Evaluating the inter-rater reliability of the Violence Risk Scale - Sexual Offense Version (VRS-SO) in a community-based treatment setting[1]

Michael J. Howell[1], Sarah M. Beggs Christofferson[1], Mark E. Olver[2]
[1]Psychology Department, University of Canterbury
[2]Department of Psychology, University of Saskatchewan

[Sexual Offender Treatment, Volume 12 (2017), Issue 1]

Abstract

Methods to assess risk of reoffending and guide rehabilitation efforts for offenders have been evolving over several decades, resulting in the development of a number of specialised tools. Evaluating the validity and reliability of these tools across contexts is imperative, in order to have confidence in the accuracy of risk assessments and the decisions they inform. One such tool is the Violence Risk Scale-Sexual Offense version (VRS-SO), designed to assess both risk of reoffending and change across treatment among sex offenders. Although validated on several incarcerated samples, to date the validity of this tool in community based settings has not been evaluated. This study reports on an exploration into VRS-SO inter-rater reliability, carried out as part of a wider community validation study. Clinician-scored pre-treatment VRS-SO ratings (n = 8) were gathered over an eight-month period, and each case was also rated independently by a researcher using file information. Rater consistency was analysed using the intraclass correlation coefficient (r_{ICC}). Unstructured interviews of clinicians were also undertaken to obtain valuable user perspectives. Our results indicate strong preliminary support for VRS-SO inter-rater reliability in a community treatment setting (e.g., $ICC_{c,2}$ = .98, $p < 0.001$for total scores), and add to the existing evidence base further given the use of prospectively collected clinical ratings. Item-level analyses suggest that modifications for the community setting and/or for clients whose offences were internet-based could further enhance reliability. The findings inform applications of the VRS-SO with community-based clients as well as potential future modifications for this context.

Keywords: Violence Risk Scale - Sexual Offense version; VRS-SO; sexual offenders, reliability; inter-rater reliability; risk; clinician; community; treatment

Introduction

Risk assessment is an important facet of the criminal justice system. It refers to evaluating the likelihood of a future event (in this context, future criminal offending) based on secondary indicator variables (Hanson, 2009). Risk assessments are widely used with respect to sentencing, parole, and monitoring recommendations, and to inform the level of treatment required for each offender in order to reduce his risk. Such decisions weigh heavily on all parties involved, and it is often difficult to balance the risks of the offender with the needs of the individual and community. Delivering an inappropriate or mismatched level of treatment to an offender (e.g., a high-intensity programme to a low-risk offender) can even have paradoxical effects, resulting in an increased risk of reoffending (Andrews & Dowden, 2006). This makes it vitally important that risk assessment tools measure as accurately as possible the level of risk posed by an individual. In the case of sex offenders, extreme care and accuracy is even more warranted when considering the impact of these kinds of crimes (e.g., see Putnam, 2003; Resick, 1993), but also given the elevated levels of attention sex offenders

receive compared to others, from both the government and the public (Levenson, 2009).

With these considerations in mind, it is important that risk assessment tools have been shown to be both reliable and valid. Reliability refers to consistency of measurement, and is a necessary precursor for validity (defined as the extent to which a tool is measuring what it is intended to measure; Kimberlin & Winterstein, 2008). This paper focuses on the reliability of a well-known sex offender risk assessment tool, the Violence Risk Scale Sexual Offense Version (VRS-SO; Wong, Olver, Nicholaichuk, & Gordon, 2003-2017), specifically on the inter-rater reliability of test scores.

Inter-rater Reliability

Inter-rater reliability refers to the degree of agreement or consensus among raters using a particular tool (McHugh, 2012). Good inter-rater reliability would mean that different raters, when observing the same stimuli, produce the same or similar ratings. If a tool has poor inter-rater reliability, this could indicate problems either with the tool, or in the raters' application of it (suggesting a training need). Inter-rater reliability can be measured in different ways, one of which is using intraclass correlations (ICC; Shrout & Fleiss, 1979). The ICC is a statistic that represents the relationship between variables of the same class (Salkind, 2010). ICC values can be interpreted as follows: 0-0.2 indicates poor agreement; 0.3-0.4 indicates fair agreement; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and > 0.8 indicates almost perfect agreement (Landis & Koch, 1977).

VRS-SO

The VRS-SO is a fourth generation risk assessment tool (see Campbell, French & Gendreau, 2009) developed for use within the sex offender population. It includes both static and dynamic risk subscales in addition to measuring risk change across interventions. The static subscale consists of seven items, based on criminal history information, while the dynamic subscale consists of 17 items measuring three broad factors: sexual deviance, criminality, and treatment responsivity. The VRS-SO was developed by incorporating two theoretical literature bases, the risk-needs-responsivity model for offender rehabilitation (Bonta & Andrews, 2007), and the Transtheoretical Model of Change (Prochaska, DiClemente, & Norcross, 1992). It is designed to be used in conjunction with therapy, allowing the clinician to identify items on the dynamic scale from which to draw treatment targets, while also allowing the clinician to indicate how far the offender has come by the end of treatment in terms of addressing these targets and changing them.

