Measuring the disparity of categorical risk among various sex offender risk assessment instruments

Sandy Jung, Anna Pham, Liam Ennis

This is an Accepted Manuscript of an article published by Taylor & Francis Group in *The Journal of Forensic Psychiatry & Psychology* on 18/06/2013, available online: http://dx.doi.org/10.1080/14789949.2013.806567

Measuring the Disparity of Categorical Risk Among Various Sex Offender Risk Assessment Measures

Sandy Jung

MacEwan University

Anna Pham

MacEwan University

Liam Ennis

Integrated Threat Risk Assessment Centre/Alberta Law Enforcement Response Teams

Author Note:

We gratefully acknowledge the support provided by the Northern Alberta Forensic Psychiatry Services and the Faculty of Arts & Science and the Research Office of Grant MacEwan University, the earlier contributions of Alberto Choy and Tarah Hook, and the research assistance of Lisa Jamieson, Bazyl Stefanski, John DeCesare, Shayla Stein, Kelly Butler, Gabriela Corabian, and Melissa Daniels. We also wish to thank two anonymous reviewers for their helpful comments and suggestions in improving this manuscript.

Correspondence concerning this article should be addressed to Sandy Jung, Department of Psychology, MacEwan University, P.O. Box 1796, Edmonton, Alberta, T5J 2P2, Canada. E-mail address: sandy.jung@macewan.ca

Abstract

The focus on reducing sexual offending has led to the development of risk assessment measures and schemes to predict reoffending, prioritize the allocation of treatment and supervision resources, and ensure public safety. However, different risk assessment approaches may not always have high agreement on the same individual. In light of the research indicating that ordinal risk rankings are most commonly used and reported in various risk communications, this study compares four risk assessment approaches, namely, the Static-99R, Static-2002R, SORAG, and SVR-20, in order to evaluate the disparities among the risk categories of these measures. The results indicate that there are disparities between all of the risk measures, but many of these can be explained by structural differences and common overlapping dimensions in the measures. Possible explanations for and implications of the discrepancies, along with some guidance on determining which approach to use, are discussed.

Measuring the Disparity of Categorical Risk Among Various Sex Offender Risk Assessment Measures

Significant empirical contributions to the literature have changed our understanding of risk assessment; be it risk for general, violent, or sexual risk to offend (e.g., Andrews, Bonta, & Wormith, 2004; Quinsey, Harris, Rice, & Cormier, 2006). The increase in the number of available tools has increased substantially with 58.2% of American correctional facilities using standardized tools to screen or assess offenders (The National Criminal Justice Treatment Practices Survey; see Taxman, Cropsey, Young, & Wexler, 2007). There are two broad approaches to risk assessment, namely, the use of actuarial risk measures and the process of structured professional judgments. Actuarial measures prescribe rules for scoring or coding information and then combine this information to generate numerical rankings, which then place individual offenders into specified risk categories (Quinsey et al., 2006). Structured professional judgment procedures are theoretically based schemes where important variables are isolated and defined. The scheme acts as an aide-memoire, to ensure that crucial factors are not overlooked in a clinical decision regarding a patient's level of risk (Bloom, Webster, Hucker, & De Freitas, 2005). Because some studies have shown that clinicians commonly report the findings from multiple risk assessment instruments and schemes (e.g., Doren, 2002) and that categorical risk is most frequently reported (e.g., Doyle, Ogloff, & Thomas, 2011), the present study aims to examine the disparity among risk assessment procedures used with sexual offenders.

The predictive accuracy of actuarial risk assessment measures and structured professional judgment schemes have been extensively examined in the literature (see Hanson & Morton-Bourgon, 2009; Langton et al., 2007; Rettenberger, Matthes, Boer, & Eher, 2010). Generally, published measures of and methods for assessing risk for sexual recidivism are quite good at predicting sexual reoffending behaviour. Moreover, well-validated approaches to risk assessment seem to already include strong risk factors predictive of sexual and violent recidivism (Kroner, Mills, & Reddon, 2005). Hence, the use of these protocols is part of best practices (ATSA, 2005).

In terms of using these protocols, it has been recommended that risk communication should be presented in terms of probabilities and percentiles rather than categorical descriptors (Hanson, 2009). However, in a study by Heilbrun, Philipson, Berman, and Warren (1999), a third of clinicians in their sample (18 of 54) used categories in communicating their conclusions about risk. The authors reported that in their second study with a separate sample of clinicians, categorical risk was preferred in communicating conclusions (i.e., mean rating of 3.9 out of 5, where 5 refers to the communication approaches as essential). Similar findings were reported in forensic psychiatric assessment reports; 91 of 122 reports where risk of recidivism was mentioned used some categorical degree of risk in their communication (Grann & Pallvik, 2002). Heilbrun, O'Neill, Strohman, Bowman, and Lo (2004) further surveyed 1000 psychologists, and of 256 respondents, descriptive risk communication that used risk categories was preferred more than using percentages or probabilities. In a more recent study by Doyle et al. (2011), 86 forensic evaluation reports of dangerous sex offenders were reviewed. Almost all of the reports (98.7%) reported a categorical method of risk communication. Moreover, the authors reported that multiple methods of risk assessment were regularly utilized (e.g., actuarial, clinical judgment, and empirically guided in more than 84% of reports). Outside of clinical judgment, multiple tools were used in 80.2% of the reports.

