

Rater Differences in Psychopathy Measure Scoring and Predictive Validity

Paige B. Harris and Marcus T. Boccaccini
Sam Houston State University

Daniel C. Murrie
University of Virginia

Although field studies reveal that some forensic evaluators tend to assign higher psychopathy measure scores to sexual offenders than others, the extent to which these findings apply to psychopathy measure scoring in other contexts is unclear. And no study has examined the impact of evaluator differences in scoring on predictive validity. We used data from the MacArthur Violence Risk Assessment Study to examine whether there were rater differences in psychopathy measure scoring and predictive effects among trained raters in a rigorous research context. The proportion of variance in Psychopathy Checklist: Screening Version (Hart, Cox, & Hare, 1995) scores attributable to raters was larger for Part 1 (14%) than Part 2 (4%) scores. The association between Facet 4 scores and future violence was stronger among evaluators who assigned higher and more variable Facet 4 scores, but there were no similar effects for other PCL:SV scores. Although there was only limited evidence for an association between PCL:SV scoring tendencies and predictive validity, findings show that mean differences in scoring have implications for score interpretation, with the cut score that indicates a high level of risk being lower when it comes from a rater who assigns relatively low scores compared to a rater who assigns relatively high scores. These findings suggest that evaluators should carefully consider their own psychopathy measure scoring tendencies across cases and the extent to which these tendencies are consistent with the normative sample scores that form the basis of their psychopathy measure score interpretations.

Keywords: evaluator differences, PCL-R, PCL:SV, psychopathy, rater agreement

Ideally, the results of a mental health evaluation reflect only the examinee's psychological functioning, and do not at all reflect the clinician who conducted the evaluation. In other words, we expect that clinicians are more or less interchangeable; that any reasonable clinician administering a psychological assessment to a particular individual will obtain the same result as any other reasonable clinician. The assumption of interchangeable evaluators is especially important in forensic assessments, when courts rely on expert opinions to help make decisions about legal matters such as competence to stand trial, sanity, sentencing, and civil commitment. For many of the most common types of forensic cases, defendants are legally entitled to only one evaluation (*Ake v. Oklahoma*, 1985), and research suggests that, at least in some types of cases, court decisions closely follow clinician conclusions (Zapf, Hubbard, Cooper, Wheelles, & Ronan, 2004). The fact that most defendants undergo only one forensic evaluation and that judges appear to rely on the evaluator's opinion makes it especially important to examine whether or not forensic clinicians are indeed generally interchangeable. If they are not interchangeable, the outcome of any individual offender's case may depend to a non-trivial degree on which clinician performed the assessment.

Findings from several recent field studies suggest that forensic evaluators may not be as interchangeable as we might expect. For example, some clinicians conducting sanity evaluations conclude that as many as 50% of defendants they evaluate meet legal criteria for insanity, whereas others find that no defendants meet criteria (Murrie & Warren, 2005). Similarly, some clinicians find as many as 60% of defendants to be incompetent to stand trial, whereas others find fewer than 2% to be incompetent (Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008). Researchers have estimated that about 12% of the variance in competence to stand trial opinions may be attributable to differences among evaluators (Murrie et al., 2008). Of course, these apparent evaluator differences may be attributable to other factors as well, and it is unclear to what degree these evaluator differences in forensic opinions generalize to other types of assessment. But, the little available research does suggest that forensic clinicians may differ in important ways, such that they are not entirely interchangeable.

Clinician Differences in Assessment Instrument Scores

One possible method for minimizing clinician differences is for clinicians to use standardized assessment instruments, which require them to gather information in the same way (standardized administration) and to assign numeric values to that information in the same way (standardized scoring). Even if we expect clinicians to differ in their tendencies to reach conclusions about competence or sanity, which are based on somewhat difficult to operationalize legal standards, it seems that clinicians should show less variability in the scores they assign on standardized assessment instruments. Indeed, instrument manuals routinely report good to strong levels of rater-agreement for instrument scores, suggesting that most of

This article was published Online First December 8, 2014.

Paige B. Harris and Marcus T. Boccaccini, Department of Psychology and Philosophy, Sam Houston State University; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia.

We thank John Monahan for his comments on a draft of this article.

