Law Hum Behav
DOI 10.1007/s10979-007-9097-5

ORIGINAL ARTICLE

Does Interrater (Dis)agreement on Psychopathy Checklist Scores in Sexually Violent Predator Trials Suggest Partisan Allegiance in Forensic Evaluations?

Daniel C. Murrie · Marcus T. Boccaccini · Jeremy T. Johnson · Chelsea Janke

Abstract Many studies reveal strong interrater agreement for Hare's Psychopathy Checklist-Revised (PCL-R) when used by trained raters in research contexts. However, no systematic research has examined agreement between PCL-R scores from independent clinicians who are retained by opposing sides in adversarial legal proceedings. We reviewed all 43 sexual-offender civil-commitment trials in one state and identified 23 cases in which opposing evaluators reported PCL-R total scores for the same individual. Differences between scores from opposing evaluators were usually in a direction that supported the party who retained their services. These score differences were greater in size than would be expected based on the instrument's standard error of measurement or the rater agreement values reported in previous PCL-R research. The intraclass correlation for absolute agreement for the PCL-R Total score from a single rater ($ICC_{1,A}$ = .39) was well below levels of agreement observed for the PCL-R in research contexts, and below published test-retest values for the PCL-R. Results raise concerns about the potential for a forensic evaluator's ''partisan allegiance'' to influence PCL-R scores in adversarial proceedings.

Keywords Psychopathy · PCL-R · Bias · Forensic evaluation · Sexually violent predator · Sex offender civil commitment

The personality construct of psychopathy has become such a well-recognized risk factor for violence and recidivism (Hemphill et al. 1998; Salekin et al. 1996) that clinicians often assess psychopathy in forensic evaluations of adult criminal offenders (Otto and Heilbrun 2002). As a result, courts in the United States are exposed to the psychopathy construct with increasing frequency (DeMatteo and Edens 2006; Walsh and Walsh 2006). Particularly when assessing risk of violence or sexual violence, clinicians often use Hare's (1991, 2003) Psychopathy Checklist-Revised (PCL-R) as part of the forensic evaluation (Archer et al. 2006). Indeed, in a survey of 64 diplomate-level forensic psychologists, most (63%) considered the PCL-R to be ''recommended'' practice for violence risk assessment; nearly all (88%) considered it at least ''acceptable'' (Lally 2003).

It is not surprising that courts have been receptive to testimony based on the PCL-R, given the strong reliability and validity data supporting the measure (for review, see Hare 2003; Patrick 2006). For example, PCL-R research has consistently revealed strong levels of rater agreement among independent raters. Hare (2003) reported that when assessing male criminal offenders (pooled N = 4,891), the intraclass correlation coefficient for a single rating (ICC_1) was .86.

Although existing research suggests strong rater agreement for the PCL-R, most available data regarding interrater agreement is based upon studies in which trained raters—often graduate students—score the same participant in an empirical study. Usually, raters in these studies score the PCL-R only after demonstrating adequate interrater agreement during training that precedes formal data collection. Might interrater agreement for the PCL-R differ in clinical settings when scored by practicing clinicians who were not involved in intensive training and reliability checks? Would interrater agreement remain as high in an adversarial contexts in which one forensic evaluator is retained by the defense while another was retained by the prosecution?

D. C. Murrie · M. T. Boccaccini (✉) · J. T. Johnson · C. Janke
Department of Psychology, Sam Houston State University, Box 2447, Huntsville, TX 77341, USA
e-mail: Boccaccini@shsu.edu

PCL-R Rater Agreement "In the Field"

Raters in research studies are likely to be highly concordant because they receive similar and extensive training. Clinicians in the field, however, may receive varied formal or informal PCL-R training. Also, it is unlikely that experts in PCL-R assessment regularly review their scores, as is often the case in research settings. Thus, we might expect PCL-R interrater agreement to be lower in clinical practice than in research studies. In contrast, an alternative perspective might argue for *greater* agreement among PCL-R raters in the field, as compared to those in research studies, because research assistants may have access to less collateral information than a clinician in the field and research assistants usually have less clinical training or experience in interviewing and diagnosis. Generally, research reveals that more experienced clinicians perform no better than less experienced clinicians with respect to most assessment and diagnostic tasks (see Garb and Boyle 2003); however no studies have specifically examined the role of rater experience in scoring the PCL-R.

Despite dozens of studies that report PCL-R interrater agreement values among research coders, only a few shed light on agreement among practicing clinicians. All of these suggest strong agreement "in the field." For example, Gacano and Hutton (1994) examined PCL-R agreement among 31 staff members at a forensic hospital who had received rigorous training on the PCL-R and found strong correlations between pairs of raters. Two more recent studies examined PCL-R scores from practicing correctional psychologists (either compared to each other or to research coders) and found excellent rater agreement, with ICC values greater than .90 (Kroner and Mills 2001; Porter et al. 2003).

Interview Timing and PCL-R Agreement

Most rater agreement values reported in the PCL-R literature reflect a procedure in which two or more clinicians review the same collateral material and witness the same interview. In other words, raters score the same content. In practice, however, it would be uncommon for two clinicians to score the same interview. Particularly in adversarial forensic contexts, it is more common for one clinician to conduct a PCL-R interview alone. An opposing evaluator might use the same PCL-R protocol—though perhaps guide the interview in a different manner—weeks or months later. Conceivably, the examinee may present differently across these two evaluations. Although the historical data required to score the PCL-R (e.g., criminal history, institutional records) is unlikely to change, one evaluator might access some collateral records that are not available to the other.

