

Field Inter-Rater Reliability of the Psychopathy Checklist–Revised

International Journal of
Offender Therapy and
Comparative Criminology
2018, Vol. 62(2) 468–481
© The Author(s) 2016

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0306624X16652452
journals.sagepub.com/home/ijo



Ghena Ismail^{1,2} and Jan Looman^{1,3}

Abstract

Strong inter-rater reliability has been established for the Hare Psychopathy Checklist–Revised (PCL-R), specifically by examiners in research contexts. However, there is less support for inter-reliability in applied settings. This study examined archival data that included a sample of sex offenders ($n = 178$) who entered federal custody between 1992 and 1998. The offenders were assessed using the PCL-R on two occasions. The first assessment occurred at Millhaven Institution, the intake unit for federally incarcerated offenders in the province of Ontario. The second assessment took place upon inmates' transfer to the Regional Treatment Center, which admits federal inmates with intense psychological and psychiatric needs. Intra-class correlation coefficients (ICCs) were calculated for item, total, factor, and facet scores. The ICC absolute agreement for the PCL-R total and factor scores from raters across both settings was slightly better than what has been previously reported by Hare. Results of this study show that the reliability of PCL-R scores in field settings can be comparable to those in research settings. Authors conclude by highlighting the importance of training, consultation, considering different scores for a given item, following the guidelines of the manual in addition to considering measures that enhance neutrality and reliability of findings in the criminal justice system.

Keywords

psychopathy, reliability, PCL-R

¹Correctional Service Canada, Ottawa, Ontario, Canada

²American University of Beirut, Lebanon

³Providence Care, Kingston, Ontario, Canada

Corresponding Author:

Jan Looman, Psychologist, Forensic Unit, Providence Care Mental Health Services, 752 King Street West, Kingston, Ontario, Canada K7L 4X3.

Email: loomanj@providencecare.ca

Definition and Rationale of Study

The Hare Psychopathy Checklist–Revised (PCL-R; Hare, 1991, 2003) is a clinical rating scale that is widely used to assess the risk for violence and recidivism, in addition to measuring traits of psychopathy. This use is supported by an empirical base that has demonstrated a relationship between psychopathy and future criminal behaviour (Campbell, French, & Gendreau, 2009; Leistico, Salekin, DeCoster, & Rogers, 2008) as well as more specifically violent behaviour (Kennealy, Skeem, Walters, & Camp, 2010; Leistico et al., 2008; Yang, Wong, & Coid, 2010) and sexual recidivism (Hawes, Boccaccini, & Murrie, 2013).

Given its established reputation as a reliable and valid instrument as a method for assessing various types of risk (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Skeem, Polaschek, Patrick, & Lilienfeld, 2011), the PCL-R, initially developed for research purposes, has become increasingly used for forensic assessment purposes in recent years. Currently, the State of California requires a PCL-R based assessment for every inmate receiving a sentence of “life with parole.” Similarly, assessments of psychopathy are mandated in the civil commitment proceedings of Sexually Violent Predator (SVP) cases in Texas. This established use extends beyond the North American judicial system to a number of European countries as well (Cooke & Michie, 2010; Hildebrand, de Ruiter, de Vogel, & van der Wold, 2002).

The use of the PCL-R in a wide range of critical decisions, including bail, sentencing, institutional placement, parole as well as designations such as “dangerous offender” and “sexually violent predator,” constitutes a deviation from the intended use of the instrument. The author of the instrument, Robert Hare (2006), warned about ignoring the different implications of using the PCL-R in real-world situations, noting that the established reliability of the measure may not necessarily translate in individual cases. The question of investigating the field reliability of the PCL-R has gained increased attention over the past years. Many of the initial studies yielded encouraging results. An intra-class correlation coefficient (ICC) of .97 was reported in a small sample of 10 offenders admitted to Canadian federal custody and examined separately by two raters (Kroner & Mills, 2001). Another study that examined a sample of 21 male offenders randomly selected from a pool of 125 offenders admitted to the Canadian federal system for committing homicide also found high inter-rater agreement, with ICCs for Total, Factor 1, and Factor 2 scores of .92, .81, and .95, respectively (Porter, Woodworth, Earle, Drugge, & Boer, 2003). Similar findings were found in a study that involved sex offenders admitted to a prison-based treatment program, the Warkworth Sexual Behaviour Clinic (WSBC; Barbaree, Seto, Langton, & Peacock, 2001). Barbaree and colleagues (2001) examined a sample of 47 sex offenders who were assessed twice—once by the WSBC team and another time at a penitentiary placement facility. Inter-rater reliability yielded a correlation coefficient of .81. It is important to keep in mind that initial examinations of the reliability of the PCL-R were examined using very small sample sizes, reducing variance and possibility of error. This may have contributed to the high ICCs reported.

