson, & Helmus, 2012b). Consequently, although Static-2002 was intended to be an improvement and potential replacement for Static-99, the research findings indicated that they should be considered different scales. Based on these findings, we (the STATIC Development Team) have recommended that evaluators use Static-99R or Static-2002R, or both (Hanson, 2014; Phenix, Helmus, & Hanson, 2015).

Since the revision of the scales, we have produced normative data to communicate risk information from the scales in various quantitative metrics, including percentiles (Hanson, Lloyd, Helmus, & Thornton, 2012), risk ratios (Babchishin, Hanson, & Helmus, 2012a; Hanson, Babchishin, Helmus, & Thornton, 2013), and absolute recidivism estimates (Hanson, Thornton, Helmus, & Babchishin, 2016).

The risk category labels were left unaddressed. Based on total scores, Static-99 had four named risk categories (0–1 = low; 2–3 = low-moderate; 4–5 = moderate-high; and 6+ = high) and Static-2002 had five (0–2 = low; 3–4 = low-moderate; 5–6 = moderate; 7–8 = moderate-high; and 9+ = high). When the scales were revised, we did not alter the risk category labels or their associated cutoff scores, with the exception that it was now possible to have scores less than zero (and these were retained in the low-risk group; Helmus, Thornton, et al., 2012). The existence of new norms, however, motivated us to carefully consider the risk categories for these measures.

The methods for developing the original Static-99 and Static-2002 risk categories were not well articulated. In her unpublished undergraduate thesis, Helmus (2007) described the development of Static-2002 risk categories, which was similar to the methods employed for Static-99. The original Static-2002 categories were guided by three general principles: (a) risk categories should encompass at least 10% of the sample, (b) there should be meaningful increases in recidivism rates between categories, and (c) the categories should maximize fit to the data (i.e., highest area under the receiver operating characteristic curve [AUC] value). For Static-2002, six different options (e.g., three categories, four categories, cut point at 4, at 5) were considered for creating risk categories, and we adopted the one with the highest AUC value. Although this method was plausible, these rules require subjective decisions and it is likely that the evidence used to select the “best” categories (i.e., AUC values) capitalized on chance features of the data. In general, the approach used to create the original STATIC risk categories was insensitive to construct validity. Furthermore, the method currently used to link Static-99R and Static-2002R scores to recidivism rates (logistic regression) assumes that risk is continuous and that there are no natural breaks separating risk categories (Hanson et al., 2010, 2013, 2016). Consequently, there is a need for more defensible categories.

Principles for Creating and Naming Risk Categories

The most useful risk categories have construct validity, that is, a set of related meanings that support professional reasoning and inferences concerning the individual being assessed (Cronbach & Meehl, 1955; Joint Committee on Standards for Educational and Psychological Testing, 2014). These meanings are determined by the intended scope of the assessment tool (what it purports to measure) and by research findings (what it actually measures). As we advance our understanding, we learn more about the latent constructs responsible for recidivism risk, such as general criminality and sexual criminality (Brouillette-Alarie, Babchishin, Hanson, & Helmus, 2016). Although criterion referenced prediction tools are primarily designed to estimate likelihoods, they can also inform decisions concerning psychological characteristics and treatment needs.

Risk categories should align with their intended purpose. Typically, evaluators use offender risk categories to communicate the urgency with which action is required (e.g., “high risk” cases require exceptional resources and attention). Response options, however, are the product of both the individual assessed and the context of the assessment. As well, not all decision makers consider the same risk level as equally serious. Nevertheless, good risk categories should convey implications for action (or inaction) that can be justified empirically (e.g., "offenders at this risk level require this amount of treatment to reduce their risk to the next lowest category," or "intervention is unnecessary because the individual's risk is already below acceptable thresholds").

For the revised STATIC category labels, a preliminary consideration was how many categories were necessary. Our assumption was that offender risk is well represented by a continuous dimension, without clear, distinct levels existing in nature (Guay, Ruscio, Knight, & Hare, 2007; Patrick, Fowles, & Krueger, 2009). Consequently, the thresholds between risk levels could not be empirically derived; instead, they would need to be determined by differences that are practically meaningful and reliable. Our decision for five levels was influenced by related discussions led by the Council of State Governments Justice Center (2014) on standardized risk categories for offender risk/need assessments tools in corrections (to be described later).

Rater Reliability

Rater reliability should be a consideration when determining the width of risk categories estimated by external judges (raters). For criterion referenced measures, reliability is determined by thresholds, not correlations across the full range of scores (Meyer, 2010). Greatest classification precision is achieved when there are few cases near the thresholds and many cases in the middle of the risk categories. Reliable criterion referenced assessment can also result in all or none of the cases surpassing a predefined threshold (e.g., everybody/nobody passes an exam; Meyer, 2010).

When Static-99R and Static-2002R are scored by trained researchers, interrater reliability (as measured by correlations with the full range of scores) is typically between .85 and .95 (e.g., McGrath, Lasher, & Cumming, 2012; Smid, Kamphuis, Wever, & van Beek, 2014; Thornton & Knight, 2015). Research with Static-99 has suggested that agreement may be lower in routine use (Boccaccini et al., 2012; Levenson, 2004), for higher scores (A. K. Rice, Boccaccini, Harris, & Hawes, 2014), or when comparing the scores of opposing experts (Murrie et al., 2009). Nevertheless, Static-99 rater reliability is typically in the .80s or higher for both field validity and research studies (see review by Phenix & Epperson, 2015). In practice, rater reliabilities in the .80s corresponds to exact agreement on Static-99 scores about half the time, and disagreement by no more than one STATIC point nine times out of 10 (Boccaccini et al., 2012; Hanson, 2001; Quesada, Calkins, & Jeglic, 2014).

Imperfect rater reliability means that the more category thresholds there are, the greater the likelihood that offenders will fall on the "wrong" side of the threshold simply by chance alone. For offenders who score one STATIC point below the threshold (e.g., 5, in which the threshold is 6), misclassification is expected about half the time. For offenders scoring 2 STATIC points below the threshold (e.g., 4, in which the threshold is 6), misclassification is expected in 1 of 10 cases. Although it would be ideal if each score supported its own unique interpretation, prudent evaluators using the STATIC measures may want to base substantive interpretations on the range defined by the assigned score and its adjacent scores (3-point range).

Quantifying Recidivism Risk

There are three quantitative metrics worth considering when creating risk categories for prediction tools: percentile ranks, absolute recidivism rates, and risk ratios (Babchishin & Hanson, 2009; Lehmann, Thornton, Helmus, & Hanson, 2016). Each has its own strengths and weaknesses as a metric for risk communication.

Percentile ranks. In psychology, the most commonly used metrics for reporting individuals' test results are based on percentile ranks, such as *z* scores, *t* scores, and IQ scores. Percentile ranks measure the unusualness of particular characteristics and are ideal for norm referenced tests (Crawford & Garthwaite, 2009). Percentile ranks can also be useful for prediction tools because they are easily calculated, easily understood, and may be sufficient for resource allocation decisions (e.g., when treatment is only provided to the riskiest 20%). Percentile ranks, however, have no intrinsic relationship to the likelihood of the outcome, which is a primary concern of prediction tools.

Absolute risk. The most commonly reported quantitative risk information for the STATIC risk tools are absolute recidivism rates (Chevalier et al., 2015). Recidivism rate tables are an intrinsic feature of actuarial risk tools (Dawes, Faust, & Meehl, 1989), and are central to decisions based on the absolute (not relative) likelihood of an outcome (e.g., is this offender more likely than not to sexually reoffend?).

Recidivism rates, however, are difficult to estimate with certainty. Only a portion of sexual offenses are detected and recorded in available databases (Falshaw, Bates, Patel, Corbett, & Friendship, 2003). Recidivism rates also vary based on features of the research design: for example, length of follow-up, sample selection characteristics, and the recidivism criteria. Even after standardizing research designs, however, the amount of variation in observed recidivism rates remains greater than would be expected by chance (Hanson et al., 2016; Helmus, Hanson, et al., 2012; M. E. Rice, Harris, & Lang, 2013).

Nevertheless, aligning risk categories with expected recidivism rates provides important information about base rates. Given the tendency of the general public to overestimate the likelihood of sexual recidivism (Center for Sex Offender Management, 2010), even approximate recidivism rate ranges (e.g., point estimates with confidence intervals; Imrey & Dawid, 2015) provide information that may otherwise have been neglected by decision makers. Confidence intervals provide a range of plausible values upon which to base decisions, and nonoverlapping 95% confidence intervals correspond to significant differences at the $p < .01$ level (Cumming & Finch, 2005). For the STATIC measures (see Table 1 and Table 2), the confidence intervals for most adjacent scores overlap, except for the well-populated scores in the middle of the risk distribution, in which there is no overlap at all. It is important to remember, however, that confidence intervals are largely determined by sample size, and even measures with little information value can produce predicted values with small confidence intervals if the sample sizes are large.

