INTERNATIONAL JOURNAL OF FORENSIC MENTAL HEALTH, 11: 9–23, 2012
Copyright © 2012 Crown copyright
ISSN: 1499-9013 print / 1932-9903 online
DOI: 10.1080/14999013.2012.667511

Routledge
Taylor & Francis Group

Developing Non-Arbitrary Metrics for Risk Communication: Percentile Ranks for the Static-99/R and Static-2002/R Sexual Offender Risk Tools

R. Karl Hanson, Caleb D. Lloyd, and Leslie Helmus
Corrections Research, Public Safety Canada, Ottawa, Canada

David Thornton
Sand Ridge Secure Treatment Centre, Wisconsin Department of Health Services, Wisconsin, USA

The aim of this article was to advance risk communication by examining percentile ranks as a non-arbitrary metric for quantifying risk. Although percentile ranks have a simple meaning, their calculation is complicated by ties (i.e., more than one offender having the same score). The strengths and weaknesses of percentile ranks are discussed, as are the options for calculating and presenting them in applied risk communication. As a demonstration, percentile ranks for Canadian sexual offenders were computed for the most popular sexual offender risk assessment tools (Static-99, Static-99R, Static-2002 and Static-2002R). The distribution of Static-99 scores was highly stable in international comparisons of sexual offenders from Canada (1990 to 2005; $n = 2{,}011$), Sweden (1993 to 1997; $n = 1{,}278$) and California (2008 to 2010; $n = 37{,}600$). The major limitation of percentile ranks is that they measure the "unusualness" of scores in a particular reference group, and may not correspond to other indicators of relative or absolute risk. Consequently, we recommend that evaluators presenting percentile ranks should consistently provide recidivism base rate information so that decision makers do not confuse the rarity of a score with estimates of absolute recidivism risk.

Keywords: risk assessment, percentiles, relative risk, sexual offenders, Static-99

Many decisions concerning the treatment and management of sexual offenders are based on their perceived likelihood of re-offending. Decision makers frequently receive risk information in the form of nominal risk categories, e.g., "low," "moderate," or "high." Both decision makers and evaluators prefer nominal categories over raw numbers (Grann & Pallvik, 2002; Heilbrun et al., 2004; Heilbrun, O'Neil, Stronhman, Bowman, & Philipson, 2000). One problem with nominal risk categories, however, is that professionals infer substantially different meanings (e.g., recidivism rates) from the same category label (Hilton, Carter, Harris & Sharpe, 2008; Monahan & Silver, 2003). With sexual offenders in particular, there is a tendency to overestimate the expected recidivism rates, especially for offenders described as "high risk" (Levenson, Brannon, Fortney, & Baker, 2007).

Nominal risk categories communicate more than expected recidivism rates. The meaning of words is based on their accepted use in a language – a linguistic structure that Wittgenstein (1976) referred to as a *language game*. Within delimited professional communities, labels such as "low risk" and "high risk" tell decision makers what to do with particular offenders. Hart and Boer (2010, p. 274), for example, explicitly defined the nominal categories of the SVR-20 (Boer, Hart, Kropp, & Webster, 1997) as signifying judgments about the degree of supervision and intervention required.

But what if the communication is between individuals participating in different language games? Would a sexual offender who is "low risk" to a parole officer also be "low risk" to a previous victim? Should an offender with a 10%

The views expressed are those of the authors and not necessarily those of Public Safety Canada or the Wisconsin Department of Health Services.

We would like to thank the following researchers for sharing their data with us: Jacques Bigras, Sasha Boer, Andy Haag, Niklas Långström, Janet Neeley, and Jean Proulx.

Address correspondence to R. Karl Hanson, Corrections Research, Public Safety Canada, 340 Laurier Avenue West, Ottawa, Ontario, Canada K1A 0P8. E-mail: karl.hanson@ps.gc.ca

chance of sexual recidivism be called "low risk" or "high risk"? It depends, of course, on what is *meant* by these labels.

We believe that violence risk communication can be improved by linking nominal categories to non-arbitrary, evidenced-based criteria. There are at least three plausible quantitative anchors for risk communication: (1) absolute recidivism rates (e.g., 10% chance of sexual reconviction within 5 years); (2) risk ratios (e.g., his risk is 2.5 times higher than that of the typical sexual offender); and (3) percentile ranks (e.g., top 15% in terms of risk for sexual recidivism). Each metric has its own strengths and weaknesses. Absolute recidivism rates are perhaps the most relevant for decision makers, but they are difficult to empirically estimate. Absolute recidivism rates vary with follow-up time, outcome criteria, year of release (Minnesota Department of Corrections, 2007) and by other factors that are not fully understood (Hanson, Helmus, & Thornton, 2010; Helmus, Hanson, Thornton, Babchishin, & Harris, in press). Risk ratios are another useful metric and are widely used in the communication of medical and health related risk (e.g., you can cut your risk of diabetes in half if you exercise regularly). There are, however, no commonly accepted conventions for reporting risk ratios in the context of offender risk assessment, and risk ratios are difficult to interpret in the absence of base rate information (Babchishin & Hanson, 2009). The current article focuses on what is perhaps the simplest metric: percentile ranks.

Percentile Ranks

In the context of risk assessment, percentile ranks indicate the proportion of offenders who score the same or worse on a risk assessment measure. For example, Mr. Jones is in the top 24% in terms of risk to reoffend. More generally, percentiles provide information concerning how common a particular pattern of results is in comparison to a reference population (Crawford & Garthwaite, 2009). Percentile ranks are ideal for describing characteristics for which there are established norms, and for which extreme values (high or low) are clinically significant.

The most commonly used metrics for describing test results in applied psychology (T-scores, IQ, STEN, percentiles; see Ley, 1972) are based on comparing the individual's results to the distribution of scores for various reference groups and normative samples. In violence risk communication, however, evaluators rarely provide a numeric estimate of the offender's standing relative to any clearly defined reference group. Although test developers present data that would allow evaluators to calculate percentiles, this information is often incidental and rarely emphasized in the recommendations for risk communication. For example, percentiles for the VRAG and SORAG are presented in Appendix C of Quinsey, Harris, Rice, and Cormier (2006), but the authors do not describe (1) how they were calculated (see below) or (2) the reference group that they are intended to represent. Hilton, Harris, and Rice (2010) are an exception to the general trend; they provide clearly described percentiles for the ODARA and DVRAG, as well as explicit recommendations for how percentiles could be used in risk communication.

The use of percentiles is not restricted to mechanical, actuarial risk tools. The manual of the Spousal Assault Risk Assessment Guide (SARA; Kropp et al., 1999), a structured professional judgement (SPJ) measure, includes percentile distributions for total scores and items. Users of the SARA are encouraged to consider these distributions when assigning their overall risk categorization based on professional judgement (Kropp & Gibas, 2010). To the extent that the meaning of the SPJ risk categories are intended to be consistent across settings, then the distribution of overall risk judgements provides valuable information for both evaluators and decision makers. To date, however, developers of SPJ measures have not emphasized normative data and certain authors have even actively opposed referring to normative data in risk reports (Webster, 2010).

There are several advantages to using percentile ranks that justify raising their prominence in violence risk communication. First, they are widely applicable, and can be used with any method that ranks offenders in terms of risk (e.g., structured or unstructured professional judgement, actuarial risk scores). For example, Barbaree, Langton, and Peacock (2006) transformed risk scores into percentiles when they wanted to compare the consistency of results across measures. A second advantage is that they have a commonsense meaning, and are easily understood by anybody with experience in ability and achievement rankings. A third advantage is that they are relatively easy to calculate with commonly available data, which means that they can be produced and updated without major investments of research resources.

Importantly, percentile ranks provide sufficient information for many decisions in the criminal justice system. For example, percentile ranks can meaningfully inform decisions concerning the treatment needs of offenders, with the higher risk offenders receiving more treatment than the lower risk offenders (the Risk principle; see Andrews & Bonta, 2010). Similarly, police could use percentiles to decide which offenders on large registries should be prioritized for follow-up and surveillance. Community supervision officers could use percentiles to decide which offenders deserve increased (or decreased) rates of reporting. Such decisions do not require accurate estimates of absolute recidivism rates; instead, they can be based on how the offender's risk compares to that of other offenders.