More recent studies, which sought larger samples or examined different populations, suggest a more nuanced picture. Some studies pointed to a possible relationship

between partisanship and the scoring of PCL-R items. In a sample of 23 sex offenders evaluated for civil commitment in Texas, the agreement between opposing evaluators (prosecution vs. defence-retained) for PCL-R total scores was .39, significantly lower than ICC values reported in the PCL-R manual (Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009). These studies have been criticized on a number of grounds, one of which pertained to the generalizability of their findings (Boccaccini, Turner, & Murrie, 2008). To address the question of generalizability, Edens, Cox, Smith, DeMatteo, and Sörman (2015) investigated forensic examiners in a large sample of Canadian cases that included different types of offenders. Poor inter-rater reliability was detected in cases of sexual offenses and non-sexual offenses alike even though inter-rater reliability was somewhat lower for non-sexual offenses. Interestingly, and in contrast to previous studies, inter-rater reliability was poor not only when comparing raters on opposing sides but also when comparing raters on the same side. Similar results were noted in other studies. For example, C. S. Miller, Kimonis, Otto, Kline, and Wasserman (2012) examined PCL-R scores given by mental health professionals hired by the Department of Children and Families (DCF), an independent body not affiliated with the courts. The evaluations were provided for SVP pertinent proceedings taking place in Florida. Overall reliability of the PCL-R (ICCA1) was .60 and the PCL-R (ICCA2) was .75. In addition, a naturalistic test-retest design among a small sample of 27 life sentenced prisoners, not specific to sex offenders, in Sweden found similar results. Prisoners were repeatedly evaluated by mental health professionals retained by an independent government authority as part of their application to receive a shortened prison term. The overall reliability of the PCL-R (ICCA1) was .70 for the total score and .62 and .76 for Factor 1 and 2 scores, respectively (Sturup et al., 2014). As such, the inter-reliability in the studies by C. S. Miller and colleagues (2012) and Sturup and colleagues (2014) appeared to be weaker than what is reported in research-based studies, but better than what is shown in some other field studies.

While Edens et al. (2015) and Sturup et al. (2014), on the one hand, sought to address the question of generalizability of SVP Texas studies to other contexts, Murrie, Boccaccini, Guarnera, and Rufino (2013) on the other hand, sought to address the role that adversarial allegiance plays in contributing to potential biases in ratings. They conducted an experiment in which 108 forensic psychologists and psychiatrists were hired to review the same offender case files. Using deception, they communicated to one group of assessors that they (the assessors) were consulting for the defence, while communicating to the other group that they were consulting for the prosecution. Murrie and colleagues (2013) simulated real-life pertinent circumstances whereby each assessor met for 15 min with the supposedly hiring party to be exposed to typical biases. Confederates pretending to be prosecutors emphasized the importance of helping the court understand that the offenders brought to trial were “a select group whom the data show are more likely than other sex offenders to reoffend.” In their turn, those who pretended to be defence attorneys emphasized the importance of helping the court appreciate the data that show that not every sex offender poses a high risk to reoffend. As may be expected, findings showed that evaluators who believed they worked for the state tended to give higher scores than those who believed they worked for the

defence. Murrie et al. (2013) convincingly argued that results of their study represented strong evidence of allegiance, especially considering that real-life situations exhibit greater biasing forces than was simulated in the experimental manipulation. For instance, interactions between evaluators and their recruiting parties—be they prosecutors or attorneys—taking place in actual forensic settings are likely to significantly exceed the 15-min contact designed in the study.