Correspondence concerning this article should be addressed to Marcus T. Boccaccini, Department of Psychology and Philosophy, Box 2447, Sam Houston State University, Huntsville, TX 77341. E-mail: boccaccini@shsu.edu

the variance in instrument scores is attributable to true differences in the traits assessed by the measure, not the clinician's personality or scoring tendencies.

Perhaps for these reasons, little research has explored clinician differences in the test scores they assign. But findings from this small body of research do raise concerns. In general clinical practice, there is evidence for significant differences in the scores clinicians assign on intelligence tests, ostensibly the most well-standardized and objective psychological assessment measures. Researchers examining the Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003) scores assigned to more than 2,500 children as part of routine clinical practice found that as much as 16% of the variance in scores was attributable to differences among clinicians (McDermott, Watkins, & Rhoad, 2014). If there are significant clinician differences for intelligence test scores, there may well be significant differences for scores from other types of standardized measures, including those more specific to forensic evaluations.

In forensic contexts, one of the most widely used, widely researched, and influential measures is Hare's Psychopathy Checklist-Revised (Hare, 2003). The PCL-R is a 20-item measure of interpersonal, affective, lifestyle, and behavioral traits, which clinicians score on the basis of an offender's records and a clinical interview. Clinicians score each item on a scale from 0 to 2, with higher scores reflecting a higher level of the psychopathic trait. PCL-R items yield scores for two factors and four facets: Factor 1 consists of an *Interpersonal* facet (Facet 1) and an *Affective* facet (Facet 2), and Factor 2 consists of an *Impulsive Lifestyle* facet (Facet 3) and an *Antisocial Behavior* facet (Facet 4).

To examine the possibility of evaluator differences in the PCL-R scores clinicians assign in risk assessment cases, researchers collected PCL-R scores assigned by 20 clinicians conducting Sexually Violent Predator (SVP) evaluations in Texas. They found that more than 30% of the variance in scores was attributable to differences among clinicians (Boccaccini, Turner, & Murrie, 2008). In other words, some evaluators assigned consistently higher PCL-R Total scores than other evaluators. For example, one evaluator assigned a mean PCL-R Total score of 17.5 across 40 cases, while another assigned a mean score of 27.1 across 60 cases, even though the evaluators saw offenders from the same population.

In a follow-up study, researchers found moderate to large evaluator differences for each PCL-R factor and facet score, with the proportion of variance attributable to evaluators ranging from 13% (Facet 2) to 25% (Factor 2; Boccaccini, Murrie, Rufino, & Gardner, 2014). Importantly, there was evidence of evaluator differences in PCL-R scoring even after using offenders' Personality Assessment Inventory (PAI, Morey, 1991) Antisocial Features scale scores to account for the possibility that some evaluators may have been randomly assigned to evaluate more antisocial or psychopathic offenders than other evaluators. Overall, these findings suggest that the PCL-R score an offender receives depends, to some degree, on the specific clinician who performs his SVP evaluation. Because those with higher PCL-R scores are more likely to be pursued for SVP commitment and committed (Murrie, Boccaccini, Caperton, & Rufino, 2012), the ultimate outcome of any offender's case may depend, to some degree, on the specific clinician assigned to perform the evaluation.

These findings raise questions about the extent to which large clinician differences are an unavoidable aspect of PCL-R scoring, or whether they are limited to certain evaluators, types of evaluations, or specific settings. Existing evidence of clinician differences in PCL-R scoring comes from a relatively small sample of clinicians ($n \approx 25$), from one state (Texas), with different levels of experience and training, conducting one type of evaluation (SVP). It may be that there is something unique about SVP evaluations or the relatively small subset of evaluators who perform these evaluations in Texas that leads to greater evaluator differences in PCL-R scores. Indeed, many PCL-R rater agreement studies report high levels of agreement, suggesting that there is little room for evaluator differences to impact scoring (see Hare, 2003). But rater agreement is lower in some settings (Lloyd, Clark, & Forth, 2010; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Sturup et al., 2014), and it is possible that there will be sizable evaluator differences in PCL-R scores in many settings, given the amount of subjectivity required to score items such as lack of remorse, superficial charm, and failure to accept responsibility (see Rufino, Boccaccini, & Guy, 2011).