Although percentile ranks provide useful information for risk communication, they have serious limitations as a metric for quantifying risk. Risk assessment measures are inherently criterion referenced, not norm referenced. Consequently, the fundamental information communicated by percentile ranks (i.e., the rarity or unusualness of scores) may be poorly linked to certain applied decision (e.g., is he too risky to release?). As well, it is easy for the casual reader to falsely assume a direct correspondence between percentile ranks and the

underlying dimension of risk measured by particular tools. Given a large number of offenders with the same score, a one-unit change to an adjacent category could result in a substantial change in percentile rank, with only a minimal change in absolute or relative risk. Despite these limitations, we believe that percentile ranks provide useful information for both research and applied decision making.

Calculating Percentile Ranks

The basic idea of percentile ranks is simple; however, their calculation is complicated by ties (i.e., more than one offender with the same score). Given ties, several different definitions of percentile ranks are plausible (Crawford, Garthwaite, & Slick, 2009; Joint Committee, 1999, p. 179). These definitions include the percentage with (1) lower scores, (2) higher scores, (3) equal or lower scores, and (4) equal or higher scores. It is also possible to calculate percentiles based on a mid-point average across the range defined by tied scores. For example, consider 100 scores distributed as follows: 30 offenders were rated "low," 50 offenders were rated "moderate," and 20 offenders were rated as "high." Offenders with a moderate score could be described with the percentile rank of 30 (percent with a lower score), 80 (percent with the same or lower score), or 55 (mid-point average = 30 + 50/2). Similarly, it would be possible to describe the percentile rank of the moderate group as a range below (30 to 80) or as a range above (20 to 70).

Each of these definitions can be useful for specific decisions. If, for example, the decision requires selecting the 20% highest risk cases, then a definition based on "the same or riskier" would be closely aligned with the task at hand. If the decision is not specified and evaluators want to report a single value, some authorities have expressed a strong preference for reporting the mid-point average (Crawford et al., 2009; Ley, 1972). The Joint Committee (1999), however, does not specify a single, preferred definition. All authorities agree that whatever percentiles are used should be described in enough depth that readers understand what they represent.

One percentile rank that deserves special consideration is the median – the score that divides the population into the bottom 50% and the top 50%. The median is a useful reference point when describing individual results, allowing specific scores to be described as "better" or "worse" than those of the "typical" sexual offender. For many risk scores, median values are preferable to means because the data is positively skewed (many low values and relatively few, extreme high values). With grouped frequency data (i.e., "ties"), two procedures can be used to calculate the median value. The simplest approach is to identify the observed value capturing the range within which the median can be found (Anderson & Sclove, 1978; Grissom & Kim, 2005). For example, if 34% of the sample had a score of 3 or less and 56% of the sample had a score of 4 or less, the median value would be "4."

The second approach estimates an exact median by assuming that the observed, discrete values are markers for a latent continuous distribution. Specifically, a fraction of the interval containing the median value is added to the lower limit of that interval to bring the cumulative total to 50% (Ferguson, 1976; Hayes, 1981). In the above example, the exact median would be 4.23 (3.5 + [50–34]/ [56–34] × [1]). Readers will note that the exact median value estimated by this approach does not correspond to any observable value.

Nominal Categories for Percentile Ranks

Given our predilection to think in terms of words rather than numbers, it is worth considering the labels that could be assigned to groups of percentile ranks. Ideally, the labels should be directly linked to the decisions at hand. If, for example, community supervision officers had the resources to conduct home visits for only 30% of their cases, the top 30% could be labelled "home visit required" and the bottom 70% could be labelled "routine; no home visit required."

Within psychology, the range of scores considered to be normal is often defined in terms of the distance from the mean of primarily normal (symptom-free) individuals. The precise distance varies across measures and across disciplines, but is typically either 1 standard deviation or 2 standard deviations (Binder, Iverson, & Brooks, 2009; Roussos et al., 2001; Taylor & Heaton, 2001). When using 1 standard deviation as the criteria, scores in the bottom 16% ($Z < -1$) would be labelled "below average", scores one standard deviation above the mean would be "above average" ($\sim$ top 16%; $Z > 1$), and scores in the middle 68% would be "average" (i.e., all ranks from 16 to 84; $-1 < Z < 1$). Two SDs would identify a much smaller group of unusual cases: the 2.3 and 97.7 percentiles.

Within medicine, the most common threshold of abnormality is $\pm$ 2 standard deviations (Lang & Secic, 2006). For example, obesity in children is typically defined as a body mass index (BMI) at two standard deviations above the mean (Reilly, 2010). However, medical practice guidelines state, quite sensibly, that the most clinically useful thresholds are those that are linked to meaningful outcomes (e.g., prognosis, response to treatment), rather than to any particular percentile (Lang & Secic, 2006).

In the above examples, the individuals in the reference group were expected to be predominantly non-problematic. Forensic assessment, however, requires differentiating between individuals who may all have significant life problems. For instance, to be scored with the Static-2002 risk tool, it is necessary to already be sentenced for a sexual offence (Phenix, Doren, Helmus, Hanson, & Thornton, 2009). Given that this happens to about 2% of adult males (Marshall, 1997), even the lowest possible Static-2002 score could still indicate that the individual is more than 2 standard deviations above the population mean in terms of risk for sexual recidivism. Consequently, it is difficult to interpret percentile ranks on offender risk scores as indicators of normality or

abnormality. Instead, they are more naturally interpreted as indicators of relative risk within correctional or forensic populations.

For offender risk communication, it seems reasonable to use ± 1 standard deviation to differentiate unusually high or unusually low scores from scores that are within the typical range. It is less clear, however, what these groups should be called. Labelling offenders with unusual scores as "low risk" or "high risk" carries meanings that may or may not be justified, and cannot be known without additional information concerning the context of risk communication and the psychometric properties of the risk scale in question. Currently, our preferred labels for these generic categories of percentile ranks are "unusually low scores," "scores within the typical range," and "unusually high scores."

Although helpful as a heuristic, clustering scores into broad categories inevitably results in a loss of precision. For certain decisions, meaningful differences in recidivism risk would be expected among the majority of offenders placed in the middle (± 1 SD) category. To the extent that greater precision is relevant to the decision at hand, evaluators may want to create more than 3 groups (e.g., quintiles, deciles), or completely avoid categories and, instead, report the percentiles (in numeric form).

Defining the Reference Group

Given that percentiles communicate how common scores are within a normative group, it is important to specify the reference group. Features that are rare among general offenders may be common among men convicted of sexual offences, and ubiquitous among men referred for sexual offender treatment. Ideally, the percentiles should represent the *population* for whom the measure is intended. More typically, the best that evaluators can hope for is a reasonably unbiased estimate of a clinically relevant reference group.

In the context of sexual offender risk assessment, a useful reference group would be the complete sample of convicted sexual offenders in a particular jurisdiction. This reference group would be expected to change slowly, and provide a reliable anchor for decisions concerning resource management. Risk assessment research has rarely used unbiased samples of this population. Most studies of actuarial risk scales have used relatively small samples ($n < 1{,}000$) selected from specific settings. Often the selection criteria are not fully known, and can change rapidly based on management decisions (e.g., a different contractor takes responsibility for a treatment program). Consequently, the frequency distributions for complete, unselected samples are likely to be more stable than the frequency distributions found for any particular setting. It is quite likely that relatively complete, unbiased samples will increasingly become available as risk assessment tools are broadly implemented, and the scores are retained in automated databases (see California data discussed below).

OVERVIEW OF CURRENT STUDY

In this paper, we calculate and compare the percentile ranks for Static-99, Static-99R, Static-2002, and Static-2002R. Static-99 (Hanson & Thornton, 2000) is the most commonly used actuarial scale to predict sexual recidivism (Archer et al., 2006; Jackson & Hess, 2007; McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). It is used for treatment planning (McGrath et al., 2010; Jackson & Hess, 2007), community supervision (Interstate Commission for Adult Offender Supervision, 2007), and civil commitment evaluations (Jackson & Hess, 2007). Static-99 is based on easily scored criminal history information, and has demonstrated predictive accuracy (discrimination between sexual recidivists and non-recidivists) as good as for other available measures (Hanson & Morton-Bourgon, 2009).