Could inter-rater differences be at least partially explained by factors other than allegiance? Boccaccini and colleagues (2008) sought to answer this question by examining individual scoring tendencies regardless of the party hiring them. They drew upon a database of PCL-R total scores for 321 offenders, who had both a PCL-R score and an identifiable assessor, and who had been referred to the Special Prosecution Unit (SPU), the legal office responsible for pursuing commitment for high-risk offenders. Boccaccini and colleagues (2008) concluded that at least 30% of the variability in PCL-R ratings was attributable to differences among assessors, noting that some consistently tended to give higher or lower scores. The role of personality traits and idiosyncratic beliefs in influencing evaluators' ratings is beginning to receive increased attention (A. K. Miller, Rufino, Boccaccini, Jackson, & Murrie, 2011), with some scholars suggesting that they may have a greater impact on PCL-R ratings than allegiance or any other single factor (Rufino, Boccaccini, Hawes, & Murrie, 2012). Rufino and colleagues (2012) examined opposing evaluator scores for the 35 offenders referred to in Murrie et al. (2009) and nine additional offenders whose cases were pursued for commitment at a later date. To determine the source of bias or error, they introduced a set of independent evaluators against which they compared ratings given by the prosecution and defence. They found that 28% of the variability in the PCL-R scores in Texas SVP cases may be attributable to allegiance, and 34% of inter-rater differences were a function of individual scoring biases.

Another possible explanation for the differences in ICCs found across studies may be due to methodological variations in both training methods and analysis used to obtain ICCs. A quick examination of the surveyed studies on the PCL-R shows that the type of training that raters undergo is often not mentioned or clarified. It is possible that variations in training may be producing differences in rater qualifications, a potential bias. Moreover, the type of ICC used to assess reliability, whether absolute agreement or consistency, is often not reported in studies. Given that the two measures of ICC do not assess the same thing, this may be a contributing factor to the discrepancies in ICCs reported (Boccaccini et al., 2008).

It is needless to say that unless the issue of inter-rater reliability in the field is addressed, we are faced with serious concerns about the legitimacy of using the PCL-R outside research settings. A key question which this study aims to address is the following: Are the scoring-related problems cited in field studies inherent in the PCL-R itself, or are they possibly related to factors external to the measure? The present study is intended to build upon the growing body of literature that examines the inter-reliability of the PCL-R measure in the field. It also aims to address factors that may impact inter-rater reliability. Scores given by assessors across two different settings within the Canadian federal prison system are compared, followed by an analysis of

total, factor, and item scores. Results of this study are then compared with the results documented by Hare (2003) and recent research.

Method

Archival Offender Data

At the time of this study, all sentenced offenders were initially admitted to the Millhaven Assessment Unit (MAU) located within the Canadian province of Ontario. All admitted offenders were subject to an intake assessment. For sex offenders this included completing a comprehensive Specialized Sexual Offender Assessment (SSOA) and administering the PCL-R. Sex offenders assessed as high risk and with high treatment needs were referred to the High Intensity Sexual Offender Treatment Program at the Regional Treatment Center (RTC) to receive treatment for their inappropriate sexual behaviour. Administration of the PCL-R was a routine part of the pre-treatment assessment at that site. Therefore, each sex offender in the sample was assessed twice—initially by a clinician at MAU and subsequently by another clinician at RTC. Intake assessments conducted at MAU serve a dual purpose of triaging for treatment intensity and providing a risk assessment that could be considered in parole-related decisions and other security-related matters such as placement. The purpose of the comprehensive SSOA completed at the RTC was to determine individual programming needs. Assessors at both institutions were trained by the Dark Stone Research Group prior to conducting risk assessments. Assessors at the MAU had an undergraduate degree in psychology and were supervised by graduate-level psychologists, while assessors at RTC were graduate-level psychologists. The staff at both sites worked independently of each other.

Sample and Procedure

The Ontario's Regional Research Committee granted approval to access the archival records of those offenders who were examined in this study. Informed consent was obtained from participants; they understood that the data gathered during their assessment/treatment may be used for research purposes and they had the opportunity to have their data excluded should they choose.