Thus, when evaluators are not basing scores on identical sets of information, it may be unreasonable to expect interrater agreement in the field to match levels reported in research studies. Instead, a more reasonable point of comparison might be test-retest values for the PCL-R. After all, when a second rater conducts a PCL-R interview for an adversarial proceeding, the situation is more similar to a test-retest reliability scenario than to the interrater reliability scenarios in which raters score the same content. Two studies allow us to make some inferences regarding the degree of rater agreement we might expect for interviews conducted by different PCL-R raters at different times.

The first study to examine test-retest values for the PCL-R reported correlations of .85 to .89 between scores at baseline and scores obtained one month later in a sample of 88 substance abuse patients (Alterman et al 1993). Moreover, the mean PCL-R total score for the group was similar for both administrations, suggesting that PCL-R scores did not systematically increase or decrease over the one month period. In a later study, researchers reported a two-year test-retest reliability value of ICC = .60 for PCL-R total scores among 200 men (Rutherford et al. 1999). The PCL-R total scores obtained after two years were significantly higher than those obtained at baseline, although the effect size for this difference was small (Cohen's d = .24, calculated from Rutherford et al., 1999, Table 1).

Thus, test-retest values for the PCL-R appear to be lower than the interrater agreement values of ICC > .85 reported in the research literature for research assistants and practicing clinicians, especially when the time between evaluations is lengthy. Although one study suggested that PCL-R scores may systematically increase over time, this effect was small and over a two-year period. No studies that report test-retest data lead us to expect a substantial systematic change in PCL-R scores in adversarial proceedings, which usually involve readministration within a period of a few months.

Do Adversarial Legal Proceedings Influence PCL-R Rater Agreement?

It is important to emphasize that *none* of the agreement values reported in the literature involved adversarial legal proceedings, in which opposing sides request evaluation and testimony from different evaluators. Might an adversarial context influence the PCL-R scores in a forensic evaluation? Brodsky (1991) described the "pull to affiliate," by which clinicians gradually shift opinions and become increasingly committed to the legal outcome pursued by the party that retained the clinician. Other authorities have also commented on "the subtle pressure to meet clients' objectives" (Grisso 1998 p.241). If opposing

clinicians in adversarial proceedings were swayed by subtle partisan pressures or the ''pull to affiliate,'' we would expect not only decreased PCL-R agreement between raters, we would also expect scores to differ in a *systematic* manner.

Therefore, in this study, we compared PCL-R scores provided by opposing experts in an adversarial context. We examined rater agreement using ICC values and by considering whether the differences we observed could be explained by the Standard Error of Measurement (SEM) for the PCL-R, which is, at most, three points (Hare 2003). Most (68%) raters assessing the same subject should arrive at scores within ± 1 SEM of each other, and the vast majority of raters (95%) should score within ± 2 SEMs of each other. Thus, two well-trained independent PCL-R raters who were scoring the same individual should arrive at scores within three points of each other most of the time. Differences of greater than six points, or two SEM units, should be rare (< 5% of cases). If we were to find ICC values similar to those in previous research and find that most PCL-R score differences were within a range expected due to SEM, we would attribute these minor differences to random measurement error. If we were to find that differences were too large to be attributed to random measurement error, but the differences were unsystematic, we would suspect that the score differences reflected a general lessening in agreement from research to ''real-world'' settings. For example, in an adversarial context in which a period of months separated evaluations by different raters, we would not be alarmed to find rater agreement values similar to published test-retest values (Rutherford et al. 1999). However, if we were to find that scores differed more than expected based on typical interrater agreement values or test-retest values, *and* consistently differed in a direction consistent with the opposing sides that retained the evaluators, we might consider whether adverserial allegiance played some part in the poor rater agreement.

Method

Context for the Present Study

Civil commitment proceedings for offenders facing commitment as Sexually Violent Predators (SVP) provided the ideal context for this study because the PCL-R is administered routinely by two or more evaluators who represent opposing sides in an adversarial legal proceeding. In Texas, evaluators in SVP cases are required by statute (Texas Health & Safety Code § 841.023 2000) to administer a measure of psychopathy, and virtually all evaluators have used the PCL-R (Amenta 2005).

Numerous resources offer detailed descriptions of the legislation, legal proceedings, and evaluations related to Sexually Violent Predator (SVP) statutes (e.g., Doren 2002; Campbell 2004; Miller et al. 2005; Schlank 2001; Winick and LaFond 2003). Briefly put, SVP statutes allow states to identify sexual offenders perceived to be at high risk for repeated sexual offenses, and civilly commit them as a precautionary measure, in order to provide treatment and protect potential victims. In Texas, SVP procedures follow a process in which a Multidisciplinary Team (MDT) determines whether inmates approaching release have two qualifying sexual offenses, and may then refer inmates to the ''The Department'' (featuring representatives from state criminal justice and mental health agencies), who commission an ''assessment for behavioral abnormality.'' These commissioned evaluations occur on a contract basis with evaluators, usually doctoral-level psychologists. To establish such a contract, evaluators must demonstrate experience and training with sexual offender assessment, as was well as training on the PCL-R. Contracted evaluation reports typically summarize: records reviewed, clinical interview, risk factors, and an overall risk estimate (Amenta 2005). Almost all report a PCL-R score, presumably due to the requirement in state statute.