The VRS-SO is now well supported as a valid risk tool in terms of predictive accuracy in relation to recidivism (e.g., Beggs & Grace, 2010; Olver et al., 2007; Olver, Nicholaichuk, Kingston, & Wong, 2014). Some of the notable validation studies have also reported promising results for the inter-rater reliability of the VRS-SO. Olver et al. (2007) reported significant single measure intraclass correlation coefficients, for the pre-treatment dynamic item total ($ICC_{c,1}$ = .74) as well as for each factor (Sexual Deviance $ICC_{c,1}$ = .72; Criminality $ICC_{c,1}$ = .77; and Treatment Responsivity $ICC_{c,1}$ = .66). The inter-rater reliability of the stages of change component was also evaluated by correlating the dynamic item change scores between two raters (r = .68). Subsequently, Beggs and Grace (2010) reported significant intraclass correlation coefficients of $ICC_{c,1}$ = .90 between independent raters who completed pretreatment ratings on the dynamic items. In sum, these findings indicate an overall high standard of inter-rater reliability across a range of different treatment groups for the VRS-SO.

Current Study

The primary aim of the current study is to further assess the inter-rater reliability of the VRS-SO, with two important departures from previous research: 1) the current data are drawn from a community-based treatment setting (part of a wider community-based prospective validation of the VRS-SO); and 2) the current study makes use of prospective clinical VRS-SO ratings. The previous studies by Olver et al. (2007) and Beggs and Grace (2010) both involved retrospective VRS-SO ratings based on archival file information from a prison-based treatment setting. The VRS-SO has been validated previously using prospectively rated clinical data (Olver et al., 2014); however, inter-rater reliability was not evaluated in that study, with the authors noting that the nature of the study's paradigm, involving clinical ratings by the individual's particular therapist who would alone have the knowledge required to derive VRS-SO scores on the basis of their clinical assessment, "did not lend itself to obtaining interrater reliability estimates of the sample" (p. 323). Our study aims to overcome this challenge inherent to the prospective clinical paradigm, to provide an important evaluation of the measure's inter-rater reliability in a 'real world' setting closely matching its ordinary intended use (as opposed to a purely research setting). Our approach mirrors that used by both Olver et al. (2007), and Beggs and Grace (2010) in their inter-rater reliability analyses, except for adaptations necessary to accommodate the nuances of utilising this 'real world' data. A full description of our procedures can be found below. Further, in order to comprehensively consider the inter-rater reliability of the VRS-SO and with a view to providing concrete suggestions for future improvements, this study also includes a secondary component (Part 2), in which clinician user perspectives are drawn on to investigate hypotheses based on close analysis of Part 1 findings regarding specific VRS-SO items showing relatively weak inter-rater reliability. It is hoped that Part 2 findings may inform as to the potential usefulness of specific adaptations to the VRS-SO for the community setting.

Both parts of this study were approved by the Human Ethics Committee of the University of Canterbury.

Part 1 Method

The purpose of this primary part of our study is to investigate the inter-rater reliability of the VRS-SO in a community-based treatment setting, using prospectively collected clinically rated data. Out of a number of possible approaches to evaluating the inter-rater reliability of a measure (see Gwet, 2014), the best fit given the type of instrument the VRS-SO is (clinician-rated), as well as the real-world clinical setting, is a fully crossed design (Hallgren, 2012). In this design, all cases submitted to the dataset are rated by both of two raters (in this case, the clinician, and independently by a researcher). This allows each item for each case to be assessed and compared across both sets of raters.

Participants

Clinician-scored VRS-SO ratings were provided for eight cases for the purposes of this study, by the team of clinicians performing regular risk assessments on their clients at a community-based sexual offending treatment site (a non-government organization located in Christchurch, New Zealand). The eight cases were all male adults, ranging in age from 19 to 61 years ($M = 40.6$), and constituted the total number of cases the clinic had received over the course of the eight-month data collection phase of this study for which sufficient information was available to enable the independent VRS-SO research ratings to be carried out from file, as per the procedure outlined below. Four cases of the eight involved contact-based offending and the remaining four involved non-contact

internet-based offending (i.e., relating to images of child sexual abuse).

VRS-SO

The VRS-SO is a 24-item scale, consisting of a 7-item static subscale and a 17-item dynamic subscale, which measures three latent factors: Sexual Deviance; Criminality; and Treatment Responsivity. Although the VRS-SO was designed to be used at both pre- and posttreatment in order to measure change, the current study involved pretreatment scores only, due to clients either not having yet completed treatment at the time of the research, or being assessment-only referrals. Cases were rated on the VRS-SO by both their clinician and the research rater independently using standard VRS-SO scoring protocols and both sets of dynamic item scores, factor totals, and dynamic total scores were gathered for analysis.