The sample included 178 male sex offenders who participated in the specialized sexual offender assessment twice, as noted above. This included men whose current conviction was a sexual or sexually related offense (e.g., murder or assault conviction with a sexual component), or who have a previous sentence for a sexual offense and who were referred for an assessment based on this history. The MAU assessments were conducted between 1992 and 1998, while the RTC assessments were conducted between 1993 and 2006. In terms of demographics, the offenders were 33.3 years of age on average ($SD = 9.5$ years) at the time of assessment at MAU. About 81% of the sample was Caucasian, 9% aboriginal, and 9% were Black. The average number of previous convictions for sexual offenses was 4.1 ($SD = 4.1$), the average number of

previous convictions for violent non-sexual offenses was 1.7 ($SD = 2.2$), and the average number of previous convictions for general offending was 13.2 ($SD = 13.3$). The average time between administrations was 35.6 months ($SD = 28.98$), with a median of 26 months and a range from 4 to 138 months.

Each assessor¹ completed a file review and a separate semi-structured interview as part of a specialized assessment. Scoring was typically done in consultation with another clinician who was familiar with the respective offender. In all cases, scoring was completed according to guidelines which appear in the manual (Hare, 1991, 2003) and with reference made to the item descriptions in the manual at the time of scoring.

Measures

PCL-R. The PCL-R (Hare, 1991, 2003) is currently the most empirically validated instrument for assessing psychopathy in correctional and forensic psychiatric populations (DeMatteo & Edens, 2006). It is a 20-item measure of interpersonal, affective, and behavioural traits that clinicians score on the basis of an offender's records and a clinical interview. PCL-R items are rated on a scale from 0 to 2, with 0 = absent, 1 = maybe/in some respect, and 2 = present. The total score is the sum of all 20 items. The maximum score that can be obtained is 40, with 30 being the cut-off score for meeting the criteria for psychopathy (Hare, 2003).

Scoring the PCL-R typically yields three scores: the total PCL-R score, Factor 1 score, and Factor 2 score. Factor 1 is traits based and covers characteristics such as glibness/superficial charm, grandiose sense of self-worth, pathological lying, lack of remorse or guilt, callousness/lack of empathy, and conning/manipulative. Factor 2, on the other hand, is behaviourally based and covers items such as need for stimulation/proneness to boredom, parasitic life style, poor behavioural controls, early behavioural problems, impulsivity, irresponsibility, juvenile delinquency, and versatility of crime.

Results

As noted above, each inmate in our sample was assessed by two evaluators, each of whom worked in a different correctional setting within Correctional Service Canada. The total, factor, and facet scores for each of the raters are displayed in Table 1. For the MAU scores the mean was 24.3 ($SD = 7.3$) with a range from 5 to 38. For the RTC the mean was 23.6 ($SD = 7.5$) with a range from 5 to 36. The average difference between the PCL-R ratings for the 179 inmates was 0.53 points ($SD = 3.3$) with 137 scores (78.3%) falling within the ± 3 point range. Differences ranged from 0 to 15 points. The relationship between time separating administrations and the difference between scores was not significant ($r = .15, p = .44$).

Rater Agreement

Rater agreement was assessed via ICCs, a measure of the proportion of variance that is attributable to the target of the measurement (McGraw & Wong, 1996). One

Table 1. Average PCL-R/Factor Scores for MAU and RTC Scorings ($N = 178$).

	MAU scoring	RTC scoring
PCL-R total	24.3	23.6*
Factor 1	9.5	9.3
Factor 2	12.6	11.0**
Facet 1	4.1	3.8*
Facet 2	5.4	5.4
Facet 3	6.2	5.9*
Facet 4	6.3	6.3

Note. PCL-R = Psychopathy Checklist-Revised; MAU = Millhaven Assessment Unit; RTC = Regional Treatment Center.

* $p < .05$. ** $p < .000$.

relevant issue to consider when using the ICC is whether to use a consistency formula (ICC2) or an absolute agreement formula (ICC1; see McGraw & Wong, 1996). The consistency formula assesses the extent to which the measurement ranks the subjects consistently (i.e., agreement on who is assigned a higher score and who is assigned a lower score), compared with the absolute agreement formula which assesses the extent to which the scores are the same. For purposes of evaluating agreement on the PCL-R scoring, the absolute coefficient (ICC1) is appropriate given that the scores have a meaning associated with them. Typically, coefficients below 0.20 indicate “poor” agreement; values from 0.21 to 0.40 indicate “fair” agreement; values from 0.41 to 0.60 indicate “moderate” agreement; values from 0.61 to 0.80 indicate “substantial” agreement; whereas values from 0.81 to 1.00 indicate “almost perfect” agreement (Landis & Koch, 1977). All analyses were conducted using SPSS version 21.