Once the Department reviews these completed evaluations, they select those whom they consider to have a ''behavioral abnormality''—which is often pedophilia, but antisocial personality disorder and psychopathy are also identified (Amenta 2005)—and refer these to the state Special Prosecution Unit (SPU), Civil Division, which has typically selected fifteen offenders per year for whom to initiate civil commitment proceedings. Again, the evaluations upon which decisions up to this point are based are not solicited directly by the petitioner (roughly analogous to the prosecutor in criminal proceedings) for purposes of trial. They are contracted third-party evaluations initially used only to screen possible candidates for civil commitment. However, evaluators understand that their evaluations and expert testimony may be required for cases that proceed to trial. During eventual civil commitment proceedings, it is the petitioner who uses the report from this evaluator to argue for civil commitment, and calls the original evaluator to serve as an expert witness.

Once the petitioner gives notice that they are initiating civil commitment proceedings, the inmate may arrange for defense counsel, which is almost always through a state-sponsored agency offering legal defense for inmates. The defense counsel for the respondent (roughly analogous to the defendant in criminal proceedings) then typically arranges for an evaluation by a psychologist. Often, as in many legal contexts, defense counsel may invite more than one evaluator to review case materials and offer preliminary opinions before hiring an evaluator for the full

evaluation. The resulting evaluations are ''defense evaluations'' in that, the evaluators were retained by the respondent for the purpose of defending against civil commitment. Unlike the original evaluations, which always result in a written report, the respondent's evaluators rarely produce a written report. Rather, the evaluator usually presents findings (including PCL-R scores) only in deposition and trial testimony. It is important to emphasize that both the original evaluator and the respondent-retained evaluator have access to essentially the same collateral materials. Both receive the same case file of correctional and law enforcement records, though individual evaluators could conceivably seek additional, external records that they deem critical (e.g., mental health records that pre-date incarceration).

Procedures and Cases Reviewed

Following approval from the affiliated university's institutional review board and permission from the relevant agency, we conducted a record review of all 43 sex offender civil commitment trials in Texas since 2000, when the relevant legislation was enacted. Thus, we reviewed the entire population of civil commitment evaluations in the state, up to the time of data collection.

To collect data, the research team reviewed written evaluations and deposition testimony for all 43 trials. We identified 23 trials in which both the petitioner and respondent reported PCL-R total scores. Although PCL-R scores were present in most (n = 42, 97.8%) of the petitioner evaluations (as required by statute), only 23 (53.5%) cases also featured PCL-R scores from the respondent's evaluation. For one case, neither side reported a PCL-R score. In the 19 cases that contained a PCL-R score from petitioner's expert, but no PCL-R score from respondent's expert, the respondent's expert did not provide (at least in deposition) a rationale for declining to administer the PCL-R. We recorded the PCL-R scores documented in the original contracted evaluation (used as evidence by the petitioner), and checked these against the PCL-R scores that the petitioner's expert reported in deposition; there were no discrepancies. Because only 5 of the 23 respondent-retained evaluators submitted written reports, we relied upon deposition testimony to identify the 23 PCL-R scores presented by respondent-retained evaluators.

Regarding the individuals facing possible civil commitment, in the 23 cases we reviewed, twelve (52.2%) were identified as Caucasian, five (21.7%) as African-American, and six (26.1%) as Hispanic/Latino. Regarding sexual offense history 7 (29.2%) had a history of convictions for offending against adults, 14 (62.5%) against children, and 2 (8.3%) against both.

SVP Evaluators

The 46 PCL-R scores (two scores per 23 cases) were produced by eleven doctoral-level psychologists. Petitioner PCL-R scores came from eight psychologists; respondent scores came from five psychologists. Two psychologists provided evaluations for both the petitioner and respondent, although each of these evaluators provided only one evaluation for each side. The remaining six psychologists who provided evaluations for the petitioner contributed an average of 3.50 (SD = 2.51, rage = 1 to 8) PCL-R scores. The three remaining respondent-retained psychologists contributed an average of 7 (SD = 4.36, range = 4 to 12) scores.

To respect privacy in this small sample, we offer only a few illustrative details on the training and qualifications of the evaluators. Three of the evaluators (one who testified for only for the petitioner, one for the respondent only, and one for both side) held diplomate status (American Board of Professional Psychology) in forensic psychology. All of the evaluators maintained some form of private practice arrangements, with most doing so as their primary employment, though at least two also held full-time academic positions.

Measure

The PCL-R (Hare 1991, 2003) is a 20 item checklist that requires a review of records and a semi-structured interview to complete. The rater assigns a score of 0 (not present), 1 (possibly present), or 2 (definitely present) to quantify the degree to which the interviewee manifests particular psychopathy criteria. The PCL-R manual (Hare 2003) reports interrater agreement values for a single evaluator (ICC_1) from .86 for male inmates to .88 for male forensic psychiatric patients. However, there is no available research documenting interrater agreement in adversarial legal contexts.

Results

Difference Scores

Table 1 lists PCL-R total scores from petitioner and respondent experts for each of the 23 inmates. Table 1 also provides a difference score, which was calculated by subtracting the respondent PCL-R score (usually lower) from the petitioner PCL-R score (usually higher). Thus, difference scores with a positive value indicate a difference in the direction predicted by adversarial bias. Difference scores ranged from –4.5 to 20, with an average of 7.81 (SD = 6.85). Table 1 reveals positive difference scores for

Table 1 Differences in PCL-R scores produced by state- versus Respondent-retained psychologists