Procedure

Data collection. Pretreatment VRS-SO clinical ratings were carried out by clinicians working at the community based treatment site as part of their regular client assessments. All clinicians who submitted client ratings had either successfully completed training in the VRS-SO by a certified VRS-SO trainer, or (in the case of new staff) performed ratings under the supervision of a trained senior colleague. In completing their ratings according to the VRS-SO scoring protocols, clinicians were informed by their clinical interview with the client, as well as documents on file such as referrals to the service, criminal conviction records, and court documents. Each case was then independently rated by the research rater (the second author, a trained and certified trainer on the VRS-SO, and experienced in using the measure for research and clinical purposes) based on all available and relevant file information, which included all the documents the clinician would have had access to, plus the case notes taken by the clinicians from their clinical interviews. To minimize any bias in scoring, both the clinicians and the research rater were blind to each other's scores when carrying out their ratings. This data collection procedure led to two sets of VRS-SO scores for each of the eight cases, one scored by the clinician rater and the other by the research rater. These scores were then entered into a database.

Planned analyses. Analyses were undertaken using SPSS software. Inter-rater reliability, our primary purpose for the study, was evaluated through computing ICC values (Shrout & Fleiss, 1979). The reliability analysis was performed for each of the 17 individual dynamic items on the VRS-SO, and for each factor and total score using a two-way mixed effects consistency model (ICCc,1). Only dynamic items were analysed in this assessment in accordance with the previous VRS-SO studies addressing inter-rater reliability (Beggs & Grace, 2010; Olver et al., 2007).

Part 1 Results

Sample Descriptives

Total scores on the VRS-SO dynamic subscale ranged from 10.6 to 39.3 (M = 22.1, SD = 8.1) for the research rater, and from 7.0 to 36.0, (M = 20.6; SD = 8.1) for the clinician raters. Total VRS-SO scores (the sum of both static and dynamic subscales) ranged from 11.1 to 52.2 (M = 27.8; SD = 11.6) for the research rater, and from 10.0 to 48.0 (M = 27.1; SD = 11.4) for the clinician raters. Both total score means fall within the "moderate-low" category based on original VRS-SO interpretive guidelines (Olver, Wong, Nicholaichuk, & Gordon, 2007). Based on updated risk categories and updated recidivism estimates (Olver et al., 2017) consistent with new common risk assessment language guidelines (Hanson et al., 2016), these means fall within the Level III "Average" category.

Inter-Rater Reliability

ICC values were computed to assess the inter-rater reliability of the 17 dynamic items (single measure), and the three factor totals, dynamic subscale total, and the overall total score (average measure) of the VRS-SO, between the two sets of ratings across all cases. ICC results are shown in Table 1.

Table 1: ICC values for VRS-SO dynamic item, factor, and total scores (pre-treatment)

VRS-SO Score Component	ICC	CI Lower	CI Upper
D1 Sexually deviant lifestyle	.44	-.32	.86
D2 Sexual compulsivity	.64*	-.05	.92
D3 Offense planning	.67*	.01	.93
D4 Criminal personality	.47	-.29	.86
D5 Cognitive distortions	.76**	.19	.95
D6 Interpersonal aggression	.95***	.75	.99
D7 Emotional control	.48	-.27	.87
D8 Insight	.59*	-.13	.90
D9 Substance abuse	.87**	.50	.97
D10 Community support	.14	-.58	.74
D11 Release to high risk situations	.18	-.53	.76
D12 Sexual offending cycle	.96***	.82	.99
D13 Impulsivity	.82**	.34	.96
D14 Compliance with community supervision	.78**	.23	.95
D15 Treatment compliance	-	-	-
D16 Deviant sexual preference	.56	-.17	.89
D17 Intimacy deficits	.73*	.13	.94
Factor 1 (Sexual Deviance)	.94***	.71	.99
Factor 2 (Criminality)	.95***	.73	.99
Factor 3 (Treatment Responsivity)	.76*	-.20	.95
Dynamic	.97***	.84	.99
Overall (Static + Dynamic)	.98***	.90	1.00

Note: * p < .05. ** p < .01. *** p < .001. Confidence intervals set at 95%.

As can be seen from Table 1, individual dynamic items produced ICCvalues ranging between .14 and .96. Of the 16 items that were able to be analysed in this dataset (all except D15), 10 were statistically significant. Of the items that were significant, ICC values were within acceptable ranges (.59 - .95) based on the Landis and Koch (1977) guidelines (items that were not significant are explored in further detail in Part 2 below). An ICC value was not calculable for D15 because there was no variance in the clinician ratings (this item is also discussed further in Part 2). Overall ICC values for the three factors, dynamic total, and overall total were all significant, with ICC values between .76 and .98. Notably, the overall total VRS-SO score showed near perfect agreement ($ICC_{c,2}$ = .98, p < 0.001; the corresponding single measure coefficient was $ICC_{c,2}$ = .96, p< 0.001). Thus, results of the current study provide excellent support for the inter-rater reliability of the VRS-SO in community-based clinical settings. Nonetheless, it is clear from the item-level ICCs shown in Table 1 that there remains room for improvement, with non-significant "fair" or even "weak" agreement on a number of items using the Landis and Koch (1977) interpretive guidelines. This therefore provided the rationale for the second part of the study, presented below.