Given that ordinal risk rankings are commonly used by evaluators, it would be important to examine the concordance among risk assessment procedures in terms of these risk categories.

Barbaree, Langton, and Peacock (2006) outline two issues in the availability of effective risk assessment approaches. They note that evaluators must make conscious choices among these approaches in each evaluation and that they must address discrepant outcomes when multiple approaches are used. Barbaree et al. examined the consistency of ranking risk among five risk assessment measures: RRASOR, Static-99, Violence Risk Appraisal Guide (VRAG), Sex Offender Risk Appraisal Guide (SORAG), and MnSOST-R. They predicted that, based on the fact that items in each instrument would represent underlying risk factors differently, rankings would not be consistent. Although 55% of the sample was identified by at least one instrument as being high risk, only 3% of the sample was identified as high risk by all five instruments. Similarly, only 4% were identified as low risk simultaneously by all of the measures. Hence, different instruments appear to yield different outcomes (Barbaree et al., 2006). One may query why do different instruments—all of which purport to predict recidivism—yield different outcomes? Perhaps one approach is better than the others. Skeem and Monahan (2011) responded to the question of whether one approach predicts better than another by examining the properties of risk measures. They draw attention to something that most researchers in the field readily acknowledge, which is the difference in the "common factors" of risk included in these measures. The structural components in existing measures are similar in some ways (e.g., often including historical offences) and different in other ways (e.g., weighing more heavily on some variables than others), and this may account for why we see discrepancies. Hence, if two measures have shared risk factors, overlapping dimensions, or common items (e.g., criminal history, criminal attitudes, substance abuse problems), then it is likely that these measures would have greater concordance. On the other hand, if these measures have non-overlapping dimensions, then we would expect little or smaller concordance between the measures.

Because evidence currently suggests that different risk appraisal instruments yield discrepant results (Barbaree et al., 2006), the present study attempts to expand on this work by using paired comparisons in two ways. First, this study explores the degree of disparity among actuarial measures—two of which should have conceptually similar dimensions—and a structured professional judgment scheme for sexual offenders. Risk measures and schemes include the Static-99R, Static-2002R, SORAG, and SVR-20, which are more commonly used to assess sexual recidivism risk (McGrath, Cumming, Burchard, Zeoli, & Ellerby, 2010). Second, we use published norms to create the risk categories to which offenders are assigned based on the same percentile cut-offs. Although rank ordering of the sample was used in Barbaree et al.'s (2006) study to determine risk category assignment, using previously published norms may allow for a greater equivalence of the normative samples. Since much of the normative samples for extant measures are based on Canadian correctional samples, using published norms is appropriate, albeit not perfect (i.e., samples are drawn from varying degrees of institutional security, e.g., maximum vs. minimum secure facilities). In this study, we attempt to address whether actuarial measures of sexual violence risk have high between-instrument agreement for categorizing sexual offenders, whether the clinician rating of categorical risk using a structured professional judgment (SPJ) scheme matches the mechanically-derived risk formed from the SPJ items, and whether actuarial measures have high agreement with an SPJ measure in categorizing sexual offenders on sexual violence risk.

Method

Participants

The setting of this study is a forensic psychiatric outpatient clinic in a Canadian city. The clinic offers court-ordered assessment services and treatment services that address offending behaviours and mental health. The clinic staff comprises an interdisciplinary team of psychologists, psychiatrists, social workers, nurses, and counselors. For this study, the clinical files of patients who were referred and/or court-mandated for an assessment of their risk and treatment needs following receipt of formal convictions for their sexual offences were reviewed. In total, 484 male sexual offenders were identified and subsequently, their files were coded. Of the total sample size of 484, 32 were missing complete scores for any of the risk assessment tools included in this study. Because not all risk variables were available in the files for all offenders, sample sizes varied depending on the analysis. Four hundred and nineteen offenders had calculable Static-99R scores, 318 had completed Static-2002R scores, and 124 had calculable SORAGs. SVR-20 protocols were coded directly from clinician ratings, and there were 74 files with completed SVR-20 protocols.

The offenders in our sample were convicted for at least one sexual offence with the average age of the sample being 36.2 years (*SD* = 13.57). The average years of education was 11.4 years (*SD* = 2.50) with the highest level of education recorded at 20 years. The majority of participants were classified as unskilled labourers (26.7%) and semiskilled worker/operators (23.1%) on the date of their index offense. Regarding marital status, most of the participants were single (43.1%), and almost a third were married (30.0%). More than half of the participants (59.7%) had previously participated in sex offender treatment, although rates of completion were unknown.

Measures

Four measures of risk for sexual recidivism were included in this study and each is described below. The three actuarial measures (Static-99R, Static-2002R, and SORAG) were scored retrospectively by trained research assistants based on historical files.

Static-99R. The Static-99R is a revised version of the Static-99 (Hanson & Thornton, 1999, 2000). The original version was developed to assess risk for sexual recidivism among adult males known to have committed at least one sexual offense. The original Static-99 contains 10 items, and total scores could range from 0 to 12, which place individuals into one of four risk categories. The Static-99 has demonstrated to have good interrater reliability, intraclass correlation (*ICC*) = .98 (Rettenberger et al., 2010), and good predictive validity for sexual, general violent, and general criminal recidivism (*AUC*s = .71, .71, and .70, respectively; Rettenberger et al., 2010). The authors of the Static-99 revised the instrument to include weights for the age at release item (Helmus, Thornton, Hanson, & Babchishin, 2011) and renamed it the Static-99R. The remainder of the items on the Static-99R is identical to the original form, but the scores on the Static-99R now range from -3 to 12.