Inmate	Offense Victims	PCL-R Total: Petitioner	PCL-R Total: Respondent	Difference	Petitioner's Psychologist	Respondent's Psychologist
1	Child	37.0	17.0	20.0	Red	White
2	Child	27.0	9.0	18.0	Green	Tan
3	Child	23.0	7.0	16.0	Blue	Orange
4	Mixed	37.0	22.0	15.0	Orange	Maroon
5	Child	29.0	15.0	14.0	White	Maroon
6	Child	36.0	23.0	13.0	Purple	Tan
7	Adult	32.0	19.0	13.0	Red	Tan
8	Child	36.8	24.0	12.8	Green	Tan
9	Child	33.0	22.0	11.0	Red	Tan
10	Mixed	21.0	11.0	10.0	Green	Black
11	Child	27.0	19.0	8.0	Purple	Tan
12	Child	30.0	23.0	7.0	Blue	Maroon
13	Child	30.0	23.0	7.0	Blue	Maroon
14	Adult	22.0	15.0	7.0	Red	Tan
15	Child	27.4	22.0	5.4	Green	Tan
16	Adult	25.0	20.0	5.0	Blue	Tan
17	Child	13.0	8.0	5.0	Blue	Tan
18	Adult	33.0	33.0	0.0	Brown	Tan
19	Adult	8.0	8.0	0.0	Brown	Black
20	Child	19.0	20.0	–1.0	Yellow	Tan
21	Adult	18.0	19.0	–1.0	Blue	Maroon
22	Child	10.0	11.0	–1.0	Blue	Black
23	Adult	20.5	25.0	–4.5	Blue	Black

Note. Mixed offenders have offended against both children and adults. Each color label represents an individual evaluator

17 (73.9%) inmates. For 14 (60.9%) inmates, difference scores were greater than two SEMs (i.e., > 6.0). In contrast, none of the negative difference scores (n = 4, 17.4%) were greater than 2 SEMs and three of the four negative difference scores were –1.0. The average PCL-R score from the petitioner was 25.86 (SD = 8.48), compared to 18.04 (SD = 6.62) from the respondent, revealing a large difference between the two sets of scores, t (22) = 5.47, p < .001, Cohen's d = 1.03.

Intraclass Correlation Coefficients for Absolute Agreement on the PCL-R

Contemporary studies tend to use intraclass correlation coefficients (ICC) to quantify evaluator agreement in a set of PCL-R scores. When multiple scores on a measure are available for the same person, the amount of variance that is attributable to the person evaluated is often converted to an ICC, which is a ratio of variance that is attributable to the person being evaluated divided by the person variance plus error (McGraw and Wong 1996; Shalverson and Webb 1991). The exact formula used to calculate an ICC depends on whether one is interested in consensus or absolute agreement (McGraw and Wong 1996; Shalverson and Webb 1991). Coefficients for *consensus* reflect only covariation in scores, regardless of the absolute values of those scores. That is, consensus coefficients consider whether the evaluators generally agree about who warrants higher scores and who warrants lower scores. For example, consensus agreement would be high if Dr. Smith reported PCL-R total scores of 4 and 11 for inmates A and B, and Dr. Jones reported scores of 27 and 36 for inmates A and B, even though the Doctors reported very different absolute values. However, coefficients for *absolute agreement* consider both co-variation and the specific value of the test score.

In absolute agreement coefficients, differences in the *specific value* of the score are considered error. It is important to use *absolute agreement* coefficients for the PCL-R because PCL-R scores have a well-established clinical meaning. For example, a score of ≥30 is often identified as a cutoff representing particularly high psychopathy (Hare 1991, 2003), and evaluators in Texas SVP cases have referenced scores ≥30 as diagnostic of psychopathy.

ICCs can be calculated for a single evaluator or for any other number of evaluators. Although all inmates in our rater-agreement analyses were assessed by two evaluators, the single evaluator values are important for the PCL-R because evaluators in SVP cases report individual scores to

the court, not the average of the two PCL-R scores from opposing evaluators. ICCs can be stepped-up for a multiple evaluator situation (e.g., two evaluators in SVP cases) using the Spearman-Brown prophecy formula (see Brennan 2001; McGraw and Wong 1996). These stepped-up values represent agreement in terms of the PCL-R score averaged across the multiple evaluators.

ICC values were calculated using SPSS 14. The ICC for a single PCL-R rating in this study was .39 (95% confidence interval = –.09 to .72). This $ICC_{A,1}$ value indicates that one can place relatively little confidence in the absolute value of a single PCL-R score in our sample of 23 cases. Indeed, researchers generally strive for rater agreement values > .85, and PCL-R ICCs for a single rater are often reported as well-above .80 (e.g., Hare 2003). The stepped-up ICC for two evaluators ($ICC_{A,2}$) was still low (.56, 95% CI = –.32 to .85), suggesting that the average of the two evaluator PCL-R scores could not provide an adequately reliable indicator of an inmates' true PCL-R score.

Another way to examine evaluator agreement in the SVP cases is to examine categorical agreement, that is, whether opposing evaluators agreed about whether an offender was a ''psychopath,'' as defined by a PCL-R total score ≥30. Although recent taxometric research (Edens et al. 2006; Marcus et al. 2004; Murrie et al., in press) does *not* support the common practice of designating offenders who score ≥30 as qualitatively distinct, 30 has been the ''cutoff score'' presented in the PCL-R manual (Hare 1991), used in numerous studies to designate a group of ''psychopaths,'' and mentioned in Texas SVP proceedings as diagnostic of psychopathy. As illustrated in Table 1, there were eight cases in which a petitioner's expert reported a PCL-R score ≥30 and the respondent's expert reported a score of less than 30. There was only one case in which both evaluators reported a score ≥30, and no cases in which the respondent's expert gave a score ≥30 and the petitioner's expert did not. The kappa coefficient for these data is .13, which indicates poor agreement. Moreover, McNemar's test of marginal homogeneity indicated that disagreements were significantly more likely to occur when the petitioner's expert gave a score of 30 or greater versus when the respondent's expert gave a score of 30 or greater ($\chi^2 = 6.00$, $p < .01$).