Static-99R has the same items as Static-99, with the exception that increased weight is accorded to advanced age as a protective factor (Helmus, Thornton, Hanson, & Babchishin, 2011). Static-2002 (Hanson & Thornton, 2003) is a separate risk scale in the same tradition as Static-99, which has also been revised with new age weights, creating Static-2002R (Helmus et al., 2012). Static-2002 was created with the goal of improving the conceptual clarity of the items and increasing its validity in adversarial assessment settings (e.g., minimizing the contribution of offender self-report). Unlike Static-99, the Static-2002/R items are organized into conceptually meaningful subscales (e.g., persistence of sexual offending, deviant sexual interests, general criminality).

The percentiles calculated in this paper considered all Canadian adults convicted of a sexual offence as the reference category. An unbiased sample of all Canadian sexual offenders was not available; however, we were able to identify four samples of sexual offenders released between 1990 and 2005 from the three major divisions of the Canadian criminal justice system: (1) community, (2) provincial prison (sentences of less than 2 years that are administered by the provinces), and (3) federal prison (sentences of 2 years or more that are administered federally by the Correctional Service of Canada). None of these samples had been explicitly pre-selected on risk relevant criteria (e.g., need for treatment, detention until warrant expiry). We then used standard survey sampling statistics (Kalton, 1983) to estimate a representative normative (Canadian) sample from these multiple independent samples.

To check the validity and generalizability of our results, we compared the distribution of Static-99 scores estimated for Canadian sexual offenders ($n = 2{,}011$) against two large samples from Sweden and California. The first comparison was all sexual offenders who were released from a prison sentence in Sweden between 1993 and 1997 ($n = 1{,}278$; Sjöstedt & Långström, 2001). The second comparison group was all registered sexual offenders serving sentences in California during the years 2008, 2009, or 2010 ($n = 37{,}600$;

State Authorized Risk Assessment Tool for Sex Offenders Review Committee, 2011). Although the sampling frames were slightly different, these comparisons, nevertheless, provide guidance as to the stability of percentile ranks across jurisdictions.

METHOD

Measures

Static-99 *(Hanson & Thornton, 2000)*

Static-99 is a 10-item actuarial measure that assesses recidivism risk of adult male sexual offenders based on frequently available criminal history information. Offenders can be placed in one of four risk categories based on their total score (ranging from 0–12): low (0, 1), moderate-low (2, 3), moderate-high (4, 5), and high (6+). Previous research has demonstrated high interrater reliability (for examples, see Bengtson & Långström, 2007; de Vogel, de Ruiter, van Beek, & Mead, 2004; G. T. Harris et al., 2003; Knight & Thornton, 2007; Langton, 2003) and moderate to high predictive accuracy (Hanson & Morton-Bourgon, 2009).

Static-99R *(Helmus et al., 2012)*

The items are identical to Static-99 (Hanson & Thornton, 2000), with the exception of updated age weights. In Static-99, a single point is according to offenders who are less than 25 at time of release; in Static-99R, the age weights are as follows: age 18–34.9 = 1, 35 to 39.9 = 0, 40 to 59.9 = –1, >60 = –3. The possible range is –3 to 12. Static-99R demonstrates moderate to high predictive accuracy (Babchishin, Hanson, & Helmus, 2011).

Static-2002 *(Hanson & Thornton, 2003; Hanson et al., 2010)*

Static-2002 is a 14-item[1] measure that assesses recidivism risk of adult male sexual offenders based on frequently available criminal history information. Offenders can be placed in one of five risk categories based on their total score (ranging from 0–14): low (0–2), low-moderate (3, 4), moderate (5, 6), moderate-high (7, 8), and high (9+; see Phenix et al., 2009). Previous research has demonstrated high interrater reliability (e.g., Haag, 2005; Langton, Barbaree, Hansen, Harkins, & Peacock, 2007) and moderate to high predictive accuracy (Hanson et al., 2010; Hanson & Morton-Bourgon, 2009).

[1]Static-2002 originally had 13 items. To increase clarity in the coding manual (Phenix et al., 2009), one item was divided into two separate items. However, this does not change the total score (the summed score of the two separate items in the new coding rules is identical to the score for the old item).

Static-2002R *(Helmus et al., 2012)*

The items are identical to Static-2002, with the exception of updated age weights. In Static-2002, age is coded as follows: 18 to 24.9 = 3, 25 to 34.9 = 2, 35 to 49.9 = 1, and > 50 = 0; for Static-2002R: 18–34.9 = 2, 35 to 39.9 = 1, 40 to 59.9 = 0, > 60 = –2. The possible range is –2 to 13. Static-2002R demonstrates moderate to high predictive accuracy (Babchishin et al., 2011).

Data Sources

Four datasets were used to estimate a representative sample of Canadian sex offenders and two datasets were used for international comparisons. Individual-level data were available for the Canadian and Swedish samples. For individual-level data, cases were deleted if more than one Static-2002/R item was missing, if any item on Static-99/R except Item 2 (single) was missing, or if data cleaning revealed errors in coding that could not be resolved after contacting the original authors. In total, 58 cases from the Canadian samples were deleted for coding inconsistencies or missing items (resulting in a reduced sample of 2,011 out of 2,069), and less than 5 cases from Sweden (leaving n = 1,278). Individual-level data were not available for California, and the analyses were based solely on the aggregated distribution of total scores.

Table 1 displays characteristics of the four samples of Canadian sex offenders (n = 2,011). Three of the samples were exclusively federal offenders whereas the fourth sample included a mix of non-custodial, provincial, and federal offenders. All offenders were adult males who committed a sexually motivated offence with an identifiable victim, either a child or non-consenting adult; sexual offences included contact and non-contact offences, such as exhibitionism and voyeurism (Category "A" offenses in the Static-99 coding rules, A. J. R. Harris, Phenix, Hanson, & Thornton, 2003). Offenders who committed a sexually motivated offense but were convicted of a non-sexual violent offense (e.g., manslaughter or a sexual assault case that was plea-bargained to assault) were also included with the exception of the 1995 Warrant Expiry Date (WED) sample and (possibly) the B.C. sample. Ethnicity was rarely provided, but it can be assumed that most offenders were Caucasian based on Canadian demographics and because approximately two-thirds of federally incarcerated offenders are Caucasian (Public Safety Canada, 2010).

Canadian Samples

1) Dynamic Supervision Project (Hanson, Harris, Scott, & Helmus, 2007)

This prospective study followed sex offenders on community supervision between 2001–2005 in all Canadian provinces and territories, and two U.S. states. For the current study, only Canadian offenders were considered.

TABLE 1
Characteristics of Canadian Samples

Sample	*N*	Age *M* (*SD*)	Offender Type: % Rapists/ % Child Molesters	Static-99 *M* (*SD*)	Static-99R *M* (*SD*)	Static-2002 *M* (*SD*)	Static-2002R *M* (*SD*)
Dynamic Supervision Project	595	42 (14)	36/54	2.6 (1.9)	2.1 (2.3)	3.8 (2.2)	3.2 (2.4)
Federal: B.C.	296	41 (12)	40/55	3.2 (2.3)	2.8 (2.8)	4.5 (2.5)	3.9 (2.7)
Federal: 1995 WED	663	41 (12)	46/52	2.8 (2.0)	2.5 (2.6)	4.6 (2.4)	4.1 (2.6)
Federal: Quebec	457	43 (12)	38/46	2.7 (2.0)	2.1 (2.4)	4.1 (2.3)	3.5 (2.5)
Total	2,011	42 (13)	40/52	2.8 (2.0)	2.3 (2.5)	4.2 (2.3)	3.7 (2.5)

Note. Age refers to age at release.

Participating probation officers ($n = 137$) were requested to submit demographic, offense history, and risk assessment information (Static-99, STABLE-2007, ACUTE-2007) on sex offenders consecutively entering their caseload. File review indicated that the cases were not always consecutive; however, the sample can be considered representative of the diverse group of sex offenders on community supervision.