Static-2002R. The Static-2002R is a revised version of the Static-2002 (Phenix, Doren, Helmus, Hanson, & Thornton, 2008). Unlike the Static-99 or Static-99R, the Static-2002 and its revised version, Static-2002R, were developed to assess some theoretically meaningful characteristics that are presumed to be the cause of recidivism. It contains 14 items in five content areas: a single age at release item, persistence of sexual offending items (3 items), deviant sexual interests items (3 items), relationship to victims items (2 items), and general criminality items (5 items). Total scores range from 0 to 14, placing individuals into one of five risk categories. Interrater reliability has been shown to be very good, *ICC* = .98 (Helmus & Hanson, 2007). Predictive validity of the Static-2002 for nonviolent, any, serious, and sexual recidivism were at the moderate level (*r*s = .65, .69, .70, and .70, respectively). The Static-2002 was revised by the developers (Helmus, Thornton, et al., 2012) to include weighted scores reflecting higher risk for younger offenders for the age at release item. The remaining items on the renamed Static-2002R are identical to the Static-2002, but the scores range from -2 to 13 (Helmus, Thornton, et al., 2012).

Modified SORAG. The SORAG (Quinsey et al., 2006) was developed to assess risk for violent recidivism (including sexual offenses involving physical contact with the victim) among adult sex offenders. The SORAG is composed of 14 items, and the total score on the SORAG can range from -27 to 51. Individuals are assigned to one of nine risk categories, ranging from 1 (lowest) to 9 (highest) according to their total score. Interrater reliability was good with an *ICC* of .93 (Rettenberger et al., 2010). Predictive validities of the SORAG for sexual, general violent, and general criminal recidivism were in the moderate range (*r*s = .69, .72, and .75, respectively; Rettenberger et al., 2010).

Due to missing data, two items were omitted from the calculation of the total SORAG score: (1) sex and age of index victim and (2) meets DSM critieria for schizophrenia. As recommended by the authors of the measure and other researchers, the Screening Scale for Pedophilic Interests (SSPI) was used to score the sexual deviance item (Seto & Lalumière, 2001), and the Child and Adolescent Taxonomy (CATS) was used as a proxy for the psychopathy item (Quinsey et al., 2006). Hence, in our study, the modified SORAG score was calculated based on 12 of the original 14 items, namely, the following: lived with both biological parents until age 16, elementary school maladjustment, history of alcohol problems, marital status, nonviolent offense history, violent offense history, sexual offense history, age at index offense, failed on prior conditional release, DSM-III diagnosis of any personality disorder, psychopathy, and sexual deviance.

SVR-20. The SVR-20 (Boer, Hart, Kropp, & Webster, 1997) is an assessment procedure based on structured professional judgment. It contains 20 items that address three themes: (1)

psychosocial adjustment, (2) sexual offenses, and (3) future planning. Similar scoring was used to produced a total score for each offender as used in other empirical endeavours (e.g., de Vogel, de Ruiter, van Beek, & Mead, 2004; Langton, 2003; Parent, Guay, & Knight, 2011), where a 0 was assigned when the factor was not present for the offender, 1 when there was some suggestive evidence, and 2 when there was clear evidence for the factor's presence. It is important to note that, although this scoring scheme has been used in multiple empirical papers, this procedure was not endorsed by the authors of the SVR-20. Using this quantified scoring scheme, the total score for the SVR-20 could range from 0 to 40. Interrater reliability was good with an *ICC* of .75 (Parent et al., 2011). The predictive validity of the SVR-20 for sexual, general violent and general criminal recidivism was considered good (*AUC*s = .77, .61, and .67, respectively; Rettenberger et al., 2010). In addition to the quantified risk score, the clinician's overall rating of low, moderate, or high risk was included in the analysis (guided by the presence of factors; note that on the original form, quantifiable scores are not used).

Procedures

All sex offender case files were retrieved from an outpatient forensic psychiatric facility in central Alberta. The case files that were reviewed contained assessment reports, criminal records, case notes, offender reports, demographic information, and some description of victim information. Files were coded retrospectively and no additional measures were administered for the purpose of the research. Because not all variables were available in the files for all offenders, sample sizes varied depending on the analysis. The present study is part of a larger database in which 406 variables were coded from the case files on each offender. To ensure we would maintain strong interrater reliability, four research assistants received a full-day of training on the variables and were examined on three cases to ensure they reliably coded the variables.

Subsequently, five offenders' case files were coded independently by two raters. For a majority of the variables included in this research (i.e., no calculations were conducted when assistants did not code a variable on more than 2 of the 5 files; this only applied to seven of the 37 variables), percentage agreements were calculated, and 26 variables had 75% to 100% agreement between coders. Four variables had an agreement of 66% or less and included number of marriages, access to mental health services, whether the offender used drugs prior to the index offence, and the number of prior convictions.