Assigning evaluative labels for sexual recidivism rates requires judgments concerning the seriousness of the outcome (Theil, 2002; Visschers et al., 2009). In the context of criminal justice, one

Table 1
Evidence-Based Risk Categories for Static-99R

	Category		Percentiles					
Static-99R score	Number	Name	Same score	Cumulative midpoint average	Risk ratio	Predicted 5- year recidivism rate	Lower CI	Upper CI
−3	I	Very low risk	2.7	1.3	.19	.9	.6	1.3
−2	I	Very low risk	3.0	4.2	.26	1.3	1.0	1.8
−1	II	Below average risk	7.9	9.7	.37	1.9	1.4	2.5
0	II	Below average risk	10.3	18.7	.52	2.8	2.2	3.5
1	III	Average risk	15.7	31.7	.72	3.9	3.3	4.7
2	III	Average risk	17.5	48.3	1.00	5.6	4.8	6.5
3	III	Average risk	17.2	65.7	1.39	7.9	7.0	8.8
4	IV-a	Above average risk	10.7	79.6	1.94	11.0	10.0	12.1
5	IV-a	Above average risk	7.4	88.7	2.70	15.2	13.8	16.6
6	IV-b	Well above average risk	3.6	94.2	3.77	20.5	18.4	22.8
7	IV-b	Well above average risk	2.5	97.2	5.25	27.2	24.0	30.7
8	IV-b	Well above average risk	1.2	99.1	7.32	35.1	30.5	40.0
9	IV-b	Well above average risk	.28	99.9	—	43.8	37.8	50.1
10+	IV-b	Well above average risk	.02	99.99	—	53.0	45.6	60.3

Note. CI = confidence interval.

plausible threshold for defining low-risk sexual offenders would be those whose risk for a new sexual offense is no different than that of offenders with no recorded history of a sexual offense (1% to 2% within the first 5 years; Bonta & Hanson, 1995; Bonta, Rugge, & Dauvergne, 2008; Duwe, 2012; Wormith, Hogg, & Guzzo, 2012). It would also be uncontroversial to consider individuals virtually certain to reoffend as high risk. Labels for other thresholds are not currently defined in natural or professional language. For sexual recidivism in particular, jurors perceive probabilities in the 15% to 30% range as sufficient to meet the "likelihood" threshold for civil commitment in the United States (Knighton, Murrie, Boccaccini, & Turner, 2014; Scurich & Krauss, 2014).

Risk ratios. Given the difficulty of estimating absolute recidivism rates, another useful quantitative metric for risk communication is the risk ratio (Babchishin et al., 2012a; Hanson et al., 2013). Risk ratios compare the recidivism rate of offenders with a particular score to the recidivism rate of a reference group (e.g., sexual offenders with a Static-2002R score of 6 are 2.6 times more likely to sexually reoffend than those in the middle of the risk distribution). There are several different ways of computing risk ratios, including rate ratios, odds ratios, and hazard ratios; each of these methods, however, provides substantively similar interpretations when the recidivism base rate is low (<20%). An important feature of risk ratios is that they are stable across samples, settings, outcome criteria, and follow-up times (Babchishin et al., 2012a; Hanson et al., 2013; Helmus, Hanson, et al., 2012; Weinberger et al., 2010).

The main weakness of risk ratios is that they appear to imply more about absolute risk than they actually do. Namely, risk ratios are only informative about absolute risk if the relevant base rate is also known and provided (Akobeng, 2008).

Standardized risk levels for offender risk tools. The problems we faced defining categories for the STATIC risk tools were

Table 2
Evidence-Based Risk Categories for Static-2002R

	Category		Percentiles					
Static-2002R score	Number	Name	Same score	Cumulative midpoint average	Risk ratio	Predicted 5-year recidivism rate	Lower CI	Upper CI
−2	I	Very low risk	2.8	1.4	.20	1.0	.6	1.7
−1	I	Very low risk	2.9	4.2	.28	1.5	.9	2.3
0	II	Below average risk	6.7	9.0	.38	2.2	1.5	3.2
1	II	Below average risk	9.7	17.3	.52	3.2	2.3	4.4
2	III	Average risk	16.0	30.1	.72	4.6	3.6	6.0
3	III	Average risk	17.9	47.1	1.00	6.8	5.5	8.2
4	III	Average risk	15.3	63.7	1.38	9.7	8.3	11.3
5	IV-a	Above average risk	13.5	78.0	1.90	13.8	12.2	15.6
6	IV-a	Above average risk	7.1	88.3	2.63	19.2	16.9	21.6
7	IV-b	Well above average risk	2.8	93.3	3.62	26.0	22.6	29.8
8	IV-b	Well above average risk	2.5	95.9	5.00	34.3	29.1	40.0
9	IV-b	Well above average risk	2.3	98.3	6.90	43.7	36.5	51.2
10	IV-b	Well above average risk	.4	99.7	—	53.5	44.4	62.4
11+	IV-b	Well above average risk	.1	99.9	—	—	—	—

Note. CI = confidence interval.

shared by other users and developers of offender risk tools. There are currently hundreds of different offender risk tools used worldwide (Singh et al., 2014), each with its own interpretive categories. Based on shared concerns, some of us (RKH, KB) have been collaborating with the U.S. Council of State Governments Justice Center to develop standardized risk levels for general offender risk tools (Justice Center, 2014, 2016). One thread of the Justice Center's discussions has been a proposal to anchor offender risk communication in five broad categories (Justice Center, 2014).

The lowest risk category (Level I) would be generally prosocial individuals who have nonetheless committed crime. They would not be expected to have the criminal backgrounds, criminogenic needs, or the prognosis typical of offenders. The recidivism rates of Level I offenders would be indistinguishable from the rates of spontaneous offending among nonoffenders (e.g., young males). Level II would be higher risk than nonoffenders, but lower risk than typical offenders. It is expected that Level II offenders would have some criminogenic needs, but that these life problems would be few and transient. Level III offenders would be the typical offenders in the middle of the risk distribution. Typical offenders have criminogenic needs in several areas, and require meaningful investments in structured programming to decrease their recidivism risk. Level IV offenders would be perceptibly higher risk than the typical offender. Most of these offenders would have chronic histories of rule violations, poor childhood adjustment, and significant criminogenic needs across multiple domains. The Justice Center's framework also included a fifth category for the highest risk offenders, defined as those virtually certain to reoffend. Level V offenders are those typically found in high-security units, in which considerable resources are devoted to managing *current* antisocial behavior.

One impediment to directly adopting the Justice Center's risk levels for the STATIC measures was that the Justice Center's levels were designed to describe general criminality, not the risk for sexual recidivism. The sexual offenders who are most likely to reoffend not only are generally criminal but also have sexual-crime-specific risk factors, such as atypical sexual preferences, emotional identification with children, and sexualized coping (Barbaree, Langton, & Peacock, 2006; Hanson & Morton-Bourgon, 2004, 2005).

Revised Risk Categories for Static-99R and Static-2002R

Based on all these considerations, the existing norms, and extensive consultation with STATIC users and trainers, we, the STATIC Development Team, developed new, improved risk categories for Static-99R and Static-2002R (summarized in Tables 1 and 2). The same principles were used to create the five categories for both Static-99R and Static-2002R, in order for category labels to have the same meaning for both instruments.

The following is a summary of our decision process. The first step was to create a middle category at the median. To account for measurement error, the middle category, Category III, was expanded up one unit and down one unit (scores from 1 to 3 for Static-99R and 2 to 4 for Static-2002R). This category included approximately half of sexual offenders.

Next, we searched for, and found in our recidivism rate tables, a category equivalent to Justice Center's Level I. Rather than defining Category I by comparison with nonoffenders, however, the STATIC Category I was defined in comparison with nonsexual offenders. Specifically, to be included in Category I, the expected sexual recidivism rates needed to be similar to the rate of spontaneous sexual offenses for offenders with no prior convictions for sexual offenses (<2% after 5 years; Bonta & Hanson, 1995; Bonta et al., 2008; Duwe, 2012; Wormith et al., 2012). Based on the recidivism rate tables (Hanson et al., 2016), this category was associated with the two lowest values in Static-99R (−3, −2) and Static-2002R (−2, −1).