Part 2 Method

Rationale

Part 2 of this study was designed to investigate more closely the relatively poor inter-rater reliability, in isolation, of particular VRS-SO items based on the Part 1 findings shown above in Table 1. Part 2 first involved the production of line graphs to compare the two independent sets of ratings at the item level, followed by interviews with community-based clinicians who had completed the item ratings to ascertain potential reasons for rating discrepancies. Because the purpose of this study was to inform potential future adaptations of the VRS-SO for use in community-based populations, feedback from clinicians currently using the measure in this context could potentially highlight specific issues within the measure or its implementation, which could then be addressed.

Participants

Participants were three clinicians at the community-based treatment site in Christchurch, New Zealand, who responded to an email inviting their participation in the current study. All clinician participants had either successfully completed training in the VRS-SO by a qualified instructor, or (in the case of a new staff member) performed ratings under the supervision of a trained senior colleague.

Procedure

From the results of Part 1, seven dynamic items in particular were noted as targets for further exploration: Sexually Deviant Lifestyle (D1), Criminal Personality (D4), Emotional Control (D7), Community Support (D10), Release to High Risk Situations (D11), Treatment Compliance (D15), and Deviant Sexual Preference (D16). Line graphs for clinician and researcher ratings on the aforementioned 7 dynamic items were generated to visually illustrate scoring discrepancies. Close inspection of these graphs along with the manual rating criteria for the items in question led to the generation of tentative hypotheses which were then investigated via unstructured interviews with clinicians using the measure in the community.

Following consent procedures an unstructured interview was undertaken either by telephone or in person with each of the clinician participants at a time and place convenient to them. Participating clinicians were approached and asked to provide input regarding their experiences of rating the

particular items that were noted for investigation. They were also then offered the opportunity to express any additional thoughts on the VRS-SO that were not captured by the initial questions. Notes based on responses to the interview were recorded during and immediately after the interviews. At the completion of all interviews, responses were evaluated in relation to the possible level of support for the explanations proposed for the rater discrepancies on low IRR items. Any other patterns or convergence of information that may be useful to the purposes of the study were also considered.

Part 2 Results

Line graphs for clinician and researcher ratings on the aforementioned 7 dynamic items and 8 cases are presented in Figure 1 (see item numeric listing in top right hand corner). The sample (by chance) included an even split of those who had been referred to the clinic due to offences involving unlawful sexual contact, and those who had committed internet-only offences. This categorization was therefore also considered when exploring the rating differences at the item-level. To facilitate these comparisons, on each of the figures presented below, the broken horizontal line separates these two groups, with participant case numbers 1 through 4 involving contact offending, and case numbers 5 through 8 involving internet-only offences. Preliminary visual inspection of the seven produced line graphs revealed noticeably similar patterns between two pairs of items, D1 and D16, and D10 and D11. These pairs are discussed together below; the remaining four items are discussed individually.