The results described above reveal greater-than-expected interrater disagreement, and give the impression of disagreement based on adverserial allegiance. Therefore, we attempted to explore two influences that may have artificially inflated this apparent disagreement: (a) the disproportionate role of individual evaluators who participated in multiple cases, and (b) the possibility of selection bias in the cases that featured opposing PCL-R scores.

Influence of Prolific Evaluators

Conceivably, one source of systematic error variance that could have contributed to the difference scores in Table 1 might be the idiosyncratic assessment practices of individual evaluators. For example, one or more evaluators might consistently rate all inmates lower on the PCL-R, or score certain PCL-R items higher, than other evaluators. Although the variance in test scores attributable to individual raters can be studied using a generalizability theory framework, the current study design does not allow us to estimate these effects (an ideal experimental design would require every evaluator to perform multiple evaluations for both petitioner and respondent so that individual evaluator effects could be separated from allegiance effects). Nevertheless, Table 1 allows for a qualitative examination of individual evaluators, labeled with codenames.

Given that a disproportionate number of PCL-R scores came from evaluator Blue for the petitioner (8 of 23) and evaluator Tan for the respondent (12 of 23), we examined whether scores from these two evaluators were the primary cause of the overall petitioner-versus-respondent-score differences. Both of these evaluators participated in cases with large and small PCL-R score differences, although differences larger than 2 SEM were present in 7 of 12 (58.3%) cases for Tan and 3 of 8 (37.5%) cases for Blue. However, difference scores of less than 2 SEM were observed in both of the cases in which these two evaluators provided PCL-R scores for the same inmate (inmates 16 and 11). Moreover, the difference scores in these cases were both 5.00, which is smaller than the average difference for the entire dataset (7.81). Thus, it appears unlikely that the overall pattern of PCL-R score differences can be attributed solely to any one or two prolific evaluators.

Influence of Selection Factors

As detailed previously, only 23 (54.8%) of the SVP trials contained scores from opposing evaluators. Thus, in 19 (45.2%) cases with a petitioner-retained PCL-R score (n = 42), the respondent-retained evaluators did *not* provide a PCL-R score. This feature of our data raises several questions about whether these 23 cases are truly representative of all the cases at trial. Therefore, we examined several ways in which selection bias might have created the appearance of poor rater agreement and adversarial allegiance.

First, perhaps respondent-retained evaluators declined to administer the PCL-R (or report a score) when they believed that the petitioner-reported PCL-R score was accurate. In other words, the 19 cases in which opposing PCL-R scores were not available might, in fact, reflect

acceptable or even perfect agreement. If this (admittedly optimistic) hypothesis were correct, the difference scores in the 23 cases discussed above would provide a grossly skewed impression of evaluator disagreement and partisan bias, and the ''true'' measures of agreement between opposing evaluators would be much more concordant. To quantify opposing-evaluator agreement under this hypothesis, we generated hypothetical, ''optimistic'' data. Specifically, we coded the respondent PCL-R score as being equal to the petitioner-reported PCL-R score for 19 cases that originally had no respondent PCL-R score. We then re-ran the difference score and ICC analyses using this hypothetical database.

Using this hypothetical data, the mean difference between the petitioner and respondent PCL-R scores was still large enough to reach statistical significance, with a moderate effect size, $t(41) = 4.35$, $p < .001$, Cohen's $d = .58$. The average difference score became 4.28 ($SD = 6.38$), rather than the 7.81 ($SD = 6.85$) reported for the sample of 23 cases. However, the improved difference score still reflects an average disagreement of greater than 1 SEM unit in the direction of adversarial allegiance. In this hypothetical ideal agreement scenario, 33.3% of the cases still reveal a difference of greater than 2 SEM units in the direction of adversarial allegiance. A difference of > 1 SEM was evident in 40.4% of the cases. So in this ideal agreement scenario, the proportion of cases with substantial evaluator disagreement, of course, decreases. But, the suggestion of adverserial allegiance would remain. In this ideal agreement scenario, the $ICC_{A,1}$ value for absolute agreement was .53 (95% confidence interval = .17 to .75), which is still well below typical PCL-R agreement values (Hare 2003), and even somewhat below two-year test-retest reliability values (Rutherford et al. 1999).

A second possible selection bias in our dataset could be that the 23 cases represented the most extreme PCL-R scores in the sample; in other words, perhaps PCL-R scores from the petitioner were lower among the cases in which the respondent provided no PCL-R scores. If this were true, we should expect lower scores in a second evaluation of these 23 cases simply due to regression to the mean, not because of any evaluator bias. However, the mean PCL-R total scores from the petitioner were nearly identical for the 23 cases in which a PCL-R score was reported by the respondent ($M = 25.86$, $SD = 8.48$) and the 19 cases in which no PCL-R score was reported by the respondent ($M = 25.22$, $SD = 6.05$), $t(40) = 0.27$, $p = .79$, Cohen's $d = .08$. Thus, the overall pattern of rater disagreement does not appear attributable to unusually high PCL-R scores in these 23 trials, as compared to the other 19 civil commitment trials.