Static-99 scores were coded prospectively by the probation officers. Static-2002 scores were coded by the third author and a graduate student based on information from Static-99 scores and Canadian Police Information Centre (CPIC) records maintained by the Royal Canadian Mounted Police (RCMP). Static-99R and Static-2002R were created by retrospectively reweighting the age variables of these measures.

Of the 595 offenders with the necessary data for the current analyses, 38 were supervised following a federal sentence ($n = 38$, 6.4%), 254 following a provincial sentence (42.7%), and 303 received solely a non-custodial sentence (e.g., probation, conditional sentence order, or in rare cases, a peace bond; 50.9%). Twenty-four offenders (4.0%) had a non-sexual violent index offense.

Interrater reliability for Static-99 was examined through file review of 88 cases coded by probation officers participating in the DSP project ($ICC = .91$). An exceptionally high interrater reliability for Static-2002 coding ($ICC = .98$, $n = 25$ cases) was observed. Coding was based upon probation officers' obtained Static-99 scores and conviction information rather than interpretation of victim information or offense circumstances. Consequently, reliability for Static-2002 scores in this study should not be considered representative or typical.

2) Canadian federal offenders: B.C. (Boer, 2003)

Archival data from the Offender Management System (OMS) maintained by Correctional Service Canada (CSC) were used to identify all federal male offenders serving a sentence for a sexual offense in British Columbia whose Warrant Expiry Date (WED; the end of their sentence) was between January 1990 and May 1994. Many offenders are granted conditional release before their WED; thus, offenders in this sample were released as early as 1986 ($n = 296$). Interrater reliability was unavailable for this sample.

3) Canadian federal offenders: 1995 WED (Haag, 2005)

OMS records were used to identify all federal sex offenders with a WED in 1995. Offenders were released as early as 1987 ($n = 663$). Interrater reliability for Static-99 and Static-2002 scores was high ($r = .92$ and .84, respectively; $n = 66$ cases) when assessed by the lead researcher (Haag) and another psychologist.

4) Canadian federal offenders: Quebec (Bigras, 2007)

This study included 94% of all sex offenders receiving a federal sentence in Quebec between 1995–2000 (6% refused participation in the research or were unable to provide consent). Static-99 and Static-2002 scores were coded from file data and offender interviews ($n = 457$). Interrater reliability was unavailable for this sample.

Samples for International Comparisons

5) Sweden (Sjöstedt & Långström, 2001)

This study included all adult men who served a prison sentence for a sex offence and were released between 1993–1997 ($n = 1,278$). In Sweden, approximately 70% of men convicted of sex offenses are sentenced to prison. Contact sex offenders are more likely to receive prison sentences than non-contact sex offenders. In this sample, 43% of offenders were rapists and 45% were child molesters. Offenders deported upon release were excluded. Risk assessment scores were coded based on file review and coding was blind to recidivism outcome. Interrater reliability was good, with an average item kappa of .90 (no item had a kappa below .83; $n = 20$).

6) California 2008–2010 (State Authorized Risk Assessment Tool for Sex Offenders Review Committee, 2011)

Static-99 data were available for all registered sexual offenders serving a sentence in California during the years 2008, 2009, and 2010 (total $N = 37{,}600$). These offenders either (1) were under the authority of the California Department of Corrections and Rehabilitation (prison or parole), (2b) received a probation sentence for a sexual offense, or (3) started a probation sentence for a non-sexual offense and had a previous conviction for a registerable sex offense. The data were collected for administrative purposes; based on a law enacted in 2006, all sexual offenders serving a sentence in California beginning in 2008 required a Static-99 (now Static-99R) score. Interrater reliability was unavailable for this sample.

Population Weights for Canadian Samples

The Canadian estimates were based on treating sentences as strata. In Canada, there are three major categories of sentences: community sentences (e.g., probation, conditional sentence, fines), prison sentences of less than two years (which are administered by the provinces), and prison sentences of two years or more (administered by the federal prison service: CSC). The distribution of these three sentence types among adult sexual offenders in Canada was drawn from the Adult Court Criminal Survey conducted by the Canadian Centre for Justice Statistics (CCJS, 2008).[2] The sentence recorded for each offender was for his or her most serious charge from adult criminal courts (for more information regarding how cases were coded, see Marth, 2008). CCJS defines sexual offenses as convictions for sexual assault (all levels), sexual interference, invitation to sexual touching, sexual exploitation, incest, anal intercourse, and bestiality (Kong, Johnson, Beattie, & Cardillo, 2003). This definition excludes Category "B" offenses (e.g., possession of child pornography or prostitution offenses; see A. J. R. Harris et al., 2003) and a small number of less common category "A" sex offenses such as trespassing at night (voyeurism) or indecent act (note that the Category "A" offenses excluded in the CCJS data were not excluded in the present study samples). CCJS statistics include female sex offenders. It is expected that approximately 3% of persons in the CCJS data were female based on known police-reported rates (Kong et al., 2003; note that only male offenders can be scored on Static-99 and Static-2002).

Data were averaged over five reporting periods (2001/2002 to 2005/2006). A small percentage of offenders (2.0%) was removed from the prison samples because the length of incarceration was unknown (e.g., "time served"); consequently, a corresponding proportion for each reporting period was removed from the community samples, resulting in a total sample of 12,819. In these years, 51.3% of convicted sexual offenders received a non-custodial sentence, 37.4% received sentences of less than 2 years (provincial), and 11.3% received sentences of more than 2 years (federal).

Plan of Analysis

Consistent with the recommendations of Crawford et al. (2009) and Ley (1972), percentile ranks were estimated as mid-point averages:

$$Percentile = \left(\frac{m + .5k}{N}\right) 100$$

Where m is the number of members of the normative sample scoring below a given score, k is the number obtaining the given score, and N is the overall size of the normative sample. We also calculated percentiles at the natural breaks between each score.

Median values were also calculated in two ways: (1) as the discrete score within which the 50th percentile is found, and (2) as an estimate of the exact value assuming a latent, continuous distribution (e.g., Ferguson, 1976, Equation 3.8):

$$median = L + \frac{{}^{N}/_{2} - m}{k} S$$

In this equation, L is the lower limit (on a continuous scale) of the interval containing the 50th percentile, and S is the size of that interval. For all STATIC scales, the scores are in increments of $S = 1$. For example, the lower limit of the interval for a score of "2" would be $L = 1.5$, and the lower limit of the interval for a score of "3" would be $L = 2.5$.

Standard formulae were used for estimating the average percentiles from stratified samples (Kalton, 1983). Specifically, the estimate of the population mean percentile was the sum of the mean percentiles calculated for each stratum after weighting the means by the relative distribution of their strata within the full population:

$$mean = \sum W_h \bar{y}_h$$

where h indicates the level of strata. In our study, the W_h were 51.3% (non-custodial), 37.4% (provincial), and 11.3% (federal).

To calculate confidence intervals for the percentile ranks, we used the adaptation of classical (frequentist) method described by Crawford et al. (2009). As in the Clopper-Pearson approach, Crawford et al.'s method treats percentile ranks as the probability parameter in a binomial distribution (the other parameter being sample size). Using these two parameters, a search procedure finds the tails of the distribution that correspond to a probability of .025. Unlike true binominal distributions, however, the precision of percentile ranks are reduced when multiple individuals share the same score. To account for ties, Crawford et al.'s (2009) method calculates

[2]We would like to thank Kelly Morton-Bourgon for locating these data.

the probability (.025) based on an average of all possible ranks within the tied category.

Crawford's *percentile_norms_int_est.exe* software was used to calculate the confidence intervals. This program was designed for a single normative sample. In order to adapt it to our purposes, we used the distribution of scores estimated from re-weighting the stratified sample; for the total sample size, however, we used the actual, observed sample size ($n = 2{,}011$). This is a conservative approach because the variance of estimates from stratified samples cannot be greater than the variance of estimates from simple (non-stratified) random samples (Kalton, 1983, p. 21).