One way to answer these questions is to examine whether there is evidence for evaluator differences in scoring in a setting in which evaluator differences should be least likely to exist: a controlled research study in which all evaluators complete the same training. Although many PCL-R researchers conduct rater-agreement studies, most report only the proportion of variance attributable to differences in psychopathy (i.e., rater-agreement). Although the remaining variance could be attributable to rater differences, other systematic sources of measurement error, or random measurement error, few studies have attempted to separate the remaining variance into proportions attributable to these other factors. One controlled rater-agreement study that did quantify the size of rater differences in PCL-R scoring used generalizability theory analyses to show that between 2% (Facet 4) and 27% (Facet 2) of the variance in PCL-R scores was attributable to rater differences (Miller, Rufino, Boccaccini, Jackson, & Murrie, 2011). But each clinician scored only four cases in that study, and they scored the PCL-R using circumscribed records and video-recorded interviews designed for training purposes, as opposed to in-person clinical interviews.

The MacArthur Violence Risk Assessment Study

Data from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) allow for a large-scale examination of rater differences in psychopathy measure scoring among trained raters scoring a PCL measure on the basis of their own clinical interview and record review. The MacArthur study was a large scale ($N = 1,136$), longitudinal, multisite study of the ability of more than 130 variables (e.g., diagnosis, psychiatric symptoms, personality traits, violent thoughts, background characteristics) to predict future violence among patients released from civil psychiatric hospitals (Monahan et al., 2001). The MacArthur researchers followed civil psychiatric patients for one year after release from inpatient hospitals, with follow-up interviews occurring every 10 weeks. Trained research assistants ($N = 24$) scored 871 of the 1,136 study participants on the Psychopathy Checklist: Screening Version (PCL:SV; Hart, Cox, & Hare, 1995), which was the strongest

predictor of future violence in the study ($AUC = .73$; Skeem & Mulvey, 2001).

Although the MacArthur study used the PCL:SV and not the PCL-R, there are several reasons why a study of rater differences in PCL:SV scoring will further our understanding of clinician differences in psychopathy measure scoring. First, the PCL:SV and PCL-R are similar in structure and overall function. Clinicians score both measures based on a combination of file review and interview, and the 12 PCL:SV items were designed to capture the content of 18 of the 20 PCL-R items. Because there was evidence of clinician differences in PCL-R scoring for each PCL-R factor and facet among Texas SVP evaluators (see Boccaccini et al., 2014), there is no reason to believe that limiting the measure to 12 items will prevent clinician differences in scoring. In other words, there is no evidence that PCL-R clinician differences are a product of the items that the two measures *do not* share.

Second, the MacArthur study differs from existing studies in important ways that have implications for understanding existing findings of clinician differences in psychopathy measure scoring (e.g., Boccaccini et al., 2014). Perhaps most importantly, the MacArthur study allows for an examination of clinician differences in a setting in which these differences should be least likely to occur. All 24 of the MacArthur study raters completed the same PCL:SV training—conducted by PCL:SV authors (Stephen Hart and Robert Hare)—and passed a reliability check before scoring patients for the study (see Skeem & Mulvey, 2001). Thus, there were explicit attempts to prevent rater differences from influencing PCL:SV scores. In contrast, Texas SVP evaluators—like most field evaluators—are not *required* to complete any specific type of training or pass a reliability check before performing evaluations (although many report having attended a formal training workshop; Boccaccini et al., 2014). A finding of only small rater differences in PCL:SV scoring in the MacArthur study would suggest that large clinician differences are not inevitable for the PCL measures, and may be minimized through the types of training and reliability checks typical of rigorous research studies.

The MacArthur study also allows for an examination of the impact of rater differences on predictive validity. The finding that some clinicians assign consistently higher scores than others raises questions about whether scores from some of these clinicians are better predictors of future violence and offending than those from other clinicians. Indeed, some field research with the PCL-R reveals that scores from some evaluators are more predictive of offending than scores from other evaluators (Murrie et al., 2012). It may be that scores from clinicians who assign especially high or especially low scores across cases are relatively weak predictors of future violence, because the scores they assign do not adequately differentiate between high and low scoring offenders. Or, it could be that scores from high and low scoring clinicians can both predict future violence, but that the scores from different clinicians may need to be interpreted differently. For example, the score that indicates “high risk” may be different for clinicians who assign high scores compared with clinicians who assign low scores. There is some evidence for this type of effect in adversarial settings, where the PCL-R score that indicates high risk is about five points higher when it comes from a prosecution-retained evaluator than a defense-retained evaluator (Boccaccini, Turner, Murrie, & Rufino, 2012).