A third possible selection bias operating in our dataset could be that the 23 cases might have represented the most extreme PCL-R scores in the much broader population of Texas sexual offenders *screened* for civil commitment. In other words, these 23 cases may have actually been selected for trial because the PCL-R scores were much higher than the other cases that were screened and not selected; in this scenario we might also expect lower PCL-R scores from the respondent due simply to regression to the mean. However, a review of PCL-R scores from a sample ($N = 99$) of the broader population of Texas sexual offenders screened for possible civil commitment (Amenta 2005) revealed a mean PCL-R Total score of 23.27 ($SD = 8.25$). This score was not significantly lower than our sample mean score of 25.86 ($SD = 8.48$) from the 23 trial cases, $t(120) = 1.12$, $p = .26$, Cohen's $d = .20$. Thus the overall pattern of rater disagreement does not appear attributable to unusually high PCL-R scores among the 23 inmates examined in this study as compared to the population of inmates screened for civil commitment in Texas.

Discussion

PCL-R scores have increasingly become a focus in many legal proceedings, including the civil commitment of sexual offenders (DeMatteo and Edens 2006). Despite strong and consistent interrater agreement across research studies, there have been no published studies addressing interrater agreement for the PCL-R as scored by opposing evaluators in adversarial proceedings. Our study of a small sample raises concerns about the reliability of the PCL-R in one adversarial setting in one state and suggests that further investigation is needed. Most of the cases we reviewed (14 of 23, or 60.9%) revealed petitioner-versus-respondent score differences that were greater than two SEM units *and* in the direction predicted by adverserial allegiance; another four cases revealed differences greater than one SEM unit in the same direction. ICC values for the PCL-R score in this context fell far below the ICC values reported for raters in research settings and even below the published test-retest values for the PCL-R. Although one earlier test-retest study of the PCL-R suggested there was a slight trend for scores to increase upon second administration (Rutherford et al.1999), our sample revealed that second administration values (as scored by defense-retained evaluators) more often decreased.

One detail to clarify is that the ''opposing'' evaluators might not be opposing in the strictest sense. As previously detailed, the respondent does not become involved until civil commitment proceedings have been initiated, and the respondent retains an evaluation specifically for purposes of defending against civil commitment. In contrast, the PCL-R scores that the petitioner presents are provided by evaluators under contract with the state correctional

department before it is determined whether any trial will take place. The department requests these evaluations for the purpose of screening which inmates meet SVP criteria; they then refer these evaluation reports to a separate unit, which selects a small subset of cases each year to pursue for commitment (thus, the vast majority of initial evaluations never become part of a trial). One might argue that the psychologists conducting screening evaluations are swayed by a financial interest in seeing the case go to trial (and therefore providing paid testimony) or receiving continued referrals; this possibility deserves further study. However, to be clear, these contracted evaluations are not retained solely for purposes of a trial, as are the respondent-retained evaluations and evaluations in many adversarial proceedings. Therefore, it is unclear how evaluator agreement might differ in situations wherein *both* the petitioner and the respondent select evaluators specifically for a case at hand, and both evaluators conduct evaluations specifically to present evaluation results at trial. This type of arrangement, too, warrants further study.

Study Limitations and Alternate Explanations for Rater Disagreement

A conclusion regarding adversarial allegiance is not to be offered lightly. Thus, we examined explanations *other than* adversarial allegiance that may have accounted for the disagreement among raters. For example, our findings did not suggest that the overall pattern of rater discrepancy could be attributed solely to one or two prolific evaluators. Nevertheless, our sample is quite small, and we would place more confidence in these findings if they held true across a greater number of evaluators.

We also considered whether the 23 cases with opposing PCL-R scores represented the most extreme discrepancies in the total population of SVP trials in our state. However, even when we created perfect agreement data for the 19 cases that did not report PCL-R scores from the respondent, there remained more score differences in the direction of partisan allegiance than would be expected based on measurement error alone. Finally, comparing our data on those cases that went to trial, versus those from a larger sample of Texas sexual offenders screened for potential civil commitment, suggested that our sample of 23 inmates did not have uniquely high PCL-R scores. These analyses argue against selection bias or regression to the mean as the primary explanation for our findings.

Although we attempted to rule out the possibility that our findings were attributable to selection bias with respect the sample, we cannot rule out the possibility that our findings were due to selection bias with respect to the evaluators. For example, we cannot determine whether the respondent evaluators who presented PCL-R results were the first evaluators that the respondent attorneys contacted, or whether respondent attorneys consulted a series of potential evaluators who reviewed case materials and offered opinions less helpful to the case, before the final evaluator who scored the PCL-R became involved. Likewise, the original screening evaluators, whose reports and testimony the petitioner later used at trial, may differ in important ways from evaluators whose cases did not proceed to trial (although our post-hoc analyses revealed they did not significantly differ in terms of the PCL-R scores they assigned).

Given that we could not examine any selection effects with respect to evaluators, it is important to re-emphasize that our results should not be used to characterize rater agreement for PCL-R scores in general clinical or correctional settings, or even to characterize rater agreement among clinicians who conduct forensic mental health evaluations. Rather, the study purpose was to examine the PCL-R *as scored by opposing evaluators participating in adversarial proceedings*. At least in the context studied here, the court is not exposed to opinions offered by the ''discarded experts'' whom attorneys screen for possible participation in a case but ultimately do not retain for a full evaluation and testimony (importantly, these evaluators may offer opinions that are quite concordant with one another or with the petitioner's evaluator). Rather, under these circumstances, the court is typically exposed only to opinions rendered by those evaluators who ultimately participate in the trial.