The next step was to identify a group that was meaningfully lower than the middle in terms of relative risk, but still higher than Category I. Meaningfully lower was defined, heuristically, as half the recidivism rate as those in the middle of the risk distribution. This definition captured the Static-99R scores of 0 and −1, and the Static-2002R scores of 0 and 1. Although the range of scores in Category II is narrow, the category was still well populated, capturing 25% to 30% of sexual offenders.

The parallel category (Category IV-a) included those who were higher risk than sexual offenders in the middle of the risk distribution (4, 5 for Static-99R; 5, 6 for Static-2002R), but were not the very highest risk sexual offenders. The final risk category (IV-b) comprised the top 8% of Static-99R scores (6+) and Static-2002R scores (7+) and was defined based on relative risk. The expected recidivism rates of offenders in the highest risk category was twice as high as for those in the previous category, and approximately 4 times higher than offenders in the middle of the risk distribution (Category III).

We also considered three (lower than average, average, above average) and four (very low, lower than average, average, above average) categories, along with the five categories. Our consultations suggested that the five categories would be more useful than three or four categories for the diverse decisions that are currently informed by the STATIC measures.

Names for the categories were based on internal discussions and more than a year of consultations with STATIC users and trainers. We considered various options, including having no descriptive labels at all. In the end, names were provided with the expectation that if we did not provide them, then a diversity of names would be provided by the users. We rejected the popular low/moderate/high labels because these terms already carried too many preexisting associations. Instead, the new names were based on relative risk because this property was considered the most empirically stable feature of risk scale scores: average, below average, above average, and well above average. The one exception to this naming convention was that the lowest risk category was named "very low risk" because we accepted the label's associations in natural language.

Given the difficulty agreeing on names, we also provided category numbers that parallel the Justice Center's standardized risk levels. However, because the observed sexual recidivism rates for the highest risk category were only 20% to 50%, we labeled the highest STATIC level Category IV-b, not Category V. Observed recidivism rates in this range do not meet the criteria for Justice Center's Level V (virtually certain to reoffend), even if the real rates are substantially higher than the observed rates (Falshaw et al., 2003).

Research Objectives

Following the creation of the categories, we examined their potential utility in a representative sample of sexual offenders aggregated from four different studies ($N > 2{,}000$). First, we expected the new categories to increase category concordance compared with the original measures. Second, we expected that the new categories would identify groups who shared more than the quantitative information (risk ratios, recidivism rates) used to create the categories. Specifically, we expected that the pattern of associated psychological features would match those articulated for the Justice Center's standardized risk levels, and be similar regardless of which instrument was used to create the categories. We expected strong differences between the risk levels on prior criminal history (any prior involvement in the criminal justice system, prior sexual offenses, stranger victims), as well as on demographic (young age, separation from parents prior to 16 years old) and psychological variables related to sexual recidivism risk (sexual preoccupation, impulsivity, lack of cooperation with supervision, and capacity for intimacy). For these variables, we expected these problems to be common (frequency >50%) in Category IVa and IVb, and infrequent (<20%) in Category I and Category II. In contrast, we expected weak associations between the categories and features that previous research (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005) has found to be only weakly related to sexual recidivism risk, such as psychiatric history, developmental delay, and having child victims of sexual offenses.

Method

Data Sources for the Revised Categories

The following sources of normative data were used to inform the development of the risk categories: percentiles for Static-99R and Static-2002R (Hanson, Lloyd et al., 2012), risk ratios for Static-99R (Hanson et al., 2013), risk ratios for Static-2002R (Babchishin et al., 2012a), and recidivism rate estimates for Static-99R and Static-2002R (Hanson et al., 2016). All of the normative data were based on meta-analyses of routine correctional samples from Canada, Europe, and the United States.

Samples for Analyses of New Categories

From the existing STATIC normative data, we used routine (i.e., unselected, representative) correctional samples that had scores available for both Static-99R and Static-2002R. As per the coding rules of the scales (A. J. R. Harris, Phenix, Hanson, & Thornton, 2003; Phenix, Doren, Helmus, Hanson, & Thornton, 2009), cases were deleted if more than one Static-2002 item was missing, any Static-99 item was missing other than Ever Lived with a Lover (Item 2), the sexual offender was less than 18 years old at time of release or less than 16 years old when they committed the index sex offense, or if the sexual offender was female. Additionally, detailed data cleaning was conducted in all data sets and inconsistencies were resolved by discussion with the original study authors. Four samples were available ($N = 2{,}395$).

Table 3 provides descriptive information for the samples. For additional information, readers are referred to other, more detailed summaries (e.g., Hanson et al., 2016; Phenix et al., 2015) or to the original studies. Three samples were from Canada and one was from Germany. The average age at release was 40 years old ($SD = 12$). Sexual offenders were released between 1976 and 2007, with a median release year of 2001. All samples were fairly evenly split between offenders with adult victims and those with child victims. For all studies, the original Static-99 and Static-2002 scores were available in the data sets, and we computed the revised versions of the scales (with updated age weights) via syntax from information on the offender's date of birth.

Bigras (2007). The original sample contained 94% of all sexual offenders receiving a federal sentence (2 or more years in custody) in Quebec between 1995 and 2000 (6% refused participation in the research or were unable to provide consent). Data on the 457 offenders in the current study were collected during their initial sentencing evaluation at the Regional Reception Centre in Sainte-Anne-des-Plaines (Quebec, Canada), a maximum-security penitentiary of the Correctional Service of Canada (CSC). Detailed offense history information was coded by trained research assistants (graduate students in psychology and criminology) from file data and interviews using a structured scoring guide. Static-99R and Static-2002R scores were subsequently calculated via SPSS syntax. The mean kappa for the individual variables used to compute STATIC scores was .90 ($SD = .2$). Recidivism information was obtained from Canadian national records (Royal Canadian Mounted Police; RCMP) in 2004, allowing for an average 4.5 years of follow-up. The period at risk excluded periods incarcerated after release from the index sexual offense (street time, not calendar time).

Boer (2003). This sample ($n = 296$) consisted of all federal offenders (CSC) serving a sentence for a sexual offense in British Columbia whose custodial sentence expired between January 1990 and May 1994, although offenders in this cohort began receiving conditional release early as 1976. The average follow-up time was 12.6 years, and recidivism was coded based on RCMP records. Sexual recidivism was defined according to the A. J. R. Harris et

Table 3
Descriptive Information for Samples

Study	*n*	*n* with adult victims	*n* with child victims	Age *M* (*SD*)	Country	Release period	Static-99R *M* (*SD*)	Static-2002R *M* (*SD*)
Bigras (2007)	457	174	211	43 (12)	Canada	1995–2004	2.1 (2.4)	3.5 (2.5)
Boer (2003)	296	120	164	41 (12)	Canada	1976–1994	2.8 (2.8)	3.9 (2.7)
Hanson et al. (2015)	710	247	342	42 (13)	Canada	2001–2005	2.4 (2.4)	3.5 (2.5)
Lehmann et al. (2013)	932	473	396	38 (12)	Germany	1994–2009	3.4 (2.2)	4.1 (1.9)
Total	2,395	1,014	1,113	40 (12)	—	1976–2009	2.8 (2.4)	3.8 (2.3)

al. (2003) Static-99 coding manual. STATIC scores were coded by the author of the original study (a graduate student in forensic psychology) from archival file information. No interrater reliability information was provided.

Hanson, Helmus, and Harris (2015). This sample included 710 sexual offenders from Canada who started a period of community supervision (probation or parole) between 2001 and 2005. Static-99 and STABLE-2000 data (described further below) were submitted by 139 supervision officers as part of their routine supervision practices. In addition to Static-99 and STABLE-2000 assessments, community supervision officers also provided information on other descriptive variables analyzed in the current study, including psychiatric history (defined as whether the offender had ever been hospitalized overnight for a psychiatric issue), whether the offender had ever been diagnosed as developmentally delayed, and whether the offender had been separated from their biological parents prior to the age of 16. Rater reliability was assessed by six experts rescoring 92 cases submitted to the study. Overall, the rater reliability for total scores was high: for Static-99, the intraclass correlation (ICC) was .91 (n = 88), and for STABLE-2000, the ICC was .89 (n = 87). The ICC for the individual STABLE-2000 items ranged from .66 to .92 (*Mdn* = .83). These findings likely overestimate rater reliability because the second raters used case files prepared by the officer who originally scored the case. As well, although the second raters were directed to make independent ratings, the original scores were contained in the case files used for the reliability analysis.

Static-2002 assessments were coded by two research assistants from victim information provided by the Static-99 assessments and from criminal history records. Rater reliability was exceptionally high (ICC = .98, n = 25), likely because the sources of information used (existing victim ratings from Static-99 assessments and criminal history records with no offense descriptions) reduced the amount of interpretation typically required for this task.