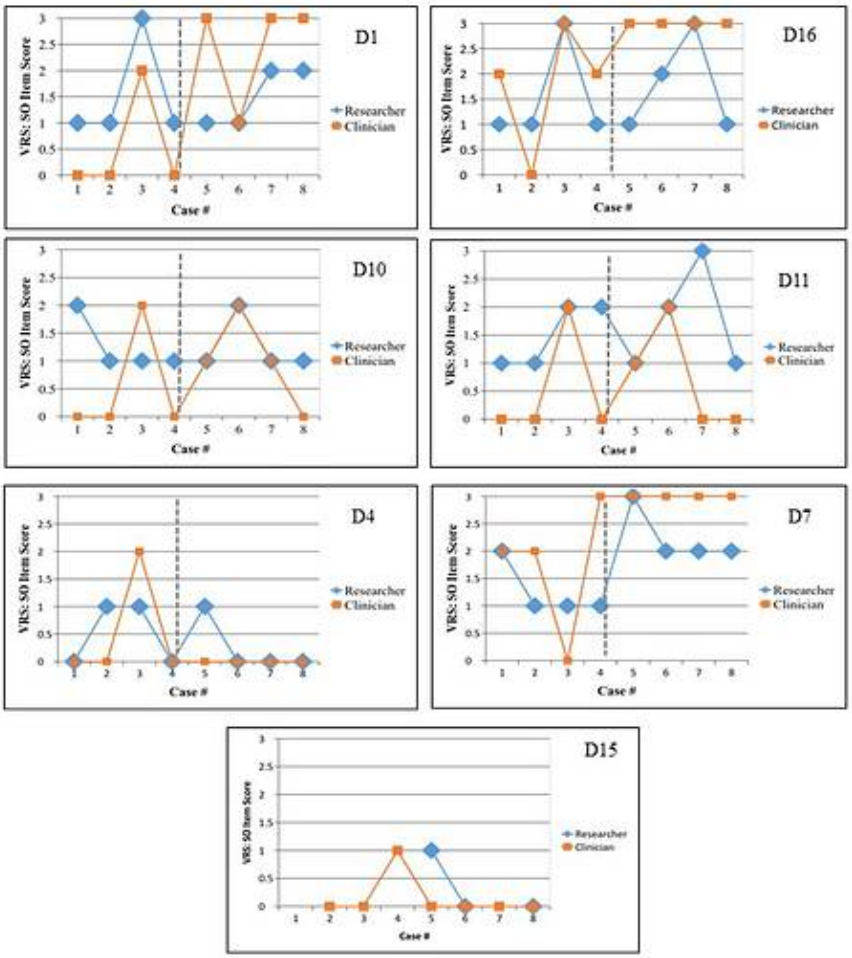

Figure 1: VRS-SO dynamic item (top right corner) scores with suboptimal IRR for each case

Sexual Deviance Items: D1 and D16

The top row of Figure 1 shows the VRS-SO scores for items D1 Sexually Deviant Lifestyle, and D16 Deviant Sexual Preference respectively, which both load onto the factor of Sexual Deviance. Both graphics show minor discrepancies in the first four cases (contact offenders) between the research rater and clinician rater; however, in cases 5 to 8 (internet-based offenders), the discrepancy was much more noticeable between the raters. Based on these observed patterns of discrepancy for these items, we anticipated that clinicians would report difficulties in applying the VRS-SO item criteria for D1 and D16 particularly when scoring the item on internet-based offenders.

Clinicians reported no general problems in assessing community-based clients on D1 (Sexually Deviant Lifestyle). Clinicians indicated that they felt the D1 criteria was clear enough to rate with community based offenders. In regards to internet-based offenders, clinicians stated that the D1 criteria could be improved by incorporating a clearer description of a sexually-deviant profile for an internet-based offender. However, the D16 (Deviant Sexual Preferences), item was viewed by clinicians as slightly more problematic to rate with their community-based clients. This was due to clients often having no previous criminal records or convictions, therefore the information supplied to rate D16 was often taken from self-reports, which can be unreliable. Therefore, D16 was considered an item that could use minor adaptations to better fit community-based clients.

Community Context Items: D10 and D11

The second row of Figure 1 depicts the VRS-SO scores of D10 (Community Support) and D11 (Release to High Risk Situations) respectively. These items, despite not being aligned on the same factor, both relate to community settings for clients. For D10, the clinician scores were generally lower than the researcher's, particularly for cases 1 to 4 (there was less discrepancy on this item for the cases involving internet-based offending). The figure shows that for D11 clinician scores were likewise generally lower, but across both categories of cases. The item criteria for both D10 and D11, consistent with the original design and intent of the measure, are written on the assumption that the client being rated is currently in prison and will be released in the future. Therefore, assessors whose clients are already in the community may find the criteria difficult to adapt for their clients. On this basis, we proposed that clinicians would report difficulties in adapting the VRS-SO manual criteria for items D10 and D11 for their community-based clients.

Clinicians reported no problems when assessing community-based clients on D10 (Community Support). When rating clients on D11 (Release to High Risk Situations), clinicians noted the difference between community-based clients that they are treating, and the prison-based clients that the VRS-SO was developed for. Specifically, prison-based clients make a conscious decision whether or not to return to a high risk environment, while community-based clients had not necessarily consciously chosen their situation, but rather were in it because it was their home. Therefore, clinicians indicated that an adaption that more suitably recognises the context of the community-based offender may be helpful in the future.

Criminal Personality Item: D4

The third row of Figure 1 (left hand side) shows the VRS-SO scores for D4 (Criminal Personality). The ICC for this item was insignificant within the current dataset, however when examining the scores closely it is evident that five out of the eight cases were in fact scored the same by both clinician and research raters. Scores that deviated between raters were within one point of each for the three cases. On this basis, we considered that the non-significant ICC score was attributable to the limited case numbers as opposed to actually poor inter-rater reliability, and thus did not pursue a follow-up discussion of ratings for this item with the clinicians.

Emotional Item: D7

Figure 1 right hand side third row shows D7 (Emotional Control) scores. There is a clear discrepancy between the raters, with clinician ratings tending to be higher than researcher ratings across almost all cases (however it is worth noting that in all but one case the discrepancy was only by a single point). This may have been due to ambiguity in the item criteria for D7, leading to raters scoring the item differently due to their own interpretations of the criteria. On this basis we anticipated that clinicians would report that the item criteria for D7 was more open to interpretation, causing difficulties in rating the item.