Inter-rater reliability for PCL-R total, factor, and facet scores, as well as for individual PCL-R items was estimated by means of the ICC (see Table 2). The single measure ICC for the PCL-R total score was .90; for Factor 1, it was .78; and for Factor 2, it was .90. These values are slightly higher than the inter-rater reliabilities reported by Hare (2003). The inter-rater reliabilities of PCL-R Factor 1 items were significantly lower than reliabilities of Factor 2 items, $z = 3.92$, $p = .0001$. This pattern is consistent with the results reported by Hare (2003) and in current literature (Edens, Boccaccini, & Johnson, 2010).

Table 3 lists inter-rater reliabilities for all PCL-R items. At the level of individual PCL-R items, in general, ICCs were good to excellent. The only item that had an inter-rater reliability less than .60 was pathological lying. The highest single measure ICC was for Item 20 (criminal versatility) followed by Item 18 (juvenile delinquency). This is consistent with Hare’s (2003) results, and it is not surprising given the quantitative nature of these items. Similarly, Items 4 (pathological lying) and 16 (failure to accept responsibility) were the items with the lowest reliabilities, also consistent with Hare’s (2003) results. It is relevant to note here, that our inter-reliability ratings on these latter two items were nonetheless higher than the ratings reported in the PCL-R manual.

Table 2. ICC Values for Total, Factor, and Facet Scores.

	Hare (2003)	Current sample
PCL-R total	0.87	0.90
Factor 1	0.75	0.78
Factor 2	0.85	0.90
Facet 1		0.76
		0.79
		0.82
		0.94

Note. ICC = Intra-Class Correlation coefficient; PCL-R = Psychopathy Checklist–Revised.

As noted above, the time between ratings ranged from 4 to 136 months. Three separate groups were formed based on time between administrations. Those whose second assessment occurred within 19 months of the first made up one group ($n = 60$), those whose second assessment was between 20 and 35 months of the first made up the second group ($n = 55$), and those whose second assessment was more than 35 months after the first ($n = 60$) made up the third group. Comparing the ICCs for the first and third groups indicated no differences in strength of inter-rater reliability .895 versus .934 respectively, $z = 1.26, p = .21$ (two-tailed), between groups.

High Scores and Rater Agreement

In this study, and as shown in Table 4, we examined inter-rater agreement for higher scores, by assessing the inter-rater reliability of offenders who scored 25 or higher on the PCL-R, in accordance with Edens et al. (2010). Scores assigned at the MAU were used for selection. Ninety-six offenders were included in these analyses. The inter-rater reliability for this group was $ICC = .63$, in comparison with .90 for the overall sample. The same trend applied to Factor scores; with Factor 1 and Factor 2, $ICC = .63$ and .54 respectively, for those scoring 25 and above in comparison with .78 and .90 for the overall sample.

As noted earlier, recent literature indicates overall lower inter-rater reliability for the PCL-R total as well as factor scores for examinees with extreme scores. Edens et al. (2010) found that adjusting for the possible effect of range restriction among the high-scoring offenders resulted in scores of the Total PCL-R and Factor 2 that are more consistent with published research. However, inter-rater agreement on Factor 1 scores continued to be marginal ($ICC = .16$), even after the adjustment.

The role of individual differences? This study sheds light on the role of individual differences in raters, and its potential effect on the reliability of the PCL-R; an issue that unfortunately was not thoroughly investigated. Examining the notes made on the score sheet, where available, suggested that the variation was a function of failure to follow

Table 3. ICC Values for PCL-R Items.

	Hare (2003) ICC values	Current sample
Item 1	.50	.75
Item 2	.53	.76
Item 3	.60	.76
Item 4	.47	.59
Item 5	.59	.66
Item 6	.57	.69
Item 7	.52	.68
Item 8	.55	.62
Item 9	.56	.80
Item 10	.57	.75
Item 11	.62	.78
Item 12	.61	.82
Item 13	.56	.67
Item 14	.49	.70
Item 15	.50	.70
Item 16	.41	.63
Item 17	.68	.74
Item 18	.74	.86
Item 19	.70	.75
Item 20	.82	.90

Note. ICC = intra-class correlation coefficient; PCL-R = Psychopathy Checklist–Revised.