Lehmann et al. (2013). This sample included 936 sexual offenders reported to the Berlin state police between 1994 and 2001 who were convicted for a violent or abusive sexual offense. Approximately 77% of the sample was German citizens, 20% was foreign nationals, and 3% had a dual citizenship. The sexual offenders were convicted of sexual abuse of children or adolescents in 42% of cases; of sexual assault, rape, or similar sexual offenses toward adults in 51% of cases, and of both in 7% of cases.

The original data set contained over 300 variables related to crime scene behavior, offense history, and recidivism. The data were coded by trained research assistants (graduates students in psychology) using a standardized coding manual. STATIC scores were subsequently computed via syntax. The percent agreement for the overall set of variables was high (median kappa values >.90; Lehmann, 2014). Recidivism data were collected from the National Conviction Registry in Germany, and sexual recidivism was defined as any reconviction for a sexual offense (including hands-off sexual offending) during the follow-up period of 9.6 years (*SD* = 3.2).

Measures

Static-99R (Hanson & Thornton, 2000; Helmus, Thornton, et al., 2012). Static-99R (see www.static99.org) is a 10-item actuarial scale that assesses recidivism risk of adult male sexual offenders who have committed a sexually motivated offense against an identifiable victim (e.g., sexual assault against adults or children, voyeurism, exhibitionism; offenses without an identifiable victim or sexual motive are excluded, such as consenting sex among similar-aged peers, mooning, or streaking without sexual motive). Static-99R contains items assessing age at release, sexual criminality (e.g., prior sex offenses, victim information), and general criminality (e.g., prior sentencing dates, nonsexual violence; see Appendix for full list of items). A meta-analysis found that Static-99R has moderate predictive accuracy for sexual recidivism (mean AUC = .70, k = 22, N = 8,055; Helmus, Hanson, et al., 2012).

Static-2002R (Hanson & Thornton, 2003; Helmus, Thornton, et al., 2012). Similar to Static-99R, Static-2002R is an empirical actuarial risk assessment tool for adult male sex offenders (see also www.static99.org). It has 14 items grouped into five main subscales: Age at Release, Persistence of Sex Offending, Sexual Deviance, Relationship to Victims, and General Criminality (see Appendix). Static-2002R also has moderate predictive accuracy for sexual recidivism (mean AUC = .70, k = 7, N = 2,609; Babchishin et al., 2012b).

STABLE-2007 (Fernandez, Harris, Hanson, & Sparks, 2014; Hanson, Harris, Scott, & Helmus, 2007). The STABLE-2007 is an empirical actuarial risk tool assessing dynamic risk factors among adult male sex offenders. The scale was developed by revising the STABLE-2000 scale based on preliminary results from the longitudinal project included in this study (Hanson et al., 2007, 2015). The STABLE-2007 has 13 items organized into five subsections (significant social influences, intimacy deficits, sexual self-regulation, general self-regulation, and cooperation with supervision). Total scores on the STABLE-2007 are calculated by summing all item scores, and can range from 0 to 26 for offenders with child victims and 0 to 24 for other sexual offender subtypes.

Results

Quantitative risk indicators for the new risk categories for Static-99R are summarized in Table 1 (Static-99R) and in Table 2 (Static-2002R). The estimated recidivism rates were 1% to 2% in the lowest risk category, up to 20% to 50% in the highest risk category. Notably, the quantitative indicators of risk information (recidivism rates, risk ratios, percentiles) were now similar for the categories of both Static-99R and Static-2002R. The Kendall's tau-b correlation between Static-99R and Static-2002R total scores was .804 (p < .001).

As shown in Table 4, the concordance of risk classification was higher for the new risk categories than the original categories. For the original risk categories, 51% of cases (n = 1,222) received the same risk category label in both scales. In contrast, 72% (n = 1,713) received the same risk category label in the revised category scheme. With the revised categories, only 13 cases (0.5%) were discrepant by more than one risk category. Specifically, 10 cases were considered well above average risk on Static-99R, but only average risk on Static-2002R, and three cases were considered well above average risk on Static-2002R, but average on Static-99R.

The ICC (two-way mixed model, single measures, absolute agreement) was significantly higher (p < .01) for the revised risk categories (ICC = .84, 95% CI [.83, .85]) compared with the original risk categories (ICC = .73, 95% CI [.62, .80]; in this

Table 4
Correspondence Between the Original and New Risk Categories for Static-99R and Static-2002R

			Static-99R score											
			Low					Low- Moderate		Moderate- High		High		
			Very low		Below average		Average			Above average		Well above average		
Static-2002R score			−3	−2	−1	0	1	2	3	4	5	6	7	8+
−2	Low	Very low risk	***29***	***2***	*1*	*1*	—	—	—	—	—	—	—	—
−1	Low	Very low risk	***6***	***29***	*4*	—	—	—	—	—	—	—	—	—
0	Low	Below average risk	—	*10*	***87***	***20***	*1*	—	—	—	—	—	—	—
1	Low	Below average risk	—	*3*	***44***	***103***	*31*	5	—	—	—	—	—	—
2	Low	Average risk	—	—	*3*	*85*	***128***	**68**	**14**	3	—	—	—	—
3	Low-Mod	Average risk	—	—	—	11	**108**	***140***	***83***	27	2	1	—	—
4	Low-Mod	Average risk	—	—	—	—	**18**	***104***	***176***	126	30	8	1	—
5	Moderate	Above average risk	—	—	—	—	1	18	110	**141**	**71**	29	5	4
6	Moderate	Above average risk	—	—	—	—	—	1	18	**53**	**83**	47	14	11
7	Mod-High	Well above average	—	—	—	—	—	—	3	*13*	*37*	**53**	**22**	**12**
8	Mod-High	Well above average	—	—	—	—	—	—	—	*2*	*13*	**18**	**19**	**15**
9	High	Well above average	—	—	—	—	—	—	—	1	1	***6***	***18***	***21***
10+	High	Well above average	—	—	—	—	—	—	—	—	1	**—**	***5***	***17***

Note. $n = 2{,}395$. The original risk categories for Static-2002R were Low, Low-Moderate, Moderate, Moderate-High, and High; for Static-99R, the original risk categories were Low, Low-Moderate, Moderate-High, and High. Bold font represent cases for which the new risk categories were the same for both scales. Italic font represent cases for which the old risk categories were the same for both scales.

analysis, to reflect absolute agreement, Static-99R risk categories were considered on a 5-point scale, with no cases in the moderate risk category). Notably, however, even the absolute agreement ICC would still be influenced by the rank ordering of risk categories. Kappa between the revised STATIC risk categories was .59; kappa cannot be calculated for the original risk categories because the two scales differ in the number of category options.

Characteristics of Category Members

As expected, category members showed strong differences on basic demographic and offense history information (see Table 5). These associated characteristics were very similar regardless of the risk tool (Static-99R or Static-2002R) used to assign category membership.

The lowest risk category (Category I) exclusively contained sexual offenders over the age of 60 with no prior sexual offenses who had offended against an acquaintance or family member. Category II offenders were younger than Category I offenders, but still older ($M = 50$ years) than average (40 years). About four of 10 Category II offenders had some prior involvement with the criminal justice system (in addition to their index sexual offense); however, it was rare for Category II offenders to have a prior sexual offense or strangers as victims. The majority of Category I and Category II offenders had sexual victimized children, not adults.

Category III contained a mixture of offenders against children and offenders against adults. Most members of Category III had some prior involvement with the criminal justice system, although only a minority had prior sexual offense convictions or stranger victims. Their age ranged from 18 to 80, with an average close to the average for the complete sample (39 to 40 years).

Table 5
Descriptive Data for the STATIC Risk Categories

		Age at release							
Category	*n*	*M*	*SD*	Range		% With child victim	% With prior involvement in criminal justice system	% With prior sex sentencing occasion	% With stranger victim
Static-99R									
I	79	67.5	5.2	60	78	79.7	20.3	.0	.0
II	359	50.5	9.6	35	84	70.3	47.6	3.3	2.8
III	1,027	39.5	10.4	18	78	46.9	74.6	14.5	16.0
IVa	604	35.0	11.0	18	74	33.8	88.9	30.5	45.4
IVb	326	36.0	10.0	18	62	38.6	99.1	59.2	72.1
Static-2002R									
I	72	67.9	5.5	60	84	79.2	12.5	.0	.0
II	304	51.0	10.0	35	84	70.7	36.5	1.0	3.9
III	1,136	39.6	10.9	18	80	45.6	73.7	11.5	15.2
IVa	606	35.0	10.4	18	74	35.2	95.9	30.7	50.7
IVb	277	37.1	10.4	18	68	44.5	99.3	78.7	69.0

Category IV-a offenders were slightly younger than average (35 years) and nine of 10 had some prior involvement with the criminal justice system. Three of 10 had prior sexual offense convictions, and 5 of 10 had victimized strangers. More rapists than sex offenders against children were found in this category (ratio of 2 to 1).