D7 (Emotional Control) was reported to be difficult to evaluate with community-based clients. Clinicians reported that clients are often reluctant or guarded about their emotions during the initial VRS-SO pretreatment scoring, and that relevant information becomes apparent after the client opens up over the course of treatment. Clinicians also reported that they often had to rely on supplemental material to rate the item, such as the client's relationship with significant others.

Treatment Item: D15

Finally, the bottom graphic of Figure 1 shows the VRS-SO item scores for D15 (Treatment Compliance). An ICC could not be calculated due to one rater (research rater) having omitted scoring the item for five of the eight cases. This discrepancy (of whether to omit or not) indicated again that there may be some problems in adapting the VRS-SO manual criteria for this item to community-based clients, most of whom may not have engaged in any prior relevant treatment at the time of assessment. Based on this finding, we hypothesized that clinicians may report having had to adapt the D15 item criteria to suit their clients, and it may be that such adaptations are not being made consistently between raters.

D15 (Treatment Compliance) was noted by clinicians to be difficult to rate for community-based clients. Clinicians reported that community-based offenders were highly likely to have not engaged in any previous treatment programs, therefore making D15 difficult to rate. However, clinicians did note that their original pre-treatment assessments appeared to be generally accurate when observing the client's progression in treatment. Clinicians also reported that this item was often omitted due to insufficient information.

Additional Patterns of Clinician Input

Clinicians noted that D14 (Compliance with Community Supervision) was difficult to rate for community-based clients. Clinicians stated that their clients were often not subjected to any forms of supervision (i.e., they were not on sentence, were not mandated for treatment, and may perhaps not have ever received a conviction for their online offending), thereby they would often rate clients as 0 on this item. Clinicians suggested that the item criteria for D14 could benefit from a community adaption to make it more applicable to their client base.

Although not in the scope of the current study, clinicians also noted difficulties in applying the static items to their community-based clients. In particular, it was often difficult or not possible to rate internet-based offenders on S3 (Sex Offender Type)[2], S5 (Unrelated Victims) and S6 (Number and Gender of Victims). This is because these offenders often do not have identifiable victims (due to primarily internet-based offences), while the VRS-SO manual criteria specifies having identifiable victims as part of the scoring criteria for these items. Community-based clients may also have no convictions, thereby making the static sub-scale unsuitable to rate in relation to them.

Overall, clinicians felt that the VRS-SO worked well overall with their clients in the community, and produced a good indication of the risk level for their clients. Clinicians noted that the factor components and stages of change model were particularly helpful in evaluating clients during the course of treatment. However, clinicians felt that some of the items mentioned could undergo some small adaptations to improve their sensitivity and fit to their community-based clients. Internet-offenders, in particular, could benefit from particular items adapted to fit their offender profile.

Discussion

This study examined the inter-rater reliability of the VRS-SO within a community-based treatment setting. The inter-rater reliability of the VRS-SO was assessed by comparing case ratings given by clinicians with ratings made independently on the same cases by an independent research rater. The overall results supported the inter-rater reliability of the VRS-SO in a community setting. ICC values for the factor, dynamic, and overall totals were all significant and reflected almost perfect

agreement, with the exception of Factor 3 which was in the strong agreement range. Because these totals are used to inform clinicians of the risk level of offenders, upon which several important decisions in the justice system are based, it is important that these produce significant and high ICC values. The results found in this inter-rater reliability study are consistent with previous reliability assessments; with ICC scores that are higher than those reported Beggs and Grace (2010) for the dynamic total, and higher ICC scores for all three factors, dynamic total, and overall total than reported in Olver et al. (2007).

Individual dynamic items were also assessed as part of this study, and produced mixed results. Ten of the 16 analysable dynamic items produced significant ICC values indicating strong agreement between the raters. It is important to note that individual dynamic items are not designed to be (and are not) used on their own to inform the risk levels of subjects or to base decisions on. Risk is well-known to be a multi-faceted construct (Bonta & Andrews, 2017), therefore the appropriate use of individual item scores is truly limited to being combined with other item scores to provide the overall totals and factor scores. Only at the total and/or factor score level can inferences be made regarding the risk level of subjects. Therefore, although six dynamic items did not produce significant results, in our view this does not reflect on the overall reliability of the tool (which, for this reason, is best judged using the results from the lower five rows of Table 1; i.e., the factor and total scores). However, in the interest of improving upon the current tool by informing possible adaptations for its use in a community-based treatment setting, all non-significant dynamic items were analysed with a high level of scrutiny in order to identify potential areas for improvement. Items D1, D4, D7, D10, D11, D15, and D16 were identified as items which may have caused some difficulties in adaptation among the raters, and all (excluding D4, which was attributed to problems with the limited sample size) were used as the basis for the hypothesized explanations explored in Part 2 of the study.