RESULTS

The percentile ranks for Canadian sexual offenders for Static-99, Static-99R, Static-2002, and Static-2002R are presented in Tables 2, 3, 4, and 5, respectively. The raw (unweighted) distributions used to calculate these percentile ranks are provided in the Appendix. For each STATIC score, Tables 2–5 provide the proportion with a lower score, the proportion with the same score, and the proportion with a higher score (these proportion always sum to 100). For example, 10.6% of Canadian sexual offenders had a Static-99 score of 4, 73.4% had a score of 3 or less, and 16.0% had a score of 5 or more. The tables also provide the percentile rank defined as a mid-point average. Readers will note that the 95% confidence intervals for the percentile ranks estimated as mid-point averages closely approximated the upper and lower limits of the observed frequency distributions. For example, 73.4% had a score lower than 4, and 84.0% had the same or lower score (observed range of 73.4 to 84.0). The mid-point average was 78.7, with a 95% confidence interval of 73.1% to 84.2%, which was virtually identical to the observed range. Given the large sample size used in the current study (> 2,000), sampling error for the raw proportions would be expected to be small. Instead, the uncertainty in the estimates of the

TABLE 2
Estimated Percentiles for Static-99 Scores for Canadian Sexual Offenders

	Observed Percentages			Percentile Rank defined as mid-point average	
Static-99 Score	Below	Same	Higher	Percentile	95% CI
0	0	12.6	87.4	6.3	0.3–12.6
1	12.6	18.7	68.7	21.9	12.8–31.3
2	31.3	21.9	46.8	42.3	31.5–53.1
3	53.2	20.2	26.6	63.3	53.3–73.3
4	73.4	10.6	16.0	78.7	73.1–84.2
5	84.0	7.0	9.0	87.5	83.6–91.3
6	91.0	5.4	3.6	93.7	90.7–96.5
7	96.4	1.9	1.7	97.3	96.0–98.6
8	98.3	1.2	0.5	98.9	98.0–99.6
9	99.5	0.48	0.02	99.7	99.3–100.0
10+	99.98	0.02	0	99.99	99.8–100.0

Note. Distribution based on an adjusted, re-weighted average of 4 samples (total sample size of $n = 2{,}011$).

TABLE 3
Estimated Percentiles for Static-99R Scores for Canadian Sexual Offenders

	Observed Percentages			Percentile Rank defined as mid-point average	
Static-99R Score	Below	Same	Higher	Percentile	95% CI
–3	0	2.7	97.3	1.3	0–2.9
–2	2.7	3.0	94.3	4.2	2.4–6.1
–1	5.7	7.9	86.4	9.7	5.7–13.9
0	13.6	10.3	76.1	18.7	13.4–24.1
1	23.9	15.7	60.4	31.7	23.8–39.7
2	39.6	17.5	42.9	48.3	39.5–57.1
3	57.1	17.2	25.7	65.7	57.0–74.3
4	74.3	10.7	15.0	79.6	74.0–85.1
5	85.0	7.4	7.6	88.7	84.6–92.5
6	92.4	3.6	4.0	94.2	91.9–96.2
7	96.0	2.5	1.5	97.2	95.6–98.6
8	98.5	1.2	0.3	99.1	98.2–99.8
9	99.7	0.28	0.02	99.9	99.5–100.0
10+	99.98	0.02	0	99.99	99.8–100.0

Note. Distribution based on an adjusted, re-weighted average of 4 samples (total sample size of $n = 2{,}011$).

TABLE 4
Estimated Percentiles for Static-2002 Scores for Canadian Sexual Offenders

	Observed Percentages			Percentile Rank defined as mid-point average	
Static-2002 Score	Below	Same	Higher	Percentile	95% CI
0	0	4.6	95.4	2.3	0.1–4.8
1	4.6	10.9	84.5	10.0	4.7–15.6
2	15.5	15.0	69.5	23.0	15.5–30.6
3	30.5	17.2	52.3	39.1	30.5–47.8
4	47.7	17.6	34.7	56.5	47.6–65.3
5	65.3	14.7	20.0	72.6	65.1–79.9
6	80.0	8.9	11.1	84.4	79.6–89.0
7	88.9	4.5	6.6	91.2	88.4–93.7
8	93.4	3.4	3.2	95.1	93.0–96.9
9	96.8	1.9	1.3	97.7	96.3–98.8
10	98.7	1.2	0.1	99.3	98.4–99.9
11	99.9	0.08	0.02	99.9	99.7–100.0
12+	99.98	0.02	0	99.99	99.8–100.0

Note. Distribution based on an adjusted, re-weighted average of 4 samples (total sample size of $n = 2{,}011$).

TABLE 5
Estimated Percentiles for Static-2002R Scores for Canadian Sexual Offenders

Static-2002R Score	Observed Percentages			Percentile Rank defined as mid-point average	
	Below	Same	Higher	Percentile	95% CI
–2	0	2.8	97.2	1.4	0–3.0
–1	2.8	2.9	94.3	4.2	2.6–6.1
0	5.7	6.7	87.6	9.0	5.5–12.8
1	12.4	9.7	77.9	17.3	12.3–22.5
2	22.1	16.0	61.9	30.1	22.2–38.3
3	38.1	17.9	44.0	47.1	38.1–56.1
4	56.0	15.3	28.7	63.7	55.9–71.4
5	71.3	13.5	15.2	78.0	71.1–84.7
6	84.8	7.1	8.1	88.3	84.3–92.1
7	91.9	2.8	5.3	93.3	91.3–95.1
8	94.7	2.5	2.8	95.9	94.2–97.4
9	97.2	2.3	0.5	98.3	96.9–99.5
10	99.5	0.4	0.1	99.7	99.3–100.0
11	99.9	0.09	0.01	99.97	99.8–100.0
12+	99.99	0.01	0	99.99	99.8–100.0

Note. Distribution based on an adjusted, re-weighted average of 4 samples (total sample size of $n = 2{,}011$).

percentile ranks was primarily based on the inherent uncertainty due to ties (i.e., multiple offenders receiving identical scores). For all the STATIC measures examined in the current study, there were substantial numbers of ties, with the most populated categories including 10% to 20% of the total sample.

For both Static-99 and Static-99R, the typical (median) value was 2, and scores of 5 or more identified unusually high scores ($> +1$ *SD*). A score of 0 was unusually low (< -1 *SD*) for Static-99, whereas the unusually low scores for Static-99R were –1 or less (–1, –2, or –3). The median value for Static-2002 was 4 and, for Static-2002R, it was 3. For both Static-2002 and Static-2002R, scores of 6 or more were unusually high. For Static-2002, 0 and 1 were unusually low scores; for Static-2002R, low scores were zero or less (0, –1 or –2; see Table 6).

International Comparisons

To examine the generalizability of the Canadian percentile ranks, the distribution of Static-99 scores were compared with the distribution of scores from California between 2008–2010 and Sweden between 1993–1997. As can be seen in Figure 1, the distributions of scores were substantially similar, although not identical. Statistical tests were not conducted because with these large sample sizes any difference of practical significance would also be statistically significant. For example, the 95% confidence interval for the proportions in the California sample would be less than $\frac{1}{2}$ of one percentage point.

TABLE 6
Distribution of STATIC Scores Among Canadian Sexual Offenders ($n = 2{,}011$)

	STATIC Measure			
	Static-99	Static-99R	Static-2002	Static-2002R
Range				
Possible	0 to 12	-3 to 12	0 to 14	-2 to 13
Observed	0 to 10	-3 to 10	0 to 12	-2 to 12
Average (SD)	2.6 (1.9)	2.1 (2.3)	3.8 (2.2)	3.2 (2.4)
Median				
Integer	2	2	4	3
Continuous	2.35	2.09	3.63	3.16
Scores < -1 SD (percent)	0 (12.6)	-3, -2, -1 (13.6)	0, 1 (15.5)	-2, -1, 0 (12.4)
Typical range (percent)	1, 2, 3, 4 (71.4)	0, 1, 2, 3, 4 (71.4)	2, 3, 4, 5 (64.5)	1, 2, 3, 4, 5 (72.4)
Scores $> +1$ SD (percent)	5 or more (16.0)	5 or more (15.0)	6 or more (20.0)	6 or more (15.2)

Compared to Canada (12.6%) and California (12.5%), the Swedish sample had a greater proportion of offenders with a score of zero (17.3%). In contrast, the Canadian distribution contained more offenders with a score of 2 (21.9% versus 19.6% in Sweden and 19.7% in California) or a score of 3 (20.2% versus 16.4% in Sweden and 17.4% in California). Nevertheless, the median was the same in all three samples (integer value of 2), as were the cut-points for unusually low scores (0), scores within the typical range (1, 2, 3, 4), and unusually high scores (5 or greater; see Table 7).