Table 4. ICC Values for Total, Factor, and Facet Scores for Those Who Score 25+.

	Edens, Boccaccini, and Johnson (2010)	Current sample
PCL-R total	0.420	0.63
Factor 1	0.157	0.54
Factor 2	0.557	0.72
Facet 1		0.62
Facet 2		0.63
Facet 3		0.74
Facet 4		0.86

Note. ICC = intra-class correlation coefficient; PCL-R = Psychopathy Checklist–Revised.

the scoring guide, despite the promotion of this practice. For instance, one examinee received a score of 2 on Item 8 (callous/lack of empathy) based merely on the nature of his offenses and the fact that he cheated on his wife. No other information was offered to justify that score.

Discussion

The current study investigated the inter-rater reliability of the PCL-R in a non-adversarial field setting. The results demonstrated that the PCL-R can be reliably scored in an applied context with properly trained and supervised raters. These results were obtained with a median period between administrations of over 2 years, and with scoring procedures that can be considered “best practice.” That is, all of the raters received training by the Dark Stone Research Group; scoring was done with consideration of the scoring criteria in the manual, to protect against drift, and in consultation with other experienced raters. Specifically, the RTC ratings were typically completed in consultation with another clinician who knew the offender after his admission to RTC. The other clinician functioned to “ground” the rater in the coding rules, asking for evidence in favor of or opposed to the presence of the trait.

While this study indicated that high levels of inter-rater reliability of the PCL-R are possible, it nonetheless highlighted a tendency even among well trained examiners to deviate from the scoring guide. For example, an offender who was arrested when he was in his mid-40s for molesting several female children was given a rating of 2 on pathological lying, based on the fact that he had lied about his offending to evade detection. When the examiner was challenged about her score, she indicated that the profile provided in the PCL-R manual was “just an example” of how to score the item. Clearly the examiner did not fully appreciate the function of prototypes, and as such, deviated from the PCL-R guidelines. Hare (2003) clearly states that each of the PCL-R items is rated “based on the degree to which the personality/behaviour of the inmate or patient matches the description of the item in the manual.” It is, also, often pointed out in PCL-R training workshops that descriptions outlined in the manual are not mere examples but rather prototypes or working templates against which examiners should consider their ratings. This point is of special significance given that during direct or cross-examination, PCL-R raters may be asked to specifically comment on the comparability of the examined offender to the normative group (Gacono & Hutton, 1994). Assessors should also review the evidence for and against a high score on each item. This could help control a common tendency to search for data that confirms one’s judgment (Douglas & Ogloff, 2003). As a last resort, evaluators are advised to consider omission. The manual clearly indicates that omission may be considered in the case of insufficient data, and also in the case of having completely divergent information with no means for determining which source of data may be more credible (Hare, 2003). This should safeguard against using some popular preconceived definition and/or possibly one’s emotional reaction as a default system to compensate for the lack of data or the presence of contradicting sets of data.

Seeking consultation with colleagues is another important measure to consider as it could help assessors justify their ratings particularly on those items which appear to require more subjective considerations. Developing a routine for discussing interpretation of items with a colleague may be an effective tool for insuring integrity of scoring. It may also be a good practice for more experienced raters, within a given institution, to mentor less experienced raters, with the purpose of ensuring use of

manual and adherence to guidelines. We consider the RTC approach to scoring as best practice in this regard, and the high reliabilities in the current sample attest to this.

Finally, with the rapidly evolving field of risk assessment and the continuous introduction of new tests and new revisions of existing instruments, it is necessary for forensic evaluators to regularly update their knowledge regarding what is utilized in the field. Assessors using the PCL-R or any other risk assessment measure will need to demonstrate an accurate and objective understanding not only of what a given construct or instrument measures, but also of what some constructs possibly fail to measure. Falling short of doing so may constitute a serious ethical breach in a forensic milieu where psychopathy testimony might outweigh other relevant mental health evidence.