Before discussing Part 2 of the study, it is important to address the confidence intervals from Part 1, displayed in Table 1. Confidence intervals indicate a range of values from which the true value (in this case, the ICC coefficient) could lie within. There are three factors that can impact the width of a confidence interval: the sample size; the variability of the characteristic being studied; and the degree of confidence selected for the study (Cumming, 2013). Our study involved a small sample size, reflective of the real-world constraints of our design, involving prospectively-collected clinical data from an actual treatment site (can only analyse cases that come through within the research period). Therefore, it is likely that sample size was the reason for the wide confidence intervals observed for the majority of items in our study. The other factor to consider is that VRS-SO individual items are rated on only a 4-point scale of 0 to 3. Despite a one-point difference in scoring on a single item being clinically negligible, it would have a large impact on the single item reliability scores for the study. The combination of these two factors may have led to the large confidence intervals seen for many of the items in the Part 1 results.

In regards to Part 2 of the study, our focus first turned to items D1 and D16, which are both part of VRS-SO Factor 1 (Sexual Deviance). We proposed that, due to disparity between researcher and clinician ratings on these items, particularly with internet-based offenders, that clinicians would report difficulties regarding D1 and D16 in adapting the VRS-SO item criteria for use with internet-based offenders; however, D1 was not considered problematic by clinicians to rate their clients. Additionally, interviewed clinicians only reported slight problems in adapting D16 for use with community-based offenders as a whole. Difficulties with D16 were attributed to insufficient information and an over-reliance on self-reports. Therefore, our hypothesised explanation regarding D1 and D16 was not supported from the current data. However, clinicians did indicate problems in adapting D16 for community-based offenders (not specifically internet-based offenders), therefore suggesting that the item should be investigated for a potential item criteria adaption for

community-based clients in the future.

D10 and D11 are items that relate to the community context of offenders, and it was anticipated that clinicians may report a need to adapt these items for their community-based clients. Clinicians reported no problems in assessing D10; however, reported that they were adapting the manual criteria from D11 to suit their clients. Because the criteria describes clients who actively choose to be placed into a high risk situation upon release from prison, clinicians were instead adapting the criteria to fit their clients by referring to their current environment rather than a future one. However, it is crucial to note that clinicians did not report difficulties in rating the item, rather that they were already re-interpreting the item criteria. Despite the item not being reported as problematic for clinicians, D11 was still noted as an item with poor inter-rater reliability. Therefore, our proposed explanation that clinicians would need to adapt the D10 and D11 item criteria for community-based clients is partially supported.

D7 refers to a client's emotional problems which relate to sexual offending. We proposed that the item criteria for D7 may be too open to interpretation, causing difficulties when rating the item. Clinicians did report that this is a difficult item to assess with community-based clients, noting problems in obtaining reliable evidence when rating this item. Clinicians were often using self-report or second-hand reports to rate the item, which is not unusual by itself but does lend itself to be open to personal biases, such as clients minimizing their issues due to the severity of the crime, leaving clinicians to judge responses rather than following an objective criterion. Based on this, it appears that our hypothesised explanation is supported for D7, and that an adaption focusing on narrowing the item-criteria down may be beneficial for clinicians.

D15 relates to treatment compliance, and was anticipated to require clinicians to adapt the item for use with their community-based clients. Clinicians reported difficulties with this item, noting that it was challenging to rate community-based offenders who often did not have any prior experience with treatment. Clinicians also indicated that this item was often omitted due to insufficient evidence, which indicates that this item may have difficulties when being used in some community-based contexts. Based on these responses, our notion that clinicians had to adapt the D15 item criteria for use with community-based clients was moderately supported.

In Part 2 of this study, some of our hypothesised explanations for rater discrepancies were not supported, or only seemed to be partially supported, which leaves the question as to what other factors may explain the less than ideal reliability of these items. As previously mentioned, strong ICCs would not be expected when taking into account the small sample size and the fact that single items are rated on a four-point scale. Our current hypotheses focused on clinicians reporting that these item-criteria would be relatively difficult to interpret in the community context. However, when this was not reported to be the case, there is the possibility that, due to the degree of subjectivity in the item criteria, despite raters feeling subjectively confident in their interpretation and application of the criteria, others may be interpreting and applying the item criteria differently, which leads to poor reliability. If this is correct, then the item criteria would still see benefits from adaptation to improve objectivity, despite raters not reporting the item as problematic. An alternative possibility is that reliability could be improved by virtue of enhanced or ongoing training (Fernandez, Harris, Hanson, & Sparks, 2012).

The overall findings of the study support the inter-rater reliability of the VRS-SO within a community-based treatment setting. Despite the VRS-SO being developed on prison-based clients, the instrument shows good potential for adaption with community-based clients. The results from Part 1 show that the ICC values for the factor, dynamic, and overall totals are at the high end of acceptable levels, supporting the inter-rater reliability of the tool. Part 2 of the study suggests that

the scoring criteria for several items, D1, D7, D10, D11, D15, and D16, could be clarified or adapted for better use within a community-based setting.