TABLE 7
International Comparison of Static-99 Scores

Jurisdiction	Canada	California (US)	Sweden
Sample size	2,011	37,600	1,278
Sentences	Prison/Probation	Prison/Probation	Prison
Years	1990–2005	2008–2010	1993–1997
Sampling	Mixed/stratified	Consecutive	Consecutive
Range	0 to 10	0 to 10	0 to 10
Average (SD)	2.6 (1.9)	Not available[a]	2.4 (2.0)
Median			
Integer value	2	2	2
Continuous estimate	2.35	2.45	2.11
Scores < -1 SD (percent)	0 (12.6)	0 (12.5)	0 (17.3)
Typical range (percent)	1 to 4 (71.4)	1 to 4 (69.0)	1 to 4 (68.3)
Scores $> +1$ SD (percent)	5+ (16.0)	5+ (18.4)	5+ (14.4)

[a]The average Static-99 score was not available for California because the publicly released data collapsed scores of 6 or more into one category.

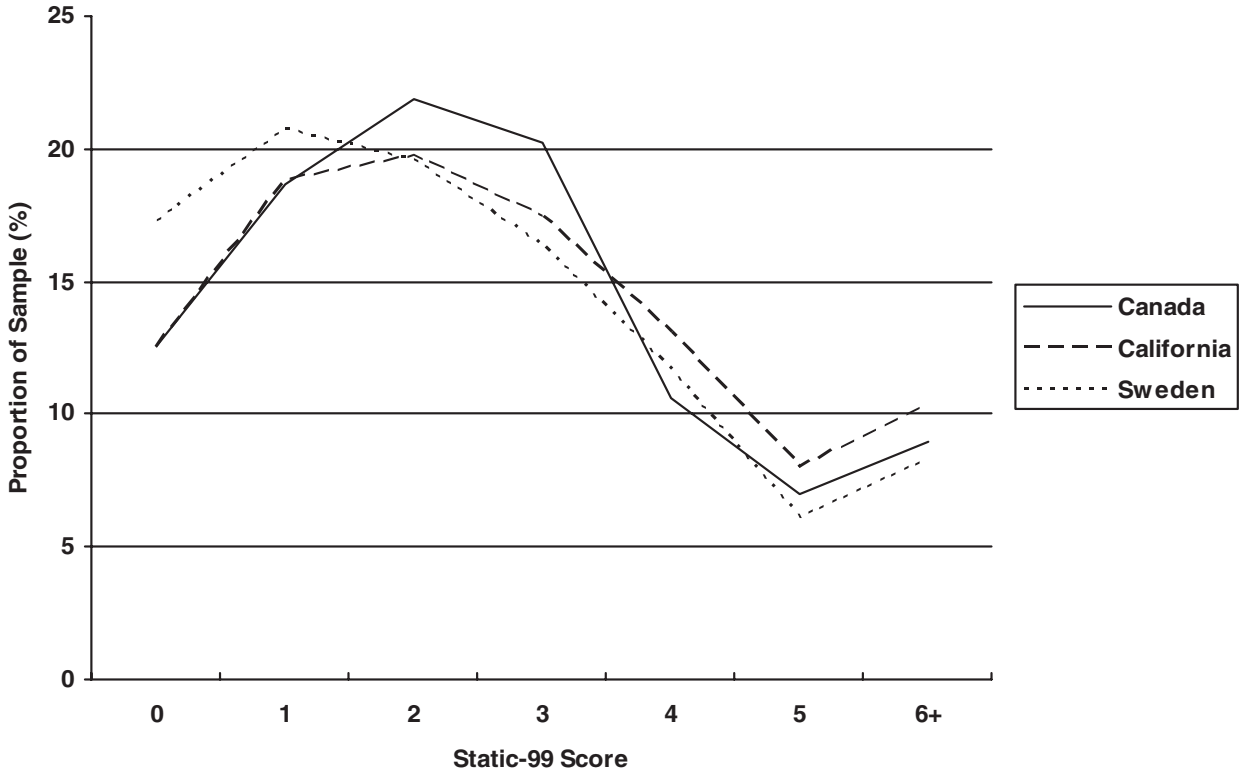

FIGURE 1 Comparison of Static-99 population percentiles for California sex offenders, Swedish prisoners, and sample estimate percentiles for Canadian sex offenders.

DISCUSSION

The aim of this paper was to advance risk communication by examining percentile ranks as a non-arbitrary metric for quantifying risk. As a demonstration, percentile ranks were computed for the most popular sexual offender risk assessment tools (Static-99, Static-99R, Static-2002, and Static-2002R). These percentile ranks should be useful to evaluators and decision makers wishing to compare offenders' scores across these different scales. For example, correctional administrators upgrading from Static-99 to Static-99R or Static-2002R could use the tables in the current study to determine cut-points that approximate the distributions for the risk categories previously used. The tables would also be useful for estimating the cost implications of changes in cut-points for routine correctional interventions. For researchers and policy-makers, the percentile tables provide an empirical reference point for statements concerning the risk presented by the "average" or "typical" sexual offenders.

Percentile ranks have both strengths and weaknesses as the basis of risk communication. In terms of strengths, percentiles are easily calculated and have an intuitive meaning to anyone experienced with skill and ability evaluations. Furthermore, they can provide a stable reference point upon which to compare results across settings and samples. We found, for example, that the distributions of Static-99 scores were substantially similar for sexual offenders in Canada (1990 to 2005), Sweden (1993 to 1997), and California (2008–2010). In each of these populations, the median value for Static-99 was identical (2) as was the typical range ($\pm$ 1 SD; 1 to 4).

The stability of percentile ranks, however, only applied to large, representative samples. When the Canadian samples were restricted to federally incarcerated sexual offenders, then the median Static-99 value was 3 (not 2), and an unusually high score was 6 (not 5; see Appendix). Although it is possible to develop and use norms based on various subsamples, communication across jurisdictions is enhanced when percentile ranks are anchored in a common population (e.g., all Canadian sexual offenders). Consequently, we recommend that evaluators privilege norms that most closely approximate unbiased samples of stable, clearly defined populations.

Despite their merits, percentile ranks have inherent limitations as a metric for quantifying the information contained in risk assessment tools. Fundamentally, percentiles quantify the unusualness of scores, and are best suited to norm-referenced tests. Risk assessment tools, in contrast, are criterion-referenced. For criterion-referenced measures, the information value of scores is determined by the association between the latent construct assessed by the measure and the outcome of interest.

For the STATIC risk scores, the fundamental quantity being assessed is relative risk. Specifically, these measures assume a cumulative stochastic model, in which a one unit increase in score is associated with an equal and consistent increase in relative risk, which can be expressed as an odds ratio or hazard ratio of approximately 1.3 (Hanson et al., 2010; Hanson & Thornton, 2003; Helmus et al., 2012). If this pattern is true, then percentile ranks are unlikely to have a linear relationship with the underlying risk measured by these scales. Given large numbers of ties, offenders with adjacent scores could have large differences in percentile ranks while having only minimal differences in STATIC risk. Consequently, research is needed to justify statements concerning the relationship between percentiles ranks and the amount of "riskiness" associated with any particular score.

The existence of ties complicates the calculation and communication of percentile ranks. When many offenders have the same score, there are several plausible definitions of percentile rank. Certain commentators recommend using only the mid-point average (Crawford et al., 2009), but this approach cannot eliminate the inherent uncertainty created by ties. Even with large sample sizes, the confidence intervals for the mid-point averages are never smaller than the range defined by the percentage lower and the percentage the same or lower. Consequently, communicating percentiles requires presenting more than a single number. Some evaluators will want to follow Crawford et al.'s (2009) recommendation and present mid-point averages with 95% confidence intervals. Other evaluators may follow the example of Hilton et al. (2010) by reporting three numbers: the percent below, the percent with the same score, and the percent with higher scores. In general, the way in which percentiles are presented should be tailored to the decisions at hand.