Finally, despite our efforts to explore and rule out alternate explanations, there may be other influences we could not identify, which contributed to the discrepancy among raters and appearance of adversarial bias. We also re-emphasize the very small and unique nature of our sample, that is, those few inmates whom prosecutors selected from among hundreds of others, in order to take to trial for civil commitment. Of course, it may be that cases selected for trial should be least prone to rater disagreement, in that prosecutors have opportunity to prioritize cases that appear clear-cut, as opposed to those that appear vulnerable to dispute. In any case, rater agreement is arguably most important in the select cases that proceed to trial, because these are the cases in which the PCL-R is introduced as evidence, and may influence substantial decisions about individual liberty and public safety.

Implications

To what extent are the discrepancies in PCL-R scores attributable to the PCL-R as an instrument? Certainly the instrument allows some room for subjectivity in scoring (Campbell 2004, 2006), and subjective interpretations may play a greater role when evaluators use fewer collateral

data sources. Item-level data was not available in our study, but future research should examine which PCL-R items most often reveal disagreement among raters in adversarial contexts. There may be less room for disagreement on historical items (juvenile delinquency, revocation of conditional release, and criminal versatility) scored from collateral records than on items related to interpersonal presentation (e.g., lack of remorse, superficial charm). Indeed, among male offenders, ICC_1 values tend to be around .50 for items relating to affective and interpersonal features, and tend to be closer to accepted levels (.70–.80) for more behaviorally based items that can be scored from collateral records (see Hare 2003, p. 64). Perhaps more narrowed, precise, or operationalized scoring criteria could reduce discrepancy. However, even if the instrument were determined to have unacceptably flexible scoring criteria, a scoring drift consistently in the direction of adversarial allegiance would appear more attributable to evaluators than to the instrument. Of course, another way in which adversarial allegiance may influence PCL-R results is in the decision of whether to use the measure at all. We were not able to examine those 19 cases in which respondent-retained evaluators chose not to administer the PCL-R. It is unclear whether their decisions reflected agreement with the PCL-R score provided by the petitioner's evaluator (as in our hypothetical analyses) or resulted from a strategic decision not to generate test data that could be harmful to the case (or perhaps both reasons).

Though limited to a small, selective sample, this study of ''real world'' rater agreement for an instrument employed in adversarial legal proceedings is important for at least three reasons. First, the lay public (Boccaccini and Brodsky 2002; Hans 1986; Silver et al. 1994) has expressed concern that at least some forensic psychologists tend to be ''hired guns'' who predictably reach only opinions that support the party who retained their services. Results from our small sample are certainly not sufficient to draw broad conclusions about wide-scale forensic practice. But, jurors who witnesses a large score discrepancy in the direction predicted by adversarial allegiance might be understandably skeptical of the evaluators involved, and perhaps skeptical of evaluators in the legal system more generally. Unfortunately, examples of evaluator agreement may be less visible to jurors and media.

The second reason these preliminary results appear important relates to the challenge of researching adversarial allegiance. Though other preliminary studies hint at the possibility of bias (Murrie and Warren 2005; Otto 1989), it is nearly impossible to conduct rigorously controlled *experimental* studies of this topic in the field (see Cornell 1987 for discussion). This study, too, was preliminary and exploratory. But, the PCL-R scores—along with published research on rater agreement in non-adversarial settings offer a clear metric for comparing agreement among clinicians in adversarial proceedings versus agreement among clinicians outside adversarial pressures.

The third important study implication relates to ethical guidance in forensic evaluation. Specialty Guidelines hold that psychologists who work in forensic contexts are bound by a ''special responsibility for fairness and accuracy'' (APA, Committee on Ethical Guidelines for Forensic Psychologists 1991). Similarly, the guidelines hold,

> (F)orensic psychologists take special care to avoid undue influence upon their methods, procedures, and products, such as might emanate from the party to a legal proceeding by financial compensation or other gains. … the forensic psychologist maintains professional integrity by examining the issue at hand from all reasonable perspectives, actively seeking information that will differentially test plausible rival hypotheses. (p. 661)

A revision of the Specialty Guidelines, currently underway, also appears to address objectivity in stringent terms (Otto 2006).

Our data are not sufficient to shed light on how evaluators arrived at discrepant scores, and we do not claim that partisan allegiance explains every discrepancy. Yet, the many greater-than-chance disagreements in the direction of partisan allegiance support Brodsky's (1991) observation about a ''pull to affiliate.'' We could identify no reason to believe that evaluators in our sample were any more vulnerable to partisan bias than other clinicians in adversarial proceedings; many had decades of experience, and several had advanced qualifications (e.g., Diplomate status). Thus, our results underscore the cautions about objectivity that authorities offer (Grisso 1998; Rogers 1987; Shuman and Greenberg 2003) and suggest that ethical guidelines such as those discussed above bear repeating.

How should the legal system respond to the finding that PCL-R scores may be influenced by the ''side'' retaining an evaluator? At present, it is unclear to what extent these findings might generalize to other jurisdictions or types of trial. However, a pattern of similar results across multiple contexts would begin to undermine the evidentiary value of the PCL-R as administered by privately retained evaluators in adversarial proceedings.

We encourage researchers to further examine the role that contextual, partisan pressures play in influencing opinions in forensic evaluations. We also encourage clinicians who offer opinions in adversarial proceedings to monitor closely their process of evaluation and opinion formation (Borum et al. 1993; Murrie and Warren 2005), given that our results suggested a pull towards adversarial allegiance even when scoring a structured and ostensibly objective instrument.

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.

Amenta, A. (2005). The assessment of sexual offenders for civil commitment proceedings: An analysis of report content. Unpublished doctoral dissertation. Sam Houston State University.