Whereas training, consultation, and conscientious practice are important steps which are likely to address some of the issues discussed above, they should be supplemented with measures at the level of policy. Evidence for the reliability and validity of the PCL-R does not necessarily mean that a single assessment will be reliable or valid. While this may be of little consequence in research settings, the implications may be serious and critical in real-world situations (Edens, Colwell, Desforges, & Fernandez, 2005; Gacono & Hutton, 1994). As such, it is incumbent that the criminal justice system takes every measure possible to improve the reliability of the PCL-R and to safeguard against potential misuse due to poor qualifications and/or partisanship effects. The justice system may benefit from some form of credentialing or mandated training procedures, especially in jurisdictions where the PCL-R is routinely used. This is, however, easier to recommend than it is to implement, especially in court cases in which opposing parties typically retain their own expert witnesses (Edens et al., 2015). Perhaps courts should consider retaining independent organizations to conduct risk assessments. At a minimum, implementing quality control procedures such as external review may contribute to reduced inconsistencies in the PCL-R ratings.

Limitations

This study has a number of limitations. First, it is restricted to a sample of sex offenders. It is not known to what extent similar results would be found with a more diverse group of offenders. While it may be argued that the population of sex offenders may pose unique challenges in terms of reliability of scoring, this has not been shown to be the case in the literature.

Second, as noted above, the raters were not purposefully blinded to the results of each other's assessments. This means that in at least some cases, the RTC rater would have been aware of the score assigned by the MAU rater. This may have led to our data overestimating the reliability of the PCL-R scores due to possible anchoring effects and confirmatory biases. Although we have no way of ascertaining in how many cases the RTC raters may have been aware of the previous rating, we should note that the clinicians at the RTC typically made a point of completing their scoring independently of the previous scoring. As such, and, because there was no established practice in reviewing previous archived data on transferred inmates, blinding was not deemed necessary nor in line

with the naturalistic design of this study. It is also relevant to point out that in some of the studies discussed in the Introduction, different raters had access to other raters' scores and, even in such cases, the inter-rater reliability has often been found to be low (e.g., Edens et al., 2010). Relatedly, recent findings suggest that assessors have little difficulty ignoring previous scores when conducting their own assessments (Edens et al., 2015). Third, and more relevant to our study, is the relationship between the amount of available information and the ratings (Alterman, Cacciola, & Rutherford, 1993). By the time the offenders transitioned from MAU to the RTC, more data were likely available, implying a possibly different context for interpreting certain traits and behaviours. For example, some offenders may have participated in correctional programming (e.g., substance abuse programming) after the MAU rating, but prior to arriving at the RTC, and information about their performance in these programs would have been available. In addition, they may have worked within the institution, been involved in substance abuse or fights, and so on. The availability of this additional information may have impacted the scoring of all items, including historical ones (e.g. criminal history).

Finally, the design of the study did not allow for investigating an important issue, namely, variables that may explain systematic even though slight variation noted between scores given by raters at MAU and those at RTC. It would have been interesting to examine the presence of consistent group or individual differences among RTC and MAU assessors. Investigating whether some raters consistently offered higher or lower scores across all the offenders, or perhaps across specific items, could offer valuable insights. Literature points to a number of individual and contextual factors that may impact scoring. Boccaccini et al. (2008) suggested that "some evaluators may assign higher PCL-R scores across all of the offenders they evaluate." Edens et al. (2010; p.265) pointed out that different evaluators, due to different interpersonal styles, may evoke different responses from offenders (Edens et al., 2010). While this question was not central to this study, it is worthwhile to study as it may shed light on issues to consider when training future clinicians.

Declaration of Conflicting Interests

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

Funding

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

Note

1. At the time the scoring of the Hare Psychopathy Checklist-Revised (PCL-R) was completed, the Millhaven Assessment Unit (MAU) had a team of four assessors supervised by a psychologist, whereas the Regional Treatment Center had five psychologists who completed assessments. Over the years in which these ratings were made, there was turnover in staff at both sites, so in total at least 15 assessors would have provided PCL-R scores used in the analyses.