Finally, it may be that variability in scoring, as opposed to elevation in scoring, is most important for understanding variability in prediction. A large standard deviation around the mean score assigned by an clinician would indicate that the clinician is willing to assign both high and low scores—ostensibly differentiating between those with high and low levels of psychopathy—whereas a smaller standard deviation would suggest that the clinician tended to assign scores very close to his or her idiosyncratic mean, regardless of whether that mean was relatively high or low. As a result, predictive validity may increase as the size of the standard deviation around the clinicians’ mean scores increases.

Current Study

We use data from the MacArthur study to examine whether there is evidence of significant rater differences in PCL:SV scoring. Because the PCL:SV raters received the same training and passed a reliability check before scoring participants for the study, we expected that the proportion of variance in scores attributable to raters would be smaller than the 20% to 30% values observed in PCL-R field studies (Boccaccini et al., 2008, 2014). But, given the amount of subjectivity required to score some PCL:SV items (see Rufino et al., 2011), we expected that there would still be some evidence of rater differences, even among these trained raters. Following existing evaluator differences research (Boccaccini et al., 2014), we used traits known to be correlated with psychopathy as covariates to control for the possibility that some raters were unintentionally assigned to score participants with higher levels of psychopathy than others. We also examined whether the relationship between PCL:SV scores and future violence varied across raters and the extent to which rater scoring tendencies (e.g., mean score and standard deviation across evaluations) might explain variability in predictive effects.

Method

Raters

There were 24 raters who scored the PCL:SV in the MacArthur study, each identified by an identification number in the publicly available dataset. Each rater scored civil psychiatric patients at one of three sites (Site 1 $n = 13$ evaluators, Site 2 $n = 6$, Site 3 $n = 5$). The number of PCL:SV scores assigned by raters ranged from 6 to 91 ($M = 36.29$, $SD = 24.33$). The majority of the PCL:SV assessments ($n = 779$) took place during the first follow-up of the MacArthur study (10 weeks). The remaining patients were scored on the PCL:SV during the second follow-up (20 weeks; $n = 77$) or the third follow-up (30 weeks; $n = 15$). To ensure that we compared only raters who had conducted a sufficient number of PCL:SV assessments to examine a pattern of scoring and predictive accuracy across cases, we limited analyses to the 18 raters who scored the PCL:SV for at least 20 patients with valid outcome data (violence at 20 weeks). After we removed patients scored by one of the less prolific raters ($n = 67$) or patients who were missing outcome data ($n = 11$), there were 793 remaining patients.

There is no reason to believe that some MacArthur study raters were assigned to score patients with higher levels of psychopathic traits than others (J. Monahan, personal communication, July 1, 2014). The mean PCL:SV Total score assigned by raters was

relatively similar at each of the three sites: Site 1 ($M = 9.15$, $SD = 5.56$, $n = 315$), Site 2 ($M = 8.60$, $SD = 5.30$, $n = 165$), and Site 3 ($M = 7.99$, $SD = 5.87$, $n = 313$). Although the difference in mean scores across sites was large enough to reach statistical significance in this large sample, $F(2, 790) = 3.34$, $p = .04$, $\eta^2_{\text{partial}} = .008$, less than one percent of the variance in scores was attributable to the evaluation site. Cohen's d values comparing mean scores from pairs of sites indicate that differences were small ($d = 0.10$ to 0.20).

Patients

The PCL:SV scores we used for this study come from 793 short-term civil psychiatric patients discharged into the community at one of three sites. Patients ranged in age from 18 to 40 years ($M = 29.94$, $SD = 6.13$) and were somewhat more likely to be male than female (56.6% male, 43.4% female). Sixty-nine percent were Caucasian, 29% were African American, and the remaining 2% were Hispanic.