Sexual offenders in the highest risk category (IV-b) were only slightly younger than average (36 to 37 years old) and had extensive criminal histories. Virtually all had prior involvement with the criminal justice system, most had prior sexual offense convictions, and most had victimized strangers.

As expected, there were also strong differences between the risk categories on the psychological characteristics associated with sexual recidivism risk, such as sexual preoccupation, lack of cooperation with supervision, and impulsivity (see Table 6). For the lowest risk categories (I and II), most (75% to 93%) did not display problems in these areas, whereas these problems were present for most (>60%) of highest risk (Category IVb) sexual offenders. In contrast, there was much less meaningful variation across categories for variables only weakly related to sexual crime, such as psychiatric history and developmental delay. Although these (largely noncriminogenic) problems were positively associated with the risk categories, the absolute rates remained low (10% to 13.8%) even for the highest risk offenders (see Table 6).

The results, however, were not entirely consistent with expectations. Although the rates of criminogenic problems were relatively lower in the lower risk categories than the higher risk categories, problems were still prevalent in the lowest risk categories. One of four of the Category I offenders had some problems with sexual preoccupation; half the Category I offenders had some problems with intimate relationships. As well, the overall density of criminogenic needs (as indicated by STABLE-2007 totals scores) was lower (not higher) for Category II compared with Category I.

Nevertheless, the overall results indicated that the frequency with which risk factors were present in the different categories was very similar regardless of which risk scale was used to infer category membership (Static-99R or Static-2002R).

Discussion

This study was motivated by the practical concern of revising risk categories for existing sexual offender risk assessment tools in light of new research findings (e.g., Should the thresholds increase if the overall recidivism rates decline? Should the range associated with moderate risk shift based on observed changes in the distribution of raw scores?). As well, we wanted a principled method of comparing the results of different risk measures (Babchishin et al., 2012b; Lehmann et al., 2013). Although there is considerable research on quantifying and communicating the results of norm referenced tests, we were unable to find similar resources for criterion referenced prediction measures. The literature reviewed in this article, however, suggested that there were certain principles that could inform standardized risk categories for criterion referenced prediction tools. We then used these principles to develop new, common risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. Although the total scores on the tools were highly correlated ($r > .80$), the concordance of the original risk categories was only 50%. The concordance increased to 70% for the new categories. More importantly, evaluators could now make the same inferences for sexual offenders in the same category regardless of which instrument was used to assign category membership. This is progress.

Although prediction tools are fundamentally justified by their relationship to the outcome of interest, risk scales contain more information than a likelihood of recidivism. Other quantitative indicators include percentile ranks (like norm referenced tests) and relative risk ratios. The most useful risk categories would also have

Table 6
Additional Descriptive Data for the STATIC Risk Categories From Hanson et al. (2015) Data Set

		STABLE-2007 score			% Offenders with at least some concern in these need areas						
Category	*n*	*M*	*SD*	*Mdn*	Major mental illness	Developmental delay	Separated from parents before 16	Sexual preocc.	Lack of cooperation with supervision	Impulsivity	Capacity for relationship stability
Static-99R											
I	28	5.9	4.7	4	3.6	.0	7.1	25.0	25.0	7.1	46.4
II	98	4.9	3.4	4	7.3	1.0	19.6	24.5	17.3	15.3	56.1
III	272	6.8	4.6	6	10.1	3.4	28.7	34.9	22.4	32.4	74.3
IVa	111	9.2	4.8	8	11.8	9.1	36.0	42.3	39.6	56.8	83.8
IVb	61	12.6	5.1	12	11.9	10.0	47.5	73.8	60.7	55.7	90.2
Static-2002R											
I	28	5.4	4.7	4	3.6	.0	7.1	25.0	17.9	10.7	46.4
II	88	5.1	3.7	4	10.3	1.1	17.0	25.0	15.9	14.8	60.2
III	270	6.8	4.6	6	7.9	3.8	26.4	36.7	22.6	29.6	74.1
IVa	125	9.0	4.7	8	13.1	6.5	38.4	40.0	37.6	56.0	81.6
IVb	59	12.4	5.5	12	13.8	12.1	54.2	67.8	66.1	61.0	84.7

Note. Cases with missing data: For history of major mental illness, 2 from Static-99R Category II, 4 from Static-99R Category III, 1 from Static-99R Category IVa, 1 from Static-99R Category IVb, 1 from Static-2002R Category II, 4 from Static-2002R Category III, 3 from Static-2002R Category IVa, and 1 from Static-2002R Category IVb. For developmental delay, 1 from Static-99R Category I, 1 from Static-99R Category II, 6 from Static-99R Category III, 1 from Static-99R Category IVa, 1 from Static-99R Category IVb, 1 from Static-2002R Category I, 6 from Static-2002R Category III, 2 from Static-2002R Category IVa, and 1 from Static-2002R Category IVb. For separation from parents before the age of 16, 1 from Static-99R Category II and 1 from Static-2002R Category III. Preocc. = preoccupation.

construct validity, that is, a theoretically integrated set of inferences relevant to the purpose for which the tool was made (in our case, the assessment and treatment of sexual offenders). We have demonstrated how these diverse features of prediction tools can inform the development of risk categories, with the result of increased interpretability.

We believe that the new STATIC categories have sufficiently improved conceptual coherence and have sufficient empirical support to replace the original categories in applied assessments. In other words, the categories in this article are the official versions, that is, those endorsed by the tests' developers. This does not mean that decision makers are compelled to change decisions based on the new compared with the previous risk category labels. The information associated with each specific score has not changed. We hope, however, that the change in risk labels motivates all decision makers in corrections, mental health, child welfare, and public safety to carefully consider the meaning of the risk categories that are currently used for sexual offenders.

In our revised risk categories, sexual offenders in different categories would be expected to be meaningfully different on risk-relevant propensities and, as such, require different intervention strategies. The lowest risk category (those with the same recidivism rates as nonsexual offenders), for example, would require no specialized interventions. In contrast, Category IVa and IVb would be expected to require intensive interventions to reduce risk. Importantly, these risk categories have explicit definitions, which can be challenged and refined based on advances in theory and research. For example, better measures and longer follow-up periods may identify a sample of sexual offenders with an expected recidivism rate of 90% or higher and, as such, a Category V could be created with the same definition as the Justice Center's Category V (virtually certain to reoffend). In addition, reclassification of offenders' risk category may be warranted based on treatment and time-free effects (e.g., Hanson, Harris, Helmus, & Thornton, 2014).

We proposed five risk categories, but three categories may be sufficient for some settings. For example, the two lowest and two highest categories could be collapsed to classify sexual offenders into those who do not need intervention (or much intervention), those who require typical intervention practices, and those who need substantial interventions. These categories represent, respectively, about one quarter, one half, and one quarter of sexual offenders in the routine validation samples of STATIC scales. Decisions regarding collapsing risk categories would be dependent on policy and jurisdiction needs, and should be explicitly noted so it is clear how the jurisdiction-specific categories relate to the ones asserted by the scale developers (i.e., someone from another setting reviewing risk assessment reports would not confuse a jurisdiction-specific classification with the five-category classification described here).

A common language for risk communication would also advance our understanding of the Risk Principle in Andrews, Bonta, and Hoge's (1990) Risk, Need, and Responsivity model of effective correctional intervention. The Risk Principle states, quite simply, that the intensity of intervention should be proportional to the risk of recidivism. Although meta-analyses have supported the Risk Principle in general (Andrews & Bonta, 2010; Hanson, Bourgon, Helmus, & Hodgson, 2009), there is very little evidence concerning how much treatment is required for which risk level. Identifying the necessary and sufficient dose of intervention requires a common metric for describing and quantifying risk, which has not previously been available. We believe that the principles used to create the risk categories in the current article could provide a solid foundation for future studies concerning how much correctional intervention should be given and to whom.

Because we propose a psychologically meaningful definition of risk categories, construct validity of these risk categories can be further researched (Cronbach & Meehl, 1955; Embretson, 2010). The current study pointed to this research agenda by demonstrating strong differences across risk levels in sex-crime-specific psychological risk factors. Further research could profitably explore how sexual offenders in these risk categories differ on the latent, risk-relevant constructs assessed by risk tools. These risk categories may be further refined to better incorporate our knowledge of risk dimensions. A focus on construct validity can also be expected to provide greater consensus in interpretation of risk categories across professionals and settings, irrespective of the risk tools being utilized.