Similar to the procedure used by Barbaree et al. (2006), percentile ranks taken from the published norms for each measure were used to establish three risk categories using a consistent percentile cut-off. The norms used in our study include those produced and reported on the Static-99 clearinghouse website (www.static99.org; November 2011 percentile tables for Static-99R and Static-2002R), in publications with normative data by Quinsey et al. (2006; SORAG) and Langton (2003; SVR-20). We created 3 categories for each measure with the low risk category including offenders with scores that fell between the 1st and 38th percentiles, moderate risk category with scores that fell between the 39th and 91st percentiles, and high risk category with scores that exceeded the 91st percentile. The cut-off scores were chosen to correspond with the Static-99R's cut-off scores for their risk categories and therefore the corresponding percentiles (e.g., scores of -3 to 1 were considered low risk while scores of 2 to 5 were considered low-moderate and moderate-high risk). Hence, the score range for each category are as follows: (a) Static-99R, low -3 to 1, moderate 2 to 5, high 6 to 12; (b) Static-2002R, low -2 to 2, moderate 3 to 6, high 7 to 13; (c) SORAG, low -27 to 2, moderate 3 to 24, high 25 to 51; and (d) SVR-20, low 0 to 17, moderate 18 to 26, high 27 to 40.

Results

Paired comparisons were made between the risk categories of the four risk measures, Static-99R, Static-2002R, SORAG, and SVR-20 using percentage agreement. For the SVR-20, both the clinician-assigned risk categories (i.e., structured professional judgment) and those systematically calculated (i.e., mechanically-derived) were compared to further assess the difference between the risk categories within the same measure. Pearson product moment correlation coefficients were used to analyze the total scores, and Spearman's rho was also used to examine the association between the rank-ordered risk categories for each comparison. Descriptive statistics of each risk assessment measure are shown in Table 1.

Overall percentage agreements among the actuarial instruments are listed in Table 2. Despite the similarity in development and items between the Static-2002R and its predecessor, the Static-99R, only moderate percentage agreements among the risk categories of these two measures were seen, ranging from 36.2% to 81.1% (see Table 3 for frequencies and percentage agreements by risk category). The total percentage agreement was 62.9%, which was lower than would be expected when examining two such closely related measures. A positive, significant correlation emerged between the total scores of the Static-99R and the Static-2002R, $r(278) = 0.64$, $p < .001$, and between the risk categories of each measure, Spearman's rho = 0.49, $p < .001$. When a paired comparison was made between the risk categories of the Static-99R and the risk categories of the SORAG, the percentage agreements were low to moderate, ranging from 10.5% to 100% (see Table 3). The total percentage agreement between these two measures was poor at 37.2%, which is to be expected given that the two instruments were created to predict somewhat different criterion (i.e., sexual vs. violent). Correlational analyses produced significant findings for the total scores, $r(76) = 0.55$, $p < .001$, and for the risk categories, Spearman's rho = 0.35, $p < .01$. Similar percentage agreements emerged in the paired comparison between the Static-2002R risk categories and the SORAG risk categories, as listed on Table 4. Percentage agreements between the two measures ranged from 11.8% to 91.3% with the total percentage agreement being 50.7%. Correlations were significant for the total score, $r(64) = 0.44$, $p <.001$, and the risk categories, Spearman's rho = 0.40, $p < .001$.

The actuarial measures were then compared with a structured professional judgment measure, the SVR-20. Frequencies and percentage agreements are listed in Tables 5 and 6. Paired comparisons of the risk categories of the Static-99R with both the clinician ratings and the systematically calculated risk categories of the SVR-20 yielded minimal agreement. The percentage agreements between the Static-99R and the SVR-20 were wide, ranging from 8.7% to 100%; the total percentage agreement was 23.2%. The association between the Static-99R and SVR-20's total scores, $r(68) = 0.44$, $p <.001$, and risk categories, Spearman's rho = 0.36, $p < .01$, were significant and positive. Strong agreement was not expected between the SVR-20 clinician rating and the Static-99R (see Table 5), since the former is a clinician's decision that is guided by empirically derived risk factors, while the latter uses a fixed algorithm that is based on empirical risk factors. As expected, the clinician-assigned risk categories had poor agreement with the Static-99R (see Table 6), but had a higher overall percentage agreement than the systematically calculated risk categories (46.3%). The correlation between the Static-99R and SVR-20 clinician rated risk categories was positive and significant for the risk categories, Spearman's rho = 0.38, $p < .01$.

The overall percentage agreements between the Static-2002R risk categories and two sets of SVR-20 risk categories were similar. With the calculated SVR-20 scores, percentage agreements ranged from 0% to 100%, with an overall agreement of 43.7% (see Table 5).

Significant correlations emerged for the total score, $r(70) = 0.43$, $p < .001$, and the risk categories, Spearman's rho = 0.31, $p < .01$. This was comparable to the paired comparison between the Static-2002R and SVR-20 clinician ratings, which ranged from 29.6% to 50% and had an overall percentage agreement of 38.2% (see Table 6). No significant correlation between the Static-2002R risk categories and the SVR-20 clinician ratings emerged, Spearman's rho = 0.19, *ns*.