Measures

PCL:SV. The PCL:SV (Hart, Cox, & Hare, 1995) is a 12-item, clinician-scored measure of psychopathic traits. The PCL:SV is similar to the 20-item PCL-R in structure and overall function, but has only 12 items and scoring rules designed to facilitate scoring for those without a significant history of criminal behavior. Each of the 12 items is scored on the same 3-point scale as the PCL-R (i.e., 0, 1, 2). The PCL:SV structure is similar to that of the PCL-R, in that both are divided into two analogous sections (PCL:SV has two Parts, while PCL-R has two Factors). The first of these two Parts/Factors is composed of interpersonal and affective traits, whereas the second Part/Factor reflects social deviance traits. To further delineate psychopathic symptomology, there are four facets within the PCL-R—two within each Factor (Hare, 2003). The Factor 1 facets are Interpersonal (Facet 1) and Affective (Facet 2) traits. The Factor 2 facets are Lifestyle (Facet 3) and Antisocial (Facet 4) traits. Using the items from the PCL-R, it is possible to create corresponding PCL:SV facets that are similar in content to the PCL-R facets (Laurell, Belfrage, & Hellström, 2010; Walters, Knight, Grann, & Dahle, 2008).

We obtained PCL:SV Total, Part 1, Part 2, and item scores from the public-domain MacArthur dataset (MacArthur Research Network, 2005). We calculated Facet 1–4 scores using PCL:SV item scores. We summed PCL:SV items 1, 2, and 3 to make the PCL-R-based Facet 1 score, items 4, 5, and 6 for the Facet 2 score, items 7, 9, and 10 for the Facet 3 score, and items 8, 11, and 12 for the Facet 4 score. We were able to calculate Total, Part 1, and Facets 1–3 scores for all 793 participants, but had Part 2 and Facet 3 scores for only 787 participants. Internal consistency (Cronbach's alpha) was good for the PCL:SV Total (.87), Part 1 (.84), and Part 2 (.82) scores, and greater than .72 for each of the four PCL:SV facet scores.

Each rater who scored the PCL:SV for the MacArthur study completed a full day of PCL:SV training provided by PCL:SV authors Stephen Hart and Robert Hare (see Skeem & Mulvey, 2001). The trainers provided an overview of the PCL measures, reviewed the description and scoring criteria for each PCL:SV item, and used three videotaped interviews to facilitate discussions regarding proper item scoring. Following the training, each rater scored more than 10

additional cases, the scores from which were “sent to Drs. Hart and Hare for reliability analyses and approval” (p. 362).

The MacArthur researchers examined rater-agreement by having raters video record five of their cases (see Skeem & Mulvey, 2001). The other raters then scored the PCL:SV for each of the cases from other raters. The researchers “defined [raters] as in agreement when their PCL:SV total scores fell within 5 points of one another” (p. 364). They concluded that “this analysis indicated that combined rates of interrater reliability on [the] PCL:SV were “good” ($K = 0.66$)” (p. 364). They did not report any other PCL:SV rater-agreement findings, but stated that they could not calculate ICC values or kappa statistics based on classifications of patients as psychopathic versus nonpsychopathic because “interrater reliability data were collected nearly a decade ago and could not be located for reanalysis” (p. 364).

Future violence. We used the presence or absence of violence during the first 20-week period as the outcome variable because this was the outcome variable in the main MacArthur study publication reporting the predictive validity of the PCL:SV (Skeem & Mulvey, 2001). The MacArthur research team collected violence outcome data using follow-up interviews of patients and collaterals (e.g., family members) and state police and hospital records (see Monahan et al., 2001). The researchers operationally defined violence as “battery that resulted in physical injury (ranging from bruises to death), sexual assaults, assaultive acts that involved the use of a weapon, or threats made with a weapon in hand” (Skeem & Mulvey, 2001, p. 361). In the current sample of 793 patients, the base rate of violence during the first 20 weeks was 18.7%.

Participant characteristics. Although there is no reason to suspect that some raters were assigned to score patients that were higher in psychopathic traits than others, such differences could have occurred by chance. To control for nonrandom assignment, we used six patient variables correlated with the PCL:SV in the MacArthur study as covariates in our rater differences analyses: prior arrests for crimes against persons ($\eta = .29$ with future violence), prior arrests for crimes against property ($\eta = .27$), chart diagnosis of Antisocial Personality Disorder ($\eta = .31$), NEO-PI-R agreeableness score ($r = -.40$), BIS Nonplanning level ($r = .23$), and the Behavioral Aspect score from the NOVACO Anger Scale ($r = .24$; effect size values from Skeem & Mulvey, 2001).