A limitation of the current study presented is that the sample of cases used in the inter-rater reliability analysis is of relatively small size for this type of study. As a function of working with 'real-world' clinical data, which has its clear advantages in terms of ecological validity and generalizability of findings, the dataset provided was small and could not be expanded upon within the research period of this study. However, the small sample size did allow finer details, such as a case by case comparison of item scores between raters, to be used within the study. Therefore, although we acknowledge the limited sample size as a limitation of the study, it also provided the opportunity to qualitatively assess the data with a higher level of scrutiny, forming the basis for Part 2 of the study. The analysis in Part 2 was only a small step in the examination of specific VRS-SO items that could be usefully adapted, but provides significant information which will prove useful for future research investigating the adaption of the VRS-SO for use in community-based treatment settings.

References

1. Bonta, J., & Andrews, D. A. (2017). The Psychology of Criminal Conduct (6th ed.). New York, NY: Routledge.
2. Andrews, D. A., & Dowden, C. (2006). Risk Principle of Case Classification in Correctional Treatment: A Meta-Analytic Investigation. International Journal of Offender Therapy and Comparative Criminology, 50, 88-100. doi: 10.1177/0306624X05282556
3. Beggs, S. M., & Grace, R. C. (2010). Assessment of dynamic risk factors: An independent validation study of the violence risk scale: Sexual offender version. Sexual Abuse: Journal of Research and Treatment, 22, 234-251. doi: 10.1177/1079063210369014
4. Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. Rehabilitation, 6, 1-22.
5. Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. Criminal Justice and Behavior, 36, 567-590. doi: 10.1177/0093854809333610
6. Cumming, G. (2013). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge.
7. Fernandez, Y., Harris, A., Hanson, R., & Sparks, J. (2012). Stable-2007 Coding Manual Revised 2012. Scoring manual. Ottawa, Ontario: Public Safety Canada.
8. Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.
9. Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8, 23-34. doi:10.20982/tqmp.08.1.p023
10. Hanson, R. K., Bourgon, G., McGrath, R. J., Kroner, D., D'Amora, D. A., Thomas, S. S., & Tavarez, L. P. (2016). A five-level risk and needs system: Maximizing assessment results in corrections through the development of a common language. Justice Center Council of State Governments: Washington, DC.
11. Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. American Journal of Health System Pharmacy, 65, 2276-2284.
12. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.
13. Levenson, J. S. (2009). Sex offense recidivism, risk assessment, and the Adam Walsh Act. Sex Offender Law Report, 10, 1-6.

14. McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22, 276-282.
15. Olver, M. E., Mundt, J. C., Thornton, D., Beggs Christofferson, S. M., Kingston, D. A., Sowden, J. N., Nicholaichuk, T. P., Gordon, A., & Wong, S. C. P. (2017). Using the Violence Risk Scale-Sexual Offense Version in sexual violence risk assessments: Updated risk categories and recidivism estimates from a multisite sample of treated sexual offenders. Psychological Assessment. Epub ahead of print doi: [tbc].
16. Olver, M. E., Nicholaichuk, T. P., Kingston, D. A., & Wong, S. C. P. (2014). A multisite examination of sexual violence risk and therapeutic change. Journal of Consulting and Clinical Psychology, 82, 312-324. dx.doi.org/10.1037/a0035340
17. Olver, M. E., Wong, S. C. P., Nicholaichuk, T. P., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. Psychological Assessment, 19, 318-329. doi: 10.1037/1040-3590.19.3.318
18. Prochaska, J. O., DiClemente, C. C., & Norcross, J. C. (1992). In search of how people change: Applications to addictive behaviors. American Psychologist, 47, 1102.
19. Resick, P. A. (1993). The psychological impact of rape. Journal of Interpersonal Violence, 8, 223-255. doi:10.1177/088626093008002005
20. Salkind, N. J. (2010). Encyclopedia of research design. Thousand Oaks, CA: SAGE Publications Ltd. doi: 10.4135/9781412961288
21. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428. doi:10.1037/0033-2909.86.2.420
22. Wong, S. C. P., Olver, M. E., Nicholaichuk, T. P., & Gordon, A. (2003-2017). Violence Risk Scale: Sexual Offender Version (VRS-SO). Saskatoon, Saskatchewan, Canada: Regional Psychiatric Centre and University of Saskatchewan.

Footnotes

[1]This study was supported by departmental research funds from the Psychology Department, University of Canterbury.
[2]Item S3 has since been renamed, Sexual Offense Victim Profile.

Author address

Sarah Beggs Christofferson
Psychology Department
University of Canterbury
Private Bag 4800
Christchurch 8140, New Zealand
sarah.christofferson@canterbury.ac.nz