The following is a generic example of how percentile ranks could be presented in an applied report: "Compared to other adult male sex offenders, Mr. X's score (6 on Static-99R) places him in the 94th percentile. Taking into account that about 4% of sex offenders shared the same score as Mr. X, this percentile means that roughly 92% of offenders scored lower than Mr. X, and that 4% scored higher." For those who like narrative accounts, it would also be possible to describe Mr. X as having an unusually high Static-99R score. Statements concerning percentile ranks should be accompanied by a brief description of the evidence upon which they are based: for example, "These percentiles are from 2,011 cases from 4 samples of Canadian sex offenders, which were reweighted to approximate the distribution of all convicted sex offenders in Canada. These percentiles appear highly stable in international comparisons with large, relatively representative samples in Sweden and California."

It is also worth considering statements that cannot be supported by percentile rank information. The finding that Mr. X has an unusually low score on a particular risk measure, for example, does not necessarily mean that he is unlikely to reoffend. Statements concerning the likelihood of recidivism can only be supported by reference to absolute recidivism rates. As well, it would be incorrect to assume that a 30% difference in percentile ranks (e.g., from 60 to 90) corresponds to a 30% difference in relative or absolute risk.

We have emphasized the need for non-arbitrary metrics for applied risk communication, but there is an equally strong need for researchers to quantify what is meant by "low," "moderate," and "high" risk. One feature of effective correctional interventions is that they direct the most resources towards offenders with moderate (or higher) risk of recidivism; services provided to low-risk offenders would be expected to have no effect, and can even increase the risk for recidivism (Andrews & Bonta, 2010; Andrews & Dowden, 2006; Bonta & Andrews, 2007). Although empirical support for this risk principle has been documented in diverse samples (see Table 2.2 from Andrews & Bonta, 2010), the enlightened practitioner wishing to apply this principle has little guidance as to where to draw the line. Different risk scales have different normative categories, and the distribution of scores in specific research studies may not generalize to other settings. Consequently, it would be useful for those evaluating correctional programs to routinely report risk ranking using the same metric. With a standardized metric, it would be possible to empirically establish the amount of treatment needed (if any) for different risk levels.

Although we believe that the distributions provided in the current study are reasonable estimates of the population distributions of these STATIC measures, our study had certain limitations. First, the Canadian estimates were based on four subsamples, none of which were random or unbiased. Furthermore, the sampling procedures for the international comparisons were not identical (e.g., community-only offenders were included in the Canadian and Californian samples but excluded from the Swedish sample). Consequently, we would expect the percentile ranks to be revised as researchers collect new and better samples. In particular, the routine implementation of Static-99 in many jurisdictions has the potential of generating excellent local norms. Nevertheless, the degree of stability observed in the current study suggests that the variation in distributions across jurisdictions will be relatively minor.

In summary, it is the responsibility of the professional community to establish plausible definitions for risk communication. Given that nominal categories are unlikely to go away, we should work towards giving nominal risk categories explicit, non-arbitrary meanings. One potential empirical reference point for definitions could be percentiles based on comprehensive, representative, unbiased samples of the population of interest. Although percentile ranks may not be the ideal metric for actuarial risk tools, they are a useful starting point that can be applied to all procedures that reliably rank offenders in terms of recidivism potential. We encourage the developers of risk measures to routinely report percentile ranks, and for evaluators and decision makers to consider this information when interpreting the results of

empirically-validated risk tools. Percentile ranks, however, should not be used in isolation. It is important for evaluators and decision makers to also consider recidivism base rates when making judgments concerning the overall "riskiness" of particular offenders.

REFERENCES

Anderson, T. W., & Sclove, S. L. (1978). *An introduction to the statistical analysis of data.* Boston: Houghton Mifflin.

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

Andrews, D. A., & Dowden, C. (2006). Risk principle of case classification for reduced recidivism: A meta-analytic investigation. *International Journal of Offender Therapy and Comparative Criminology*, *50*, 88–100. doi: 10.1177/0306624X05282556

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 84–94. doi: 10.1207/s15327752jpa8701_07

Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Moving beyond the "low", "moderate", and "high" typology of risk communication. *Crime Scene*, *16*(1), 11–14.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2011). *The RRASOR, Static-99R, and Static-2002R all add incrementally to the prediction of recidivism among sex offenders* (Corrections Research User Report No. 2011–02). Ottawa, ON: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/2011–02-aipraso-eng.pdf

Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006). Different actuarial risk measures produce different risk rankings for sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 423–440. doi: 10.1007/s11194–006-9029–9

Bengtson, S., & Långström, N. (2007). Unguided clinical and actuarial assessment of re-offending risk: A direct comparison with sex offenders in Denmark. *Sexual Abuse: A Journal of Research and Treatment*, *19*, 135–153. doi: 10.1177/107906320701900205

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)*.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*, 31–46. doi:10.1093/arclin/acn001

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.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk–20: Professional guidelines for assessing risk of sexual violence*. Vancouver, British Columbia, Canada: British Columbia Institute Against Family Violence.

Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation* (Corrections Research User Report 2007–06). Ottawa, ON: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/risk_need_200706-eng.aspx

Canadian Centre for Justice Statistics. (2008). *Adult Criminal Court Survey*.

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. doi: 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. doi: 10.1080/13854040902795018

de Vogel, V., de Ruiter, C., van Beek, D., & Mead, G. (2004). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior*, *28*, 235–251. doi: 10.1023/B:LAHU.0000029137.41974.eb

Ferguson, G. A. (1976). *Statistical analysis in psychology and education* (4th ed.). New York: McGraw-Hill.

Grann, M., & Pallvik, A. (2002). An empirical investigation of written risk communication in forensic psychiatric evaluations. *Psychology, Crime & Law*, *8*, 113–130. doi: 10.10801068316020840181 2

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* New York: Erlbaum.

Haag, A. M. (2005). Do psychological interventions impact on actuarial measures: An analysis of the predictive validity of the Static-99 and Static-2002 on a re-conviction measure of sexual recidivism. *Dissertations Abstracts International*, *66* (*08*), 4531B. *(UMI No. NR05662)*

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 2007–05). Ottawa, ON: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/crp2007–05-en.pdf

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. doi: 10.1007/s10979–009-9180–1

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

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

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

Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, ON: Solicitor General Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/_fl/2003–03-stcc de-eng.pdf

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumière, M. L., Boer, D., & Lang, C. (2003). A multi-site comparison of actuarial risk instruments for sex offenders. *Psychological Assessment*, *15*, 413–425. doi: 10.1037/1040–3590.15.3.413

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

Hayes, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.

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

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

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M. & Harris, A. J. R. (in press). 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.*

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

Hilton, N. Z., Carter, A., Harris, G. T., Sharpe, A. J. B. (2008). Does using nonumerical terms to describe risk aid violence risk communication? *Journal of Interpersonal Violence*, *23*, 171–188. doi: 10.1177/0886260507309337

Hilton, N. Z., Harris, G. T., & Rice, M. E. (2010). *Risk assessment for domestically violent men: Tools for criminal justice, offender interventions, and victim services*. Washington, DC: American Psychological Association.

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

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. *Sexual Abuse: A Journal of Research and Treatment*, *19*, 425–448. doi: 10.1007/s11194–007-9062–3

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, The American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage Publications.

Knight, R. A., & Thornton, D. (2007). *Evaluating and improving risk assessment schemes for sexual recidivism: A long-term follow-up of convicted sexual offenders* (Document No. 217618). Submitted to the U.S. Department of Justice. Retrieved from http://www.ncjrs.gov/pdffiles1/nij/grants/217618.pdf

Kong, R., Johnson, H., Beattie, S., & Cardillo, A. (2003). Sexual offences in Canada. *Juristat*, *23*(*6*), 1–26.