Results

The mean patient PCL:SV Total score across all raters was 8.58 ($SD = 5.65$), with scores being somewhat lower for Part 1 ($M = 3.15$, $SD = 3.02$) than Part 2 ($M = 5.42$, $SD = 3.33$). Scores also tended to be lower for Facets 1 ($M = 1.43$, $SD = 1.61$) and 2 ($M = 1.72$, $SD = 1.71$), than Facets 3 ($M = 2.93$, $SD = 1.87$) and 4 ($M = 2.50$, $SD = 1.87$).

Table 1 lists the mean PCL:SV Total, Part 1, and Part 2 score assigned by each of the 18 raters. The mean scores in Table 1 suggest that some raters did assign consistently higher scores than others. For example, Rater 13 assigned a mean Total score of 14.56 across 24 evaluations, whereas Rater 31 assigned a mean Total score of 5.70 across 26 evaluations. These scores fall at approximately the 75th and 30th percentiles, respectively (see Hart et al., 1995). There were similar differences for Part 1 and Part 2 scores, although the differences appear to be more pronounced for Part 1 scores (see Table 1). The mean Part 1 scores assigned by evaluators differed by as much as

Table 1
Descriptive Statistics for PCL:SV Total, Part 1, and Part 2 Scores Assigned by Each MacArthur Study Rater

Rater	n	Total		Part 1		Part 2	
		M	SD	M	SD	M	SD
13	24	14.56	6.88	7.21	3.96	7.17	3.28
19	20	11.08	5.92	4.15	3.18	6.96	3.14
22	31	10.23	4.60	4.23	2.65	6.01	2.93
16	46	10.09	6.57	3.78	3.76	6.28	3.33
12	47	9.73	5.94	3.53	3.73	6.20	3.45
23	77	9.42	5.38	3.56	2.81	5.85	3.01
32	88	9.37	6.51	3.19	3.31	6.16	4.12
33	91	8.64	6.43	3.99	2.92	4.69	3.84
4	27	8.26	4.96	2.93	3.10	5.33	2.75
18	33	8.21	4.75	2.70	2.51	5.52	3.07
9	28	7.75	4.13	2.68	2.00	5.07	2.71
17	44	7.75	4.23	2.61	2.10	5.15	2.91
7	22	7.50	4.42	2.00	2.76	5.50	2.26
8	24	7.25	3.05	1.63	1.35	5.66	2.90
35	56	6.95	4.37	1.93	2.08	5.02	2.94
36	52	6.79	5.33	2.75	2.56	4.04	3.24
24	57	6.61	5.02	2.33	2.61	4.28	2.94
31	26	5.70	3.78	1.12	1.28	4.59	2.67

Note. Rater numbers are those listed in the publically available MacArthur study dataset.

6 points (range = 1.12 to 7.21), compared with about 3 points for Part 2 scores (range = 4.04 to 7.17).

Rater Differences in PCL:SV Scoring

We used multilevel linear modeling (MLM) to more thoroughly examine rater differences in PCL:SV scoring. In the MacArthur study, patients are nested under evaluators and the PCL:SV scores they receive are, to some extent, a product of the specific rater assigning the score. MLM models provide a variance component that can be used to calculate the proportion of variance in scores attributable to this nesting. We began by specifying a separate unconditional random effects model for each PCL:SV score (i.e., total, parts, and facets). Results from each model allowed us to calculate the proportion of variance in scores attributable to differences among raters, as opposed to other factors (e.g., patients differing in psychopathy, random error). This proportion of variance is an intraclass correlation coefficient (ICC). We used Stata 13.0 for all MLM analyses.

Table 2 provides the rater difference ICC values for PCL:SV total, factor, and facet scores from the MacArthur study, as well as those from existing clinician differences research for comparison (Boccaccini et al., 2014). Overall, the proportion of variance in scores attributable to MacArthur study raters ranged from 14% (Part 2) to 3% (Facet 4). These proportions were larger for the scores evaluators assigned on Part 1 (14%) and its facets (both 12%), than Part 2 (4%) and its facets (5% and 3%). Each of these ICC values for rater differences from the MacArthur study was notably lower than the corresponding values from PCL-R field research (e.g., Boccaccini et al., 2014).