The paired comparison of the SORAG and the quantified SVR-20 risk rankings showed moderate to high agreement of 71.4% and wide ranging agreements for each risk category (0 to 92.9%; see Table 5), although the sample size was quite small ($n = 21$). The correlation between the SORAG and the SVR-20's total score was significant, $r(22) = 0.63$, $p < .01$, as was the correlation between the SORAG risk rankings and the SVR-20's clinician-assigned risk categories, Spearman's rho = 0.47, $p < .05$. The percentage agreement between the risk categories of the SORAG and the clinician-rated SVR-20 risk categories, ranged from 25% to 100% with an overall percentage agreement of 46.7% (see Table 6). Again, the sample size was very small ($n = 15$) and did not produce a significant correlation, Spearman's rho = 0.49, *ns*.

An additional paired comparison was made between the clinician ratings of the SVR-20 and the systematically calculated risk categories of the SVR-20. Although this was a within-instrument comparison of the same risk measure, the results were not much different from comparisons of different measures. The percentage agreements were wide, ranging from 12% to 100%, with an overall percentage agreement of 46.4% (see Table 6). However, the correlation between the systematically rated and empirically-derived risk rankings was significant, Spearman's rho = 0.56, $p < .001$.

Discussion

The current study examined the agreement in ordinal risk rankings among risk assessment approaches. Four measures were examined, and they included the Static-99R, Static-2002R, SORAG, and SVR-20, along with clinician's ranking empirically derived from the SVR-20. Specifically, we addressed three research questions: (1) do actuarial measures of sexual violence risk have high between-instrument agreement for categorizing sexual offenders, (2) does the clinician rating of categorical risk using an SPJ match the mechanically-derived risk formed from the SPJ items, and (3) do actuarial measures have high agreements with an SPJ measure in categorizing sexual offenders on sexual violence risk.

As recently suggested by Hanson, Lloyd, Helmus, and Thornton (2012), percentile ranks have some strengths as a non-arbitrary metric for risk communication. However, results from the current study revealed a large discrepancy even in closely related risk measures, using percentile ranks to determine cut-off scores. The paired comparison between the risk categories of the Static-99R and the Static-2002R showed a moderate percentage agreement of 62.9%. In spite of the fact that the Static-2002R was developed with the same intent as the Static-99R and by the same authors, which would suggest the two instruments are highly related, the results show that these two actuarial measures are not so similar when it comes to risk rankings, based on percentile cut-offs using large normative samples of the same individuals. Nonetheless, the Static-99R had the largest concordance with the Static-2002R as one would expect of the risk schemes included.

When pairings included the SORAG, the Static-2002R had greater concordance than with the Static-99R. The results from these comparisons seem consistent with the results from Barbaree et al.'s (2006) study. Their study found that the five actuarial instruments (VRAG, SORAG, RRASOR, MnSOST-R, and Static-99) did not produce consistent risk rankings for the same evaluations. The measures were expected to produce similar rankings of offenders. However, based on the substantial weight of evidence in favour of multiple underlying risk dimensions (Hanson & Morton-Bourgon, 2004); the discrepancy between the risk measures analyzed in this study is not surprising. Hanson and Morton-Bourgon (2005) have identified characteristics of persistent sexual offenders in their meta-analysis of recidivism studies. They found two characteristics that are commonly present in persistent sexual offenders. The first is deviant sexual interests, which are enduring attractions to sexual acts that are illegal or highly unusual. The second is antisocial orientation/lifestyle instability, which are antisocial traits and a history of rule violation. Both of these factors are present to varying degrees in these actuarial measures, and they are measured differently. If we examine the overlapping dimensions in each of these measures, there appear to be more similarities between the SORAG and the Static-2002R. For instance, deviant sexual interests is measured on the SORAG using the SSPI score, which comprises 4 items (see Seto & Lalumière, 2001), and the Static-2002R includes several items that in common with the SSPI. There seems to be fewer overlapping features between the SORAG and the Static-99R, which is reflected in our finding of a lower concordance rate of 37.2%.

When we examined two different types of risk assessment approaches (i.e., actuarial vs. structured professional judgment), varying results emerged. Low concordance was found (albeit significant) between the Static measures and the quantified SVR-20. Again, some dimensions are measured differently. For example, in the Static measures, sexual deviance is assessed with items, such as male victims and stranger victims, while the SVR-20 includes a single item of sexual deviance coded on the basis of clinicians’ rating of diagnostic criteria defined in the SVR-

20 manual (or phallometric testing, if available). A higher concordance emerged between the SVR-20 and the SORAG. This finding is interesting given that the SORAG and SVR-20 are designed to assess different criminal outcomes: violent (including sexual) vs. sexual re-offending. However, if we revisit Skeem and Monahan's (2011) emphasis on common factors, both the SORAG and the SVR-20 include psychopathy and sexual deviance. In contrast, it is notable that the Static measures heavily weigh on criminal history, hence suggesting that structural similarities could explain the larger concordance seen in the ordinal risk rankings between the SVR-20 and the SORAG.

Our study also examined the concordance between the clinician's rating empirically guided by the SVR-20 and the actuarial measures and the quantified SVR-20 total. In using the SVR-20, a clinician who scores an offender on each of the SVR-20 items would typically use these scores to empirically guide him/her in determining the ordinal risk category for the same offender. The percentage agreements were relatively low (all fell below 50% concordance). It was surprising that clinicians' ratings were higher in concordance with the Static-99R than the quantified SVR-20 total. However, the difficulty in interpreting these results is that we are unaware of the weighting that a clinician may have placed on some factors over others. Moreover, on the SVR-20, there is the ability to code a variable's relevance (Boer et al., 1997), and this was not coded in our study. Therefore, any given clinician could have placed more weight on certain variables over others given the imminence or seriousness of the variable in a particular offender's case, hence, rendering differences seen in the risk classification between the quantified SVR-20 ranking and the clinician's empirically derived ranking.