Our hope is that standardized risk categories will focus scientific and professional discussion on the psychological attributes of offenders, and away from the risk scale scores. Routinely, we receive STATIC scoring questions in which a disputed difference of a single STATIC point could result in lifetime incapacitation or freedom in the community. Instead of defining policy relevant risk levels in terms of Static-99R scores, it would be better to define risk levels based on standardized risk categories. By focusing debates on the risk categories, it is possible to ask questions about the characteristics associated with the different categories, and the accuracy of different assessment procedures for assigning category membership. Previously, the lack of clearly defined risk levels made it difficult for evaluators and decision makers to know what information external to a risk scale was relevant to risk classification decisions.

As found in the current study, one predictable attribute of the lowest risk offenders is age. With the STATIC measures, all the Category I offenders were over 60, and all the Category II offenders were over 35. This pattern is expected given the strong decline in sexual recidivism risk with advanced age (Barbaree et al., 2007, 2009; Hanson, 2006). This finding does not mean, however, that only sexual offenders over the age of 60 are in the very-low-risk group. The STATIC risk tools only address a limited range of demographic and criminal history variables among recently released sexual offenders, and it is possible that evaluations based a broader set of variables (e.g., current community adjustment, dynamic risk factors) could identify sexual offenders in their 20s who also fit the Category I profile. For example, there are strong, predictable declines in the risk for sexual recidivism based on years offense-free in the community (Hanson et al., 2014). Although structured methods for empirically identifying Category I sexual offenders have yet to be developed, it is likely that many sexual offenders would merit this label given a few years of positive community adjustment, and that the risk of most sexual offenders would reduce to that of nonsexual offenders after 10 years sex-offense-free in the community (Hanson et al., 2014).

The current study focused on creating common categories for two sexual offender risk assessment tools; however, we hope this study motivates other test developers and users to consider the meaning of the risk categories for the diverse offender risk tools now widely used in corrections and forensic mental health. As we demonstrated in the current study, disagreements between measures can arise simply from different risk categories, and need not be related to differences in the information implied by the scales. Standardized risk categories should equally apply to empirical actuarial measures as well as SPJ measures. The only difference would be the methods used to assign category membership (empirical/mechanical vs. professional judgment).

Test developers interested in mechanical methods of developing risk categories could easily follow the heuristics described in the current study for sexual recidivism measures, or the heuristics for general recidivism measures being developed by the Justice Center (2016). These approaches are relatively simple to compute given large recidivism studies (100+ recidivists) with representative samples of the population of interest. Such norming projects, however, raise further questions concerning the population to which these measures apply (e.g., All offenders? All male offenders? All male offenders in the United States? All convicted male sexual offenders in Missouri?). These are important questions that need to be addressed in future research and policy development.

The purpose of this study was to articulate principles for developing offender risk levels and to demonstrate their utility in guiding the construction of new risk categories for Static-99R and Static-2002R. Although the principles were expected to apply to a wide range of offender risk measures, the tools examined in the current study were very similar (both predicted sexual recidivism using static factors). Consequently, the concordance of the new categories was less noteworthy (.72) than the low concordance (.51) of the original categories. Further research is needed to determine whether applying these principles to more diverse risk tools would similarly increase the concordance of risk classification.

Another limitation was that the new risk category levels for the STATIC measures assumed that sexual offender risk is a single dimension, whereas research indicates that it is better represented by at least two dimensions (Hanson & Morton-Bourgon, 2005) and probably more (Brouillette-Alarie et al., 2016). Furthermore, the STATIC risk scales focus on only one aspect of risk, that is, likelihood of sexual recidivism, and did not consider other elements of risk, such as severity, imminence, or other types of recidivism.

Although we showed that the STATIC categories correlated as expected with psychologically meaningful characteristics, only a limited set of variables were available in the current study. Many of the proposed features of the standardized risk categories have yet to be examined. For example, the measure of criminogenic needs used in the current study (STABLE-2007) did not discriminate between the two lowest risk categories (Level I vs. Level II). STABLE-2007 items are scored according to the presence or absence of clinically significant problems. It is likely, however, that the absence of problems is not synonymous with prosocial strengths. Consequently, it is worth exploring whether risk tools that explicitly measures strengths and good community adjustment (e.g., SAPROF; de Vries Robbé, de Vogel, Koster, & Bogaerts, 2015) help evaluators separate out Level I offenders from the larger group of Level II offenders.

Another untested assertion concerns the required intensity of intervention and supervision for offenders in the different risk levels. Based on the limited research available (see review by Hanson & Yates, 2013), we believe our proposals are plausible. Nevertheless, further research using standardized risk categories may conclude that substantially more or substantially less intervention is required to successfully address the recidivism potential of sexual offenders.

Conclusions

Like temperature measurement in the 17th century, the field of offender risk assessment currently lacks a common language by which to communicate recidivism risk. We believe that such a language is possible. We further believe that such a language should be informed by the quantitative information implied by risk scale scores (e.g., percentiles, risk ratios, recidivism rates), by a scientific understanding of the constructs being assessed, and by the purposes for which the measures are intended to be used. Standardized risk categories would allow professionals to better understand offenders' risk of recidivism, allow evaluators to compare results of different risk tools, and allow decision makers within the criminal justice, mental health, and child welfare systems to better understand risk assessment reports. We hope that this article motivates us all to carefully consider the language we use, and to advance our understanding of the riskiness we attribute to others.

Although this article focused on offender risk assessment, the problem of risk category labels must be addressed by all prognostic and prediction measures. Whereas there is professional consensus that norm referenced measures position individuals within groups, we have yet to develop a similar consensus for criterion referenced prediction measures. The preferred language for communicating the likelihood of an outcome will obviously vary based on the context, the outcome, and the audience; nevertheless, we believe that the principles and metrics presented in this article can be used to advance how we understand and communicate the findings of criterion referenced prediction measures from diverse fields of psychological assessment.

References

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-68560-1

Akobeng, A. K. (2008). Communicating the benefits and harms of treatments. *Archives of Disease in Childhood, 93,* 710–713. http://dx.doi.org/10.1136/adc.2008.137083

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). New Providence, NJ: LexisNexus Mathew Bender.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19–52. http://dx.doi.org/10.1177/0093854890017001004

Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Moving beyond the low, moderate, and high typology of risk communication. *Crime Scene, 16,* 11–14. Retrieved from http://www.cpa.ca/cpasite/userfiles/Documents/Criminal%20Justice/Crime%20Scene%202009-05(1).pdf

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012a). Communicating risk for sex offenders: Risk ratios for Static-2002R. *Sexual Offender Treatment, 7,* 1–12.

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

Barbaree, H. E., Langton, C. M., & Blanchard, R. (2007). Predicting recidivism in sex offenders using the VRAG and SORAG: The contribution of age-at-release. *International Journal of Forensic Mental Health, 6,* 29–46. http://dx.doi.org/10.1080/14999013.2007.10471247

Barbaree, H. E., Langton, C. M., Blanchard, R., & Cantor, J. M. (2009). Aging versus stable enduring traits as explanatory constructs in sex offender recidivism: Partitioning actuarial prediction into conceptually meaningful components. *Criminal Justice and Behavior, 36,* 443–465. http://dx.doi.org/10.1177/0093854809332283

Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006). Different actuarial risk measures produce different risk rankings for sexual offenders. *Sexual Abuse: Journal of Research and Treatment, 18,* 423–440.

Berman, A. L., & Silverman, M. M. (2014). Suicide risk assessment and risk formulation part II: Suicide risk formulation and the determination of levels of risk. *Suicide and Life-Threatening Behavior, 44,* 432–443. http://dx.doi.org/10.1111/sltb.12067

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1,* 257–269. http://dx.doi.org/10.1002/for.3980010305

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)

Blais, J., & Forth, A. E. (2014). Prosecution-retained versus court-appointed experts: Comparing and contrasting risk assessment reports in preventative detention hearings. *Law and Human Behavior, 38,* 531–543. http://dx.doi.org/10.1037/lhb0000082

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41. http://dx.doi.org/10.1037/0003-066X.61.1.27

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A. K., & Jeglic, E. L. (2012). Implications of Static-99 field reliability findings for score use and reporting. *Criminal Justice and Behavior, 39,* 42–58. http://dx.doi.org/10.1177/0093854811427131

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

Bonta, J., & Hanson, R. K. (1995, August). *Violent recidivism of men released from prison*. Paper presented at the 103rd Annual Convention of the American Psychological Association, New York, NY.