Although each of the ICC values was large enough to reach statistical significance ($p < .01$) in the large MacArthur study sample, several of the ICC values were relatively small. For example, one way to think about these ICC values is by subtracting

them from 1.00, with the result providing information about the proportion of variance attributable to factors other than raters. If only 8% of the variance in PCL:SV Total scores is attributable to rater differences in scoring, as much as 92% could be due to patients differing in true levels of psychopathy (although other sources of error likely explain some of that variance). If we want the proportion of variance attributable to differences in psychopathy to be at least .80 (i.e., the rater-agreement coefficient), rater differences in the MacArthur study may be tolerable for PCL:SV Total (.08), Part 2 (.04), and Facet 3 (.05) and 4 (.03) scores. Because there will always be some variability due to random measurement error, the ICC values for Part 1 (.14), Facet 1 (.12) and Facet 4 (.12) scores may be more problematic.

One possible explanation for the Part 1, Facet 1, and Facet 2 rater difference findings is that some raters may have been assigned to score patients with higher levels of these psychopathic traits than

Table 2
Proportion of Variance in Psychopathy Checklist Scores Attributable to Raters

PCL-SV/R score	Proportion of variance in scores attributable to raters	
	MacArthur	Texas SVP
Total	.08	.32
Part 1/Factor 1	.14	.23
Part 2/Factor 2	.04	.25
Facet 1	.12	.18
Facet 2	.12	.13
Facet 3	.05	.24
Facet 4	.03	.19

Note. All values in this table are significantly different from zero at $p < .01$. For the MacArthur study, $n = 793$ for Total and Part 1 scores, and $n = 787$ for Part 2 scores. Texas SVP values from Boccaccini et al. (2014).

others. To account for this possibility, we used participant background variables correlated with psychopathy as fixed-effect covariates in the MLM models to control for nonrandom assignment. These analyses allowed us to calculate the proportion of variance in PCL:SV scores attributable to raters after accounting for the possibility that some raters may have been assigned to score patients who were less agreeable, more impulsive, or had more prior arrests than those assigned to other raters. To maximize sample size, we used a separate model for each PCL:SV score with each covariate. If we had examined each of the five covariates in the same model, the sample size would have decreased from 793 to 544 participants because of missing data (see Table 3 for n values).

Table 3 provides the ICC values for PCL:SV scores after controlling for patient characteristics known to be associated with PCL:SV scores in the MacArthur study. The findings in Table 3 suggest that nonrandom assignment may explain some, but not all of the rater differences effect, especially for Part 1 scores. For example, more than 10% of the variance in Part 1 scores was attributable to raters even after controlling for covariates. Across PCL:SV scores, the largest reduction in variance attributable to raters occurred when we entered self-reported agreeableness as a covariate. After accounting for patient agreeableness, only 5% of the variance in PCL:SV Total scores and only 1% of the variance in Part 2 scores was attributable to rater differences.

Rater Differences in PCL:SV Scoring and the Predictive Validity of PCL:SV Scores

Findings of rater differences in PCL:SV scoring—especially Part 1 scoring—raise questions about whether these differences may influence the predictive validity of PCL:SV scores. For example, might rater scoring tendencies explain why scores from some raters are better predictors of future violence than scores from other raters? Or, do rater differences in PCL:SV scoring suggest that scores from some evaluators need to be interpreted differently than scores from other evaluators?

To examine the relationship between rater differences in PCL:SV scoring and PCL:SV score predictive validity, we first calculated predictive effects (Cohen's d) for each PCL:SV score, separately for each rater. Table 4 lists these predictive effects for Part 1 and Part 2 scores. The values in Table 4 suggest that scores from some raters did indeed predict future violence better than those from others. In the overall sample, both Part 1 ($d = .72$) and

Part 2 ($d = .94$) scores were moderate to strong predictors of future violence. But the predictive effects for scores from individual raters ranged to from $d = 1.51$ (Rater 19) to $d = .00$ (Rater 8) for Part 1 scores, and $d = 1.69$ (Rater 19) to $d = .07$ (Rater 4) for Part 2 scores.