It is important to draw attention to the limitations of the current study. The range of scores from the database did not fully encompass the range of possible total scores for each measure analyzed especially at the high-risk polarity. For example, the total scores of the SORAG can range from -27 to 51, but the scores obtained from the current sample ranged from -19 to 25. The entire sample was large, but the actual analyses were limited by relatively little overlap in the availability of each pairing of measures; hence, larger sample sizes would have greatly improved the external validity of the study. One of the strengths of the current study was that the categories were determined from previously published norms (in contrast to Barbaree et al., 2006); however, normative samples for each measure may not be equivalent in terms of average risk level. For example, the Static-99R and the Static-2002R percentile norms were based on Canadian sex offenders where many would not have necessarily received prison sentence for their index offence (Helmus, Hanson, et al., 2012), whereas the SORAG norms were based on offenders referred for assessment at a maximum security penitentiary (Quinsey et al., 2006) and the SVR-20 norms were based on sexual offenders who were referred to a moderate intensity treatment program (Langton, 2003). Nonetheless, the patterns of disagreements appear to be consistent with the differences in the normative groups. For example, when comparing the SORAG and the Static-99R, there were 49 individuals who were off-diagonal (i.e., risk level did not match), and 48 were rated as lower by the SORAG than the by the Static-99R. Further, given these comparative norms, it is not surprising to find that the sample in the current study was lower in risk overall compared to the range of scores available on the SORAG, given the difference in the sample recruitment.

The differences in resulting risk categories from different risk assessment measures for the same individual illustrate ongoing issues on how risk assessments should be communicated and continual debates on how they should be selected or how discrepancies should be resolved. Although several prominent researchers in the field have urged that evaluators use numerical descriptors and provide some discussion of psychologically meaningful causal risk factors in their reports (Hanson, Babchishin, Helmus, & Thornton, 2013; Monahan & Steadman, 1996), as previously mentioned, this is not common practice in the forensic evaluation field. Several studies have already demonstrated that clinicians highly favored categorical statements about recidivism risk (Heilbrun et al., 1999; 2004; Heilbrun, O'Neill, Strohman, Bowman, & Philipson, 2000), and such risk rankings are commonly used when communicating risk in evaluation reports (Doyle et al., 2011; Grann & Pallvik, 2002; Heilbrun et al., 1999). However, the problem with low, moderate, and high descriptors is that they have no inherent scientific meaning (i.e., there was no benefit shown in including these categorical descriptors with numerical probabilities of risk for harm; Hilton, Carter, Harris, & Shape, 2008). Because these categories are prone to divergent interpretations (Hilton et al., 2008), some researchers have proposed that evaluators should offer clear definitions of these risk categories in their usage of such descriptors (Babchishin & Hanson, 2009).

In addition to the issues regarding how we should communicate risk, literature to date has provided little consistency on how evaluators should interpret differences among outcomes arising from the use of different risk assessment approaches. Clearly there are structural differences among existing validated measures, including the approaches examined in the present study. Given the knowledge that disparities exist among various risk assessment measures, it is clear that clinicians should be cautious when choosing risk measures and/or schemes and that this vigilance should extend to the interpretation of discrepant results from the use of multiple approaches. Some have suggested that clinicians should limit the use of different measures since structure has been shown to be unnecessary (Kroner et al., 2005; Seto, 2005); for example, if certain circumstances prohibit the use of one instrument, choosing the next appropriate measure to assess individual offenders should be considered, or if the normative sample on which the measure or scheme was based is more similar to the context or type of offender, then it would make sense to use a particular risk assessment approach. To illustrate, when choosing the appropriate risk measure to assess general violence recidivism of sexual offenders, the SORAG should be chosen over the SVR-20 because studies have shown that the SVR-20 fails to predict general violence (e.g., Rettenberger et al., 2010).

Several other researchers have since contested this single-measure view of risk assessment. In a large-scale meta-analysis using 20 samples and 7491 sex offenders, Babchishin, Hanson, and Helmus (2012) found that averaging estimations of risk using more than one measure produced better discrimination. They conclude that evaluators may benefit from using multiple risk measures. Furthermore, Skeem and Monahan (2011) address the question of determining which measure to use and instruct evaluators to ask what is the purpose of their evaluation (i.e., risk assessment vs. risk reduction) to guide them in their decision. Given the disparity reported in this present study, the authors would point to a combination of Babchishin's work (2009; 2012) and place emphasis on the importance of providing clear definitions in any reporting of one's findings. If averaging were to be used, it would be critical to ensure the reader (i.e., recipient of the risk assessment report) is made explicitly aware of the evaluator's interpretation of the scores or findings. Similarly, if consensus is to be attained (through multiple raters), similar explicit explanations and definitions should be given in the report as to the meaning of the findings—whether it be the unfavourable use of ordinal risk rankings or the more favoured used of numerical probabilities (and/or percentiles).