References

- Alterman, A. I., Cacciola, J. S., & Rutherford, M. J. (1993). Reliability of the Revised Psychopathy Checklist in substance abuse patients. *Psychological Assessment, 5*, 442-448.
- Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84-94.
- Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior, 28*, 490-521.
- Boccaccini, M. T., Turner, D. B., & Murrrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14*, 262-283.
- 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.
- 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.
- DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist-Revised in court: A case law survey of U.S. courts (1991-2004). *Psychology, Public Policy, and Law, 12*, 214-241.
- Douglas, K. S., & Ogloff, J. R. P. (2003). The impact of confidence on the accuracy of structured professional and actuarial violence risk judgments in a sample of forensic psychiatric patients. *Law and Human Behavior, 27*, 573-587.
- Edens, J. F., Boccaccini, M. T., & Johnson, D. W. (2010). Inter-rater reliability of the PCL-R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioural features? *Behavioral Sciences & the Law, 28*, 106-119.
- Edens, J. F., Colwell, L. H., Desforges, D. M., & Fernandez, K. (2005). The impact of mental health evidence on support for capital punishment: Are defendants labeled psychopathic considered more deserving of death? *Behavioral Sciences & the Law, 23*, 603-625.
- Edens, J. F., Cox, J., Smith, S. T., DeMatteo, D., & Sörman, K. (2015). How reliable are Psychopathy Checklist-Revised scores in Canadian criminal trials? A case law review. *Psychological Assessment, 27*, 447-456.
- Gacono, C. B., & Hutton, H. E. (1994). Suggestions for the clinical and forensic use of the Hare Psychopathy Checklist-Revised (PCL-R). *International Journal of Law and Psychiatry, 17*, 303-317.
- Hare, R. (1991). *The Hare Psychopathy Checklist-Revised Manual*. Toronto, Ontario, Canada: Multihealth Systems.
- Hare, R. (2003). *Hare Psychopathy Checklist-Revised (PCL-R): Technical manual* (2nd ed.) Toronto, Ontario, Canada: Multi-Health Systems.
- Hare, R. (2006). Psychopathy: A clinical and forensic overview. *Psychiatric Clinics of North America, 29*, 709-724.
- Hawes, S. W., Boccaccini, M. T., & Murrrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist-Revised. *Psychological Assessment, 25*, 233-243.
- Hildebrand, M., de Ruiter, C., de Vogel, V., & van der Wolf, P. (2002). Reliability and factor structure of the Dutch language version of Hare's Psychopathy Checklist-Revised. *International Journal of Forensic Mental Health, 1*, 139-154.

- Kennealy, P. J., Skeem, J. L., Walters, G. D., & Camp, J. (2010). Do core interpersonal and affective traits of PCL-R psychopathy interact with antisocial behavior and disinhibition to predict violence? *Psychological Assessment, 22*, 569-580.
- Kroner, D. G., & Mills, J. F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior, 28*, 471-489.
- Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.
- Leistico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28-45. doi:10.1007/s10979-007-9096-6
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. doi:10.1037//1082-989X.1.1.30
- Miller, A. K., Rufino, K. A., Boccaccini, M. T., Jackson, R. L., & Murrie, D. C. (2011). On individual differences in person perception: Raters' personality traits relate to their Psychopathy Checklist-Revised scoring tendencies. *Assessment, 18*, 253-260.
- Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24*, 944-953. doi:10.1037/a0028411
- Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24*, 1889-1897.
- Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis) agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior, 32*, 352-362.
- Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*, 19-53.
- Porter, S., Woodworth, M., Earle, J., Drugge, J., & Boer, D. (2003). Characteristics of sexual homicides committed by psychopathic and nonpsychopathic offenders. *Law and Human Behavior, 27*, 459-470.
- Rufino, K. A., Boccaccini, M. T., Hawes, S. W., & Murrie, D. C. (2012). When experts disagreed, who was correct? A comparison of PCL-R scores from independent raters and opposing forensic experts. *Law and Human Behavior, 36*, 527-537. doi:10.1037/h0093988
- Skeem, J. L., Polaschek, D. L. L., Patrick, C. J., & Lilienfeld, S. O. (2011). Psychopathic personality: Bridging the gap between scientific evidence and public policy. *Psychological Science in the Public Interest, 12*, 95-162.
- Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist-Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38*, 315-324.
- Yang, M., Wong, S., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*, 740-767. doi:10.1037/a0020473