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.

Boccaccini, M. T., & Brodsky, S. L. (2002). Believability of expert and law witnesses: Implications for trial consultation. *Professional Psychology: Research & Practice, 33*, 384–388.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry & Law, 21*, 35–76.

Brodsky, S. L. (1991). *Testifying in court: Guidelines and maxims for the expert witness*. Washington, DC: American Psychological Association.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Campbell, T. (2004). *Assessing sex offenders: Problems and pitfalls*. Springfield, IL: Thomas.

Campbell, T. (2006).The validity of the Psychopathy Checklist-Revised in adversarial proceedings. *Journal of Forensic Psychology Practice, 6*, 43–53.

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty Guidelines for Forensic Psychologists. *Law and Human Behavior*, *15*, 655–665.

Cornell, D. G. (1987). Role conflict in forensic clinical psychology: A reply to Arcaya. *Professional Psychology: Research & Practice, 18*, 429–432.

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*, 215–241.

Doren, D. M. (2002). *Evaluating sex offenders: A manual for civil commitment and beyond*. London: Sage Publications.

Edens, J., Marcus, D., Lilienfeld, S., & Poythress, N. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology, 115*, 131–144.

Gacono, C., & Hutton, H. (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.

Garb, H. N., & Boyle, P. (2003). Understanding why some clinicians use pseudoscientific methods: Findings from research on clinical judgment. In S.O. Lilienfeld, S.J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology*. (pp. 17–38). New York. Guilford.

Grisso, T. (1998). *Forensic evaluation of juveniles*. Sarasota, Florida: Professional Resources Press.

Hans, V. P. (1986). An analysis of public attitudes towards the insanity defense. *Criminology, 24*, 393–414.

Hare, R. D. (1991). *The Hare Psychopathy Checklist Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist Revised – Second Edition*. Toronto, Ontario, Canada: Multi-Health Systems.

Hemphill, J. F., Templeman, R., Wong, S., & Hare, R. D. (1998). Psychopathy and crime: Recidivism and criminal careers. In D. Cooke, A. Forth, & R. Hare, (Eds.), *Psychopathy: Theory, research, and implications for society* (pp. 375–398). Dordrecht, Netherlands: Kluwer Academic.

Kroner, D. G., & Mills, J. F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and newconvictions. *Criminal Justice and Behavior, 28*, 471–489.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice, 5*, 491–498.

Marcus, D. K., John, S., & Edens, J. F. (2004). A taxometric analysis of psychopathic personality. *Journal of Abnormal Psychology, 113*, 626–635.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Miller, H. A., Amenta, A. E., & Conroy, M. A. (2005). Sexually violent predator evaluations: Empirical evidence, strategies for professionals, and research directions. *Law and Human Behavior, 29*, 29– 54.

Murrie, D. C., & Warren, J. I. (2005).Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice, 36*, 519–524.

Murrie, D.C., Marcus, D.K., Douglas, K.S., Salekin, R.T., Lee, Z., & Vincent, G., (in press). Youth with psychopathy features are not a discrete class: A taxometric analysis. *Journal of Child Psychology and Psychiatry.*

Otto, R. K. (1989). Bias and expert testimony of mental health professionals in adversarial proceedings: A preliminary investigation. *Behavioral Sciences and the Law, 7*, 267–273.

Otto, R. K. (2006). *Discussion of the Forensic Specialty Guidelines*. St. Petersburg, Florida: Panel presented at the annual meeting of the American Psychology-Law Society.

Otto, R. K., & Heilbrun, K. (2002). The practice of forensic psychology: A look toward the future in light of the past. *American Psychologist, 57*, 5–18.

Patrick C. (Ed.) (2006). *Handbook of psychopathy*. New York: Guilford.

Porter, S., Woodworth, M., Earle, J., Drugge, J., & Boer, D. P. (2003). Characteristics of violent behavior exhibited during sexual homicides by psychopathic and non-psychopathic murderers. *Law and Human Behavior, 27*, 459–470.

Rogers, R. (1987). Ethical dilemmas in forensic evaluations. *Behavioral Sciences & the Law, 5*, 149–160.

Rutherford, M., Cacciola, J. S., Alterman, A. I., McKay, J. R., & Cook, T. G. (1999). The 2–year test-retest reliability of the Psychopathy Checklist-Revised in methadone patients. *Assessment, 6*, 285–291.

Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the psychopathy checklist and psychopathy checklist-revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice, 3*, 203–215.

Schlank A. (Ed.) (2001). *The sexual predator: Legal issues, clinical issues, special populations* (2nd Ed). Kingston, NJ: Civic Research Institute, Inc.

Shalverson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shuman, D. W., & Greenberg, S. A. (2003). The expert witness, the adversary system, and the voice of reason: Reconciling impartiality and advocacy. *Professional Psychology: Research & Practice, 34*, 219–224.

Silver, E., Cirincione, C., & Steadman, H. (1994). Demythologizing inaccurate perceptions of the insanity defense. *Law and Human Behavior, 18*, 63–70.

Texas Health & Safety Code *§ 841.000 – 841.150* (2000).

Walsh, T., & Walsh, Z. (2006). The evidentiary introduction of *Psychopathy Checklist-Revised* assessed psychopathy in U.S. courts: Extent and Appropriateness. *Law and Human Behavior, 30*, 493–507.

Winick, B. J., & LaFond, J. Q. (2003). *Protecting society from sexually dangerous offenders: Law, justice, and therapy*. Washington, DC: American Psychological Association.