Bonta, J., Rugge, T., & Dauvergne, M. (2008). Sexual recidivism of 11,909 Canadian Federal offenders with and without a prior conviction for a sexual offence. [Unpublished raw data].

Brouillette-Alarie, S., Babchishin, K. M., Hanson, R. K., & Helmus, L. (2016). Latent constructs of the Static-99R and Static-2002R: A three-factor solution. *Assessment, 23,* 96–111. http://dx.doi.org/10.1177/1073191114568114

Budescu, D. V., Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change, 113,* 181–200. http://dx.doi.org/10.1007/s10584-011-0330-3

Center for Sex Offender Management. (2010). *Exploring public awareness and attitudes about sex offender management: Findings from a national public opinion poll*. Washington, DC: Center for Effective Public Policy.

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior, 39,* 209–218. http://dx.doi.org/10.1037/lhb0000114

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist, 23,* 193–204. http://dx.doi.org/10.1080/13854040801968450

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist, 23,* 1173–1195. http://dx.doi.org/10.1080/13854040902795018

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180. http://dx.doi.org/10.1037/0003-066X.60.2.170

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

de Vries Robbé, M., de Vogel, V., Koster, K., & Bogaerts, S. (2015). Assessing protective factors for sexually violent offending with the SAPROF. *Sexual Abuse: Journal of Research and Treatment, 27,* 51–70. http://dx.doi.org/10.1177/1079063214550168

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20^{V3}: Assessing risk for violence: User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K. S., & Ogloff, J. R. P. (2003). Multiple facets of risk for violence: The impact of judgmental specificity on structured decisions about violence risk. *The International Journal of Forensic Mental Health, 2,* 19–34. http://dx.doi.org/10.1080/14999013.2003.10471176

Doyle, D. J., Ogloff, J. R. P., & Thomas, S. D. M. (2011). An analysis of dangerous sexual offender assessment reports: Recommendations for best practice. *Psychiatry, Psychology and Law, 18,* 537–556. http://dx.doi.org/10.1080/13218719.2010.499159

Duwe, G. (2012). Predicting first-time sexual offending among prisoners without a prior sex offense history: The Minnesota Sexual Criminal Offending Risk Estimate (MnSCORE). *Criminal Justice and Behavior, 39,* 1436–1456. http://dx.doi.org/10.1177/0093854812453911

Embretson, S. E. (Ed.). (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12074-000

Falshaw, L., Bates, A., Patel, V., Corbett, C., & Friendship, C. (2003). Assessing reconviction, reoffending and recidivism in a sample of UK sexual offenders. *Legal and Criminological Psychology, 8,* 207–215. http://dx.doi.org/10.1348/135532503322362979

Fernandez, Y., Harris, A. J. R., Hanson, R. K., & Sparks, J. (2014). *STABLE-2007 coding manual: Revised 2014* [Unpublished manual]. Ottawa, Ontario, Canada: Public Safety Canada.

Gillies, D. (2000). Varieties of propensity. *The British Journal for the Philosophy of Science, 51,* 807–835. http://dx.doi.org/10.1093/bjps/51.4.807

Gjerdrum, D., & Peter, M. (2011). The new international standard on the practice of risk management–A comparison of ISO 31000: 2009 and the COSO ERM framework. *Risk Management, 31,* 8–13.

Greenland, S. (1998). Probability logic and probabilistic induction. *Epidemiology (Cambridge, Mass.), 9,* 322–332. http://dx.doi.org/10.1097/00001648-199805000-00018

Guay, J. P., Ruscio, J., Knight, R. A., & Hare, R. D. (2007). A taxometric analysis of the latent structure of psychopathy: Evidence for dimensionality. *Journal of Abnormal Psychology, 116,* 701–716. http://dx.doi.org/10.1037/0021-843X.116.4.701

Hanson, R. K. (2001). *Note on the reliability of Static-99 as used by the California Department of Mental Health evaluators* [Unpublished report]. Sacramento, CA: California Department of Mental Health.

Hanson, R. K. (2006). Does Static-99 predict recidivism among older sexual offenders? *Sexual Abuse: Journal of Research and Treatment, 18,* 343–355.

Hanson, R. K. (2014). Risk assessment. In M. S. Carich & S. E. Mussack (Eds.), *The Safer Society handbook of sexual abuser assessment and treatment* (pp. 45–62). Brandon, VT: Safer Society.

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

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

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66,* 348–362. http://dx.doi.org/10.1037/0022-006X.66.2.348

Hanson, R. K., Harris, A. J., Helmus, L., & Thornton, D. (2014). High-risk sex offenders may not be high risk forever. *Journal of Interpersonal Violence, 29,* 2792–2813. http://dx.doi.org/10.1177/0886260514526062

Hanson, R. K., Harris, A. J. R., Scott, T.-L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (User Report No. 2007–05). Ottawa, Ontario, Canada: Public Safety Canada.

Hanson, R. K., Helmus, L. M., & Harris, A. J. R. (2015). Assessing the risk and needs of supervised sexual offenders: A prospective study using STABLE-2007, Static-99R, and Static-2002R. *Criminal Justice and Behavior, 42,* 1205–1224. http://dx.doi.org/10.1177/0093854815602094

Hanson, R. K., Helmus, L., & Thornton, D. (2010). Predicting recidivism amongst sexual offenders: A multi-site study of static-2002. *Law and Human Behavior, 34,* 198–211. http://dx.doi.org/10.1007/s10979-009-9180-1

Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk scales. *The International Journal of Forensic Mental Health, 11,* 9–23. http://dx.doi.org/10.1080/14999013.2012.667511

Hanson, R. K., & Morton-Bourgon, K. E. (2004). *Predictors of sexual recidivism: An updated meta-analysis* (Corrections Research User Report No. 2004–02). Ottawa, ON: Public Safety and Emergency Preparedness Canada.

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73,* 1154–1163. http://dx.doi.org/10.1037/0022-006X.73.6.1154

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. http://dx.doi.org/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. http://dx.doi.org/10.1023/A:1005482921333

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

Hanson, R. K., Thornton, D., Helmus, L. M., & Babchishin, K. M. (2016). What sexual recidivism rates are associated with Static-99R and Static-2002R scores? *Sexual Abuse, 28,* 218–252. http://dx.doi.org/10.1177/1079063215574710

Hanson, R. K., & Yates, P. M. (2013). Psychological treatment of sex offenders. *Current Psychiatry Reports, 15,* 348. http://dx.doi.org/10.1007/s11920-012-0348-x

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

Hart, S. D., & Boer, D. P. (2010). Structured professional judgment guidelines for sexual violence risk assessment: The Sexual Violence Risk-20 (SVR-20) and Risk for Sexual Violence Protocol (RSVP). In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 269–294). New York, NY: Routledge.

Heilbrun, K., O'Neill, M. L., Strohman, L. K., Bowman, Q., & Philipson, J. (2000). Expert approaches to communicating violence risk. *Law and Human Behavior, 24,* 137–148. http://dx.doi.org/10.1023/A:1005435005404

Heilbrun, K., Philipson, J., Berman, L., & Warren, J. (1999). Risk communication: Clinicians' reported approaches and perceived values. *Journal of the American Academy of Psychiatry and the Law, 27,* 397–406.

Helmus, L. (2007). *A multi-site comparison of the validity and utility of the Static-99 and Static-2002 for risk assessment with sexual offenders* (Unpublished honors thesis). Carleton University, Ottawa, Ontario, Canada.

Helmus, L. & Babchishin, K. M. (in press). What statistics should we use to evaluate the information provided by risk scales? *Criminal Justice and Behavior.*

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

Hilton, N. Z., Carter, A., Harris, G. T., & Sharpe, A. J. B. (2008). Does using nonnumerical terms to describe risk aid violence risk communication? Clinician agreement and decision making. *Journal of Interpersonal Violence, 23,* 171–188. http://dx.doi.org/10.1177/0886260507309337

Hilton, N. Z., Scurich, N., & Helmus, L. M. (2015). Communicating the risk of violent and offending behavior: Review and introduction to this special issue. *Behavioral Sciences & the Law, 33,* 1–18. http://dx.doi.org/10.1002/bsl.2160

Imrey, P. B., & Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy, 2,* 1–18. http://dx.doi.org/10.1080/2330443X.2015.1029338

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, 19,* 425–448.

Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Justice Center. (2014). *A common language for risk assessment: Experts convene in Washington*. Retrieved from http://csgjusticecenter.org/reentry/posts/a-common-language-for-risk-assessments-experts-convene-in-washington/

Justice Center. (2016). *A five-level risk and needs system: Maximizing assessment results through the development of a common language*. New York, NY: Author.