In conclusion, the present study provides further evidence of the disparity seen among risk measures and schemes when using ordinal risk rankings. This is obviously a problematic issue in court when making decisions based on the interpretations of different risk assessment approaches and when an individual evaluator uses more than one method in their risk communication. It is important to be aware of the structural differences between measures and the need for explicating interpretations of risk rankings in the evaluation report.

References

Andrews, D. A. & Bonta, J., & Wormith, J. S. (2004). *The Level of Service/Case Management Inventory (LS/CMI): User's Manual*. Toronto, Canada: Multi-Health Systems.

Association for the Treatment of Sexual Abusers (ATSA). (2005). *Practice standards and guidelines for members of the Association for the Treatment of Sexual Abusers*. Beaverton, OR: Author.

Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Going beyond "low," "moderate", and "high" in risk communication. *Crime Scene, 16*, 11-14.

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

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: 10.1177/107906320601800408

Bloom, H., Webster, C., Hucker, S., & De Freitas, K. (2005). The Canadian contribution to violence risk assessment: History and implications for current psychiatric practice. *Canadian Journal of Psychiatry, 50,* 3-11.

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

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior, 34,* 259-274. doi:10.1007/s10979-009-9176-x 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

Doren, D. M. (2002). *Evaluating sex offenders: A manual for civil commitments and beyond*. Thousand Oaks, CA: Sage.

Doyle, D. J., Ogloff, J. R. P., & Thomas, D. M. (2011). Designated as dangerous: Characteristics of sex offenders subject to post-sentence orders in Australia. *Australian Psychologist, 46,* 41-48. doi: 10.1111/j.1742-9544.2010.00006.x

Grann, M., & Pallvik, A. (2002). An empirical investigation of written risk communication in forensic psychiatric evaluations. *Psychology, Crime & Law, 8*, 113-130. doi:l0.1080/10683160290000923

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology, 50*, 172-182. doi: 10.1037/a0015726

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

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

Hanson, R. K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis*. Ottawa, Canada: Public Works and Government Services Canada. Cat. No.: PS3-1/2004-2E-PDF. ISBN: 0-662-36397-3.

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*(6), 1154-1163. doi: 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. doi: 10.1037/a0014421

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders* (User Report 2003-01). Ottawa, Canada: Department of the Solicitor General of Canada.

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

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

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. doi:10.1023/A:1005435005404

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

Helmus, L. & Hanson, K. (2007). Predictive validity of the Static-99 and Static- 2002 for sex offenders on community supervision. *Sexual Offender Treatment, 2,* 1-14.

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

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). 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., Harris, G. T., Rice, M. E., & Sharpe. A. J. B. (2008). Does using nonnumerical terms to describe risk and violence risk communication: Clinician agreement and decision making. *Journal of Interpersonal Violence, 23,* 171-188. doi: 10.1177/0886260507309337

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A Coffee Can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry, 28,* 360-374. doi: 10.1016/j.ijlp.2004.01.011

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*. Unpublished doctoral dissertation, University of Toronto, Toronto, Canada.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior, 34*, 37-59. doi:10.1177/0093854806291157

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

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. doi:10.1037/0003-066X.51.9.931

Mossman, D. (2006). Another look at interpreting risk categories. *Sexual Abuse: A Journal of Research and Treatment, 18,* 41-63. doi:10.1177/107906320601800104

Parent, G., Guay, J, & Knight, R. A. (2011). An assessment of long-term risk of recidivism by adult sex offenders: One size doesn't fit all. *Criminal Justice and Behavior, 38*, 188-209. doi:10.1177/0093854810388238

Phenix, A., Doren, D., Helmus, L., Hanson, R. K., & Thornton, D. (2008). *Coding Rules for Static-2002.* Ottawa, Canada: Public Safety Canada.

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

Rettenberger, M., Matthes, A., Boer, D. P., & Eher, R. (2010). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology, 54,* 169-186. doi:10.1177/0306624X08328755

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17*, 156-167. doi:10.1037/1040-3590.17.2.156

Seto, M.C., & Lalumière, M.L. (2001). A brief screening scale to identify pedophilic interests among child molesters. *Sexual Abuse: A Journal of Research and Treatment, 13*, 15–25. doi:10.1177/107906320101300103

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20,* 38-42. doi:10.1177/0963721410397271

Taxman, F. S., Cropsey, K. L., Young, D. W., & Wexler, H. (2007). Screening, assessment, and referral practices in adult correctional settings: A national perspective. *Criminal Justice and Behavior, 34,* 1216-1234. doi:10.1177/0093854807304431

Table 1

Descriptive statistics on the Static-99R, Static-2002R, SORAG, and SVR-20

Measure	*n*	*M*	*SD*	Range of scores	Converted risk categories (*n*)		
					Low (0-38%ile)	Moderate (39-90%ile)	High (91-100%ile)
Static-99R (-3 – 12)	361	3.46	2.17	-2 to 11	64 (17.7%)	243 (67.3%)	54 (15%)
Static-2002R (-2 – 13)	345	3.93	2.34	0 to 12	107 (31.0%)	187 (54.2%)	51 (14.8%)
SORAG (-27 – 51)	82	1.35	11.59	-19 to 25	44 (53.7%)	35 (42.7%)	3 (3.7%)
SVR-20 (0 – 40)	74	10.35	6.38	1 to 31	62 (83.8%)	9 (12.2%)	2 (2.7%)
SVR-20 (clinician rated)					27 (38.6%)	32 (45.7%)	11 (15.7%)