Kropp, R. P. & Gibas, A. (2010). The Spousal Assault Risk Assessment Guide (SARA). In R.K. Otto & K.S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 227–250). New York: Taylor & Francis.

Kropp, R. P., Hart, S. D., Webster, C. W., & Eaves, D. (1999). *Spousal Assault Risk Assessment: User's Guide.* Toronto, ON, Canada: Multi-Health Systems, Inc.

Lang, T. A., & Secic, M. (2006). *How to report statistics in medicine: Annotated guidelines for authors, editors, and reviewers* (2nd ed.). Philadelphia, PA: American College of Physicians.

Langton, C. M. (2003). Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy. *Dissertations Abstracts International*, *64* (04), 1907B. (UMI No. NQ78052).

Langton, C. M., Barbaree, H. E., Hansen, K. T., Harkins, L., & Peacock, E. J. (2007). Reliability and validity of the Static-2002 among adult sexual offenders with reference to treatment status. *Criminal Justice and Behavior*, *34*, 616–640. doi: 10.1177/0093854806296851

Levenson, J. S., Brannon, Y. N., Fortney, T., & Baker, J. (2007). Public perceptions about sex offenders and community protection policies. *Analyses of Social Issues and Public Policy*, *7*, 137–161.

Ley, P. (1972). *Quantitative aspects of psychological assessment: An introduction.* London, UK: Duckworth.

Marshall, P. (1997). *The prevalence of convictions for sexual offending. Home Office Research and Statistics Directorate Research, Finding No. 55*. London: UK Home Office.

Marth, M. (2008). Adult criminal court statistics, 2006/2007. *Juristat*, *28*(5), 1–21.

McGrath, R. J., Cumming, G. F., & Burchard, B. L. (2003). *Current practices and trends in sexual abuser management: The Safer Society 2002 nationwide survey*. Brandon, VT: The Safer Society Foundation, Inc.

Minnesota Department of Corrections (2007). *Sex offender recidivism in Minnesota: April 2007*. St. Paul, MN: *Author*. *Retrieved from http:*//www.doc.state.mn.us/publications/documents/04–07SexOffenderReport-Recidivism.pdf

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal of Forensic Mental Health*, *2*, 1–6.

Phenix, A., Doren, D., Helmus, L., Hanson, R. K., & Thornton, D. (2009). *Coding rules for Static-2002*. Ottawa, ON: Public Safety Canada. Retrieved from http://www.publicsafety.gc.ca/res/cor/rep/sttc-2002-eng.aspx

Public Safety Canada (2010). *Corrections and conditional release statistical overview: Annual report 2010*. Ottawa, ON: Author.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

Reilly, J. J. (2010). Assessment of obesity in children and adolescents: Synthesis of recent systematic reviews and clinical guidelines. *Journal of Human Nutrition and Dietetics*, *23*, 205–211. doi:10.1111/j.1365–277X.2010.01054.x

Roussos, A., Francis, K., Zoubou, V., Kiprianos, S., Prokopiou, A., & Richardson, C. (2001). The standardization of Achenbach's Youth Self-Report in Greece in a national sample of high school students. *European Child & Adolescent Psychiatry*, *10*, 47–53. doi:10.1007/s007870170046

Sjöstedt, G., & Långström, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior*, *25*, 629–645. doi: 10.1023/A:1012758307983

State Authorized Risk Assessment Tool for Sex Offenders Review Committee. (February 10, 2011). [*California Department of Corrections and Rehabilitation and Facts of Offence reported Static-99 scores for the years 2008, 2009, and 2010*]. Unpublished raw data: Sacramento, California.

Taylor, M. J., & Heaton, R. K. (2001). Sensitivity and specificity of WAIS-III/WMS-III demographically corrected factor scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *7*, 867–874. doi:10.1017/S1355617701766118

Webster, C. (2010, June). *From flipping coins to looking at both sides of them: Assessing violence risk and strengths over the short-term*. Presentation at the Annual Convention of the Canadian Psychological Association, Winnipeg, MB, Canada.

Wittgenstein, L. (1976). (G.E.M. Anscombe, Trans.). *Philosophical investigations.* Oxford: Basil Blackwell.

APPENDIX

TABLE A1
Static-99 Frequency Distributions for the Probation (Non-custodial), Provincial, and Federal Samples

	Probation		Provincial		Federal	
Static-99 Score	Frequency	Cumulative Percent	Frequency	Cumulative Percent	Frequency	Cumulative Percent
n	303		254		1,454	
0	39	12.9	30	11.8	198	13.6
1	69	35.6	36	26.0	230	29.4
2	73	59.7	51	46.1	258	47.2
3	54	77.6	62	70.5	247	64.2
4	32	88.1	24	79.9	215	79.0
5	18	94.1	20	87.8	129	87.8
6	12	98.0	18	94.9	85	93.7
7	4	99.3	6	97.2	57	97.6
8	2	100.0	4	98.8	26	99.4
9	–	–	3	100.0	6	99.8
10	–	–	–	–	3	100.0
11	–	–	–	–	–	–
12	–	–	–	–	–	–

TABLE A2
Static-99R Frequency Distributions for the Probation (Non-custodial), Provincial, and Federal Samples

	Probation		Provincial		Federal	
Static-99R Score	Frequency	Cumulative Percent	Frequency	Cumulative Percent	Frequency	Cumulative Percent
n	303		254		1,454	
–3	8	2.6	7	2.8	35	2.4
–2	13	6.9	4	4.3	41	5.2
–1	27	15.8	15	10.2	141	14.9
0	36	27.7	21	18.5	134	24.2
1	54	45.5	35	32.3	190	37.2
2	50	62.0	51	52.4	186	50.0
3	53	79.5	45	70.1	214	64.7
4	28	88.8	30	81.9	192	77.9
5	20	95.4	20	89.8	139	87.5
6	9	98.3	9	93.3	97	94.2
7	4	99.7	10	97.2	49	97.5
8	1	100.0	6	99.6	20	98.9
9	–	–	1	100.00	13	99.8
10	–	–	–	–	3	100.0
11	–	–	–	–	–	–
12	–	–	–	–	–	–

TABLE A3
Static-2002 Frequency Distributions for the Probation (Non-custodial), Provincial, and Federal Samples

	Non-Custodial		Provincial		Federal	
Static-2002 Score	Frequency	Cumulative Percent	Frequency	Cumulative Percent	Frequency	Cumulative Percent
n	303		254		1,454	
0	17	5.6	9	3.5	59	4.1
1	39	18.5	22	12.2	121	12.4
2	54	36.3	32	24.8	159	23.3
3	57	55.1	40	40.6	211	37.8
4	56	73.6	44	57.9	207	52.1
5	36	85.5	46	76.0	233	68.1
6	19	91.7	29	87.4	190	81.2
7	13	96.0	9	90.9	129	90.0
8	10	99.3	7	93.7	78	95.4
9	1	99.7	10	97.6	33	97.7
10	1	100.0	6	100.0	23	99.2
11	–	–	–	–	8	99.8
12	–	–	–	–	3	100.0
13	–	–	–	–	–	–
14	–	–	–	–	–	–

TABLE A4
Static-2002R Frequency Distributions for the Probation (Non-custodial), Provincial, and Federal Samples

	Non-Custodial		Provincial		Federal	
Static-2002R Score	Frequency	Cumulative Percent	Frequency	Cumulative Percent	Frequency	Cumulative Percent
n	303		254		1,454	
–2	10	3.3	6	2.4	26	1.8
–1	12	7.3	4	3.9	32	4.0
0	21	14.2	16	10.2	108	11.4
1	37	26.4	17	16.9	128	20.2
2	54	44.2	38	31.9	159	31.2
3	61	64.4	42	48.4	178	43.4
4	43	78.5	43	65.4	216	58.3
5	34	89.8	41	81.5	208	72.6
6	15	94.7	22	90.2	178	84.8
7	8	97.4	4	91.7	109	92.3
8	6	99.3	7	94.5	56	96.2
9	2	100.0	12	99.2	30	98.2
10	–	–	2	100.0	18	99.5
11	–	–	–	–	7	99.9
12	–	–	–	–	1	100.0
13	–	–	–	–	–	–