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: Applied, 10,* 25–41. http://dx.doi.org/10.1037/1076-898X.10.1.25

King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., Weich, S., . . . Nazareth, I. (2008). Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: The PredictD study. *Archives of General Psychiatry, 65,* 1368–1376. http://dx.doi.org/10.1001/archpsyc.65.12.1368

Knighton, J. C., Murrie, D. C., Boccaccini, M. T., & Turner, D. B. (2014). How likely is "likely to reoffend" in sex offender civil commitment trials? *Law and Human Behavior, 38,* 293–304. http://dx.doi.org/10.1037/lhb0000079

Landsberg, H. E. (1964). A note on the history of thermometer scales. *Weather, 19,* 2–6. http://dx.doi.org/10.1002/j.1477-8696.1964.tb02713.x

Lehmann, R. J. B. (2014). *Using crime scene information for risk assessment in sexual offenders* (Unpublished doctoral dissertation). Freien Unversität, Berlin, Germany.

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

Lehmann, R. J. B., Thornton, D., Helmus, L. M., & Hanson, R. K. (2016). Developing non-arbitrary metrics for risk communication: Norms for the Risk Matrix 2000. *Criminal Justice and Behavior*. Advance online publication. http://dx.doi.org/10.1177/0093854816651656

Levenson, J. S. (2004). Reliability of sexually violent predator civil commitment criteria in Florida. *Law and Human Behavior, 28,* 357–368. http://dx.doi.org/10.1023/B:LAHU.0000039330.22347.ad

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9,* 563–564. http://dx.doi.org/10.3758/BF03327890

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

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with static-99R. *Sexual Abuse: Journal of Research and Treatment, 24,* 431–458. http://dx.doi.org/10.1177/1079063211432475

Meyer, P. (2010). *Reliability*. Oxford, UK: Oxford University Press.

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *The International Journal of Forensic Mental Health, 2,* 1–6. http://dx.doi.org/10.1080/14999013.2003.10471174

Monahan, J., & Steadman, H. J. (1996). Violent storms and violent people: How meteorology can inform risk communication in mental health law. *American Psychologist, 51,* 931–938. http://dx.doi.org/10.1037/0003-066X.51.9.931

Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *British Medical Journal (Clinical Research Ed.), 338,* b375. http://dx.doi.org/10.1136/bmj.b375

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15,* 19–53. http://dx.doi.org/10.1037/a0014897

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior, 41,* 1406–1421. http://dx.doi.org/10.1177/0093854814548449

Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment, 23,* 191–202. http://dx.doi.org/10.1177/1073191115580638

Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology, 21,* 913–938. http://dx.doi.org/10.1017/S0954579409000492

Phenix, A., Doren, D., Helmus, L., Hanson, R. K., & Thornton, D. (2009). *Coding rules for Static-2002*. Ottawa, Ontario, Canada: Public Safety Canada.

Phenix, A., & Epperson, D. L. (2015). Overview of the development, reliability, validity, scoring, and uses of the Static-99, Static-99R, Static-2002, and Static-2002R. In A. Phenix & H. M. Hoberman (Eds.), *Sexual offending: Predisposing conditions, assessments, and management* (pp. 437–455). New York, NY: Springer.

Phenix, A., Helmus, H., & Hanson, R. K. (2015). *Static-99R and Static-2002R evaluators' workbook* [Unpublished manual]. Retrieved from www.static99.org

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10,* 25–42. http://dx.doi.org/10.1093/bjps/X.37.25

Quesada, S. P., Calkins, C., & Jeglic, E. L. (2014). An examination of the interrater reliability between practitioners and researchers on the static-99. *International Journal of Offender Therapy and Comparative Criminology, 58,* 1364–1375. http://dx.doi.org/10.1177/0306624X13495504

Rice, A. K., Boccaccini, M. T., Harris, P. B., & Hawes, S. W. (2014). Does field reliability for Static-99 scores decrease as scores increase? *Psychological Assessment, 26,* 1085–1094. http://dx.doi.org/10.1037/pas0000009

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the VRAG and SORAG: The Violence Risk Appraisal Guide-Revised (VRAG-R). *Psychological Assessment, 25,* 951–965. http://dx.doi.org/10.1037/a0032878

Rockhill, B., Byrne, C., Rosner, B., Louie, M. M., & Colditz, G. (2003). Breast cancer risk prediction with a log-incidence model: Evaluation of accuracy. *Journal of Clinical Epidemiology, 56,* 856–861. http://dx.doi.org/10.1016/S0895-4356(03)00124-0

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology, 2,* 191–201. http://dx.doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2

Scurich, N., & Krauss, D. A. (2014). The presumption of dangerousness in sexually violent predator commitment proceedings. *Law Probability and Risk, 13,* 91–104. http://dx.doi.org/10.1093/lpr/mgt015

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., . . . Otto, R. K. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *The International Journal of Forensic Mental Health, 13,* 193–206. http://dx.doi.org/10.1080/14999013.2014.922141

Singh, J. P., Fazel, S., Gueorguieva, R., & Buchanan, A. (2013). Rates of sexual recidivism in high risk sex offenders: A meta-analysis of 10,422 participants. *Sexual Offender Treatment, 7,* 44–57.

Singh, J. P., Fazel, S., Gueorguieva, R., & Buchanan, A. (2014). Rates of violence in patients classified as high risk by structured risk assessment instruments. *The British Journal of Psychiatry, 204,* 180–187. http://dx.doi.org/10.1192/bjp.bp.113.131938

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24,* 271–296. http://dx.doi.org/10.1023/A:1005595519944

Smid, W. J., Kamphuis, J. H., Wever, E. C., & Van Beek, D. J. (2014). A comparison of the predictive properties of nine sex offender risk assessment instruments. *Psychological Assessment, 26,* 691–703. http://dx.doi.org/10.1037/a0036616

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293. http://dx.doi.org/10.1126/science.3287615

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: A meta-analysis. *Journal of Risk Research, 5,* 177–186. http://dx.doi.org/10.1080/13669870110038179

Thomas, D. A. (2003). Predicting law school academic performance from LSAT scores and undergraduate grade point averages: A comprehensive study. *Arizona State Law Journal, 35,* 1007–1028.

Thornton, D. (2006). Age and sexual recidivism: A variable connection. *Sexual Abuse: Journal of Research and Treatment, 18,* 123–135.

Thornton, D., & Knight, R. A. (2015). Construction and validation of SRA-FV need assessment. *Sexual Abuse: Journal of Research and Treatment, 27,* 360–375. http://dx.doi.org/10.1177/1079063213511120

Varela, J. G., Boccaccini, M. T., Cuervo, V. A., Murrie, D. C., & Clark, J. W. (2014). Same score, different message: Perceptions of offender risk depend on Static-99R risk communication format. *Law and Human Behavior, 38,* 418–427. http://dx.doi.org/10.1037/lhb0000073

Visschers, V. H., Meertens, R. M., Passchier, W. W., & de Vries, N. N. (2009). Probability information in risk communication: A review of the research literature. *Risk Analysis, 29,* 267–287. http://dx.doi.org/10.1111/j.1539-6924.2008.01137.x

Weinberger, D. M., Harboe, Z. B., Sanders, E. A., Ndiritu, M., Klugman, K. P., Rückinger, S., . . . Lipsitch, M. (2010). Association of serotype with risk of death due to pneumococcal pneumonia: A meta-analysis. *Clinical Infectious Diseases, 51,* 692–699. http://dx.doi.org/10.1086/655828

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

Appendix

Items of the Static-99R and Static-2002R

STATIC-99R	STATIC-2002R
1. Offender's age at release	1. Offender's age at release
2. Four or more prior sentencing dates	2. Prior sentencing occasions for anything
3. Number of prior sexual offence charges and convictions	3. Prior sentencing occasions for sexual offences
4. Any unrelated victims of sexual assaults	4. Any unrelated victims of sexual assaults
5. Any male victims of sexual assaults	5. Any male victims of sexual assaults
6. Convictions for non-contact sexual offences	6. Convictions for non-contact sexual offences
7. Any stranger victims of sexual assaults	7. Any stranger victims of sexual assaults
8. Conviction for non-sexual violence prior to the Index Offence	8. Prior violent non-sexual sentencing occasion
9. Conviction for non-sexual violence at the time of the Index Offence	9. Any prior involvement with the criminal justice system
10. Ever lived with an intimate partner for two consecutive years	10. Any young, unrelated victims
	11. Rate of sexual offences
	12. Any community supervision violation
	13. Arrests for sexual offences as both an adult and a juvenile
	14. Years free prior to Index

Received September 25, 2015
Revision received June 28, 2016
Accepted June 28, 2016 ■