Table 2

Overall percentage agreement between risk categories of the Static-99R, Static-2002R, SORAG, SVR-20, and clinician rated SVR-20

	Static-2002R	SORAG	SVR-20 (Quantified)	SVR-20 (Clinician Rating)
Static-99R	62.9% (280) 0.49***	37.2% (78) 0.35**	23.2% (69) 0.36**	46.3% (67) 0.38**
Static-2002R		50.7% (78) 0.40***	43.7% (71) 0.31**	38.2% (68) 0.19
SORAG			71.4% (21) 0.47*	46.7% (15) 0.49
SVR-20 (Quantified)				46.4% (56) 0.56***

Note. Overall percentage agreement, total N (in parentheses), and Spearman's rho are provided. **p*<.05, ***p*<.01, ****p*<.001.

Table 3

Percentage agreement between risk categories of the Static-99R with the Static-2002R and the SORAG

	Static-99R			Percentage Agreement
	Low	Moderate	High	
Static-2002R (*N* = 280; *Spearman's rho* = 0.49, *p* < .001)				
Low	**34**	57	3	**36.2%**
Moderate	9	**120**	19	**81.1%**
High	2	14	**22**	**57.9%**
Percentage Agreement	**75.6%**	**62.8%**	**50.0%**	**62.9%**
SORAG* (*N* = 78; *Spearman's rho* = 0.35, *p* < .01)				
Low	**6**	31	6	**14.0%**
Moderate	0	**21**	11	**65.6%**
High	0	1	**2**	**66.7%**
Percentage Agreement	**100%**	**39.6%**	**10.5%**	**37.2%**

*SORAG risk categories are based on the total score of 12 SORAG items (excludes items 8, victim injury, and 12, diagnosis of schizophrenia, as these were not available for a majority of the sample)

Table 4

Percentage agreement between risk categories of the Static-2002R with the Static-99R and the SORAG

	Static-2002R			Percentage Agreement
	Low	Moderate	High	
SORAG* (*N* = 78; *Spearman's rho* = 0.40, *p* < .001)				
Low	**12**	19	5	**46.7%**
Moderate	2	**15**	10	**55.6%**
High	0	1	**2**	**66.7%**
Percentage Agreement	**91.3%**	**42.9%**	**11.8%**	**50.7%**

*SORAG risk categories are based on the total score of 12 SORAG items (excludes items 8, victim injury, and 12, diagnosis of schizophrenia, as these were not available for a majority of the sample)

Table 5

Percentage agreement between risk categories of the SVR-20 total score with the Static-99R, Static-2002R, and the SORAG

	SVR-20			Percentage Agreement
Actuarial Risk Measure	Low	Moderate	High	
Static-99R (*N* = 69; *Spearman's rho* = 0.36, *p* < .01)				
Low	**10**	0	0	**100%**
Moderate	42	**4**	0	**8.7%**
High	8	3	**2**	**15.4%**
Percentage Agreement	**16.7%**	**57.1%**	**100%**	**23.2%**
Static-2002R (*N* = 71; *Spearman's rho* = 0.31, *p* <.01)				
Low	**25**	0	0	**100%**
Moderate	29	**6**	1	**16.7%**
High	7	3	**0**	**0%**
Percentage Agreement	**41.0%**	**66.7%**	**0%**	**43.7%**
SORAG (*N* = 21; *Spearman's rho* = 0.47, *p* < .05)				
Low	**13**	1	0	**92.9%**
Moderate	4	**2**	0	**33.3%**
High	0	1	**0**	**0%**
Percentage Agreement	**76.5%**	**50.0%**	**-**	**71.4%**

Table 6

Percentage agreement between the clinician rated risk categories from the SVR-20 with the risk categories of the Static-99R, Static-2002R, SORAG, and SVR-20

	SVR-20 Clinician ratings			Percentage Agreement
	Low	Moderate	High	
Static-99R (*N* = 67; *Spearman's rho* = 0.38, *p* < .01)				
Low	**4**	7	0	**36.4%**
Moderate	22	**20**	3	**44.4%**
High	0	4	**7**	**63.6%**
Percentage Agreement	**15.4%**	**64.5%**	**70.0%**	**46.3%**
Static-2002R (*N* = 68; *Spearman's rho* = 0.19, *ns*)				
Low	**8**	14	1	**34.8%**
Moderate	18	**13**	4	**37.1%**
High	1	4	**5**	**50.0%**
Percentage Agreement	**29.6%**	**41.9%**	**50.0%**	**38.2%**
SORAG (*N* = 15; *Spearman's rho* = 0.49, *ns*)				
Low	**2**	5	1	**25.0%**
Moderate	0	**4**	2	**66.7%**
High	0	0	**1**	**100%**
Percentage Agreement	**100%**	**44.4%**	**25.0%**	**46.7%**
SVR-20 (based on total score) (*N* = 56; *Spearman's rho* = 0.56, *p* < .001)				
Low	**21**	0	0	**100%**
Moderate	22	**3**	0	**12.0%**
High	3	5	**2**	**20.0%**
Percentage Agreement	**45.7%**	**37.5%**	**100%**	**46.4%**