But the findings in Table 4 suggest that rater scoring differences may not explain much of the variability in predictive effects. For example, the predictive effects for scores from some raters were nearly identical, despite the raters assigning noticeably different PCL:SV scores. For instance, the scores assigned by Raters 19 and 35 were the strongest predictors of future violence for both Part 1 and part 2 scores, but Rater 19 assigned consistently higher scores than Rater 35. Moreover, there was no evidence that scores from more extreme scoring raters (e.g., high mean, low mean) were less predictive than scores from other raters. For example, scores from low scoring raters led to both relatively large (Rater 31) and small (Rater 8) predictive effects. Finally, Table 4 appears to show a similar amount of variability in predictive effects for Part 1 and Part 2 scores, despite our earlier finding that there were larger rater differences in Part 1 scoring than Part 2 scoring. If rater differences in scoring were directly related to rater differences in predictive validity, we would have expected more variability in Part 1 predictive effects than Part 2 predictive effects.

Meta-analysis. We used meta-analysis to provide more direct examination of the possible relationship between rater scoring tendencies and variability in predictive effects. We treated each evaluator as a “study” and used SPSS metaregression macro programs (Lipsey & Wilson, 2001) to examine whether the mean score that the rater assigned across evaluations, or the standard deviation of those scores, explained a significant amount of the variance in PCL:SV predictive effects (e.g., the d values listed in Table 4). We used Hedges' (1981) correction to account for upward bias in effect sizes based on a smaller number of assessments and weighted each effect by its inverse variance weight. We used a random-effects model for each analysis. Because predictive effects for PCL:SV scores tended to be larger for evaluators at Site 3 (PCL:SV Total score $d = 1.20$, 95% CI = .88, 1.53) than Site 1 ($d = .78$, 95% CI = .55, 1.06) and Site 2 ($d = .76$, 95% CI = .37, 1.16), we included two dummy coded variables (with Site 3 as the reference group) in each regression equation to represent evaluation site. We ran separate evaluator mean and standard deviation models for each PCL:SV score (total, part, facet).

Table 3

Proportion of Variance In PCL:SV Scores Attributable to Raters After Controlling for Participant Characteristics

Covariate	Total	Part 1	Part 2	Facet 1	Facet 2	Facet 3	Facet 4
None	.08	.14	.04	.12	.12	.05	.03
Prior arrests person crimes	.07	.11	.04	.10	.08	.06	.03
Prior arrests property crimes	.07	.11	.04	.10	.09	.05	.03
NEO-PI-R Agreeableness	.05	.11	.01*	.09	.11	.03	.01**
BIS Nonplanning	.09	.15	.04	.13	.12	.05	.04
NAS Behavioral	.08	.14	.04	.12	.12	.04	.03

Note. Values in the table are intraclass correlation coefficients (ICCs), which indicate the proportion of variance in PCL:SV scores attributable to raters. Unless otherwise indicated, ICC values are significantly greater than zero at $p < .01$. Because of missing data, $n = 787$ to 793 for models with no covariates, 653 to 658 for models prior arrest variables, 667 to 671 for models with the NEO-PI-R, 768 to 774 for models with the Barratt Impulsivity Scale (BIS), and 785 to 791 for models with the Novaco Anger Scale (NAS).

* $p = .052$. ** $p = .23$.

Table 4
Predictive Validity of PCL:SV Part 1 and 2 Scores For Each Rater

Rater	<i>n</i>	Yes future violence		No future violence		Cohen's <i>d</i>	95% CI
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
PCL:SV Part 1 Scores							
19	20	7.20	2.49	3.13	2.75	1.51	[.37, 2.62]
35	56	4.00	2.24	1.63	1.90	1.22	[.39, 2.04]
24	57	4.55	3.08	1.80	2.22	1.14	[.45, 1.83]
9	28	4.40	1.67	2.30	1.89	1.13	[.10, 2.13]
31	26	2.33	1.15	0.96	1.22	1.13	[−.13, 2.36]
18	33	4.50	2.67	2.12	2.20	1.03	[.18, 1.85]
17	44	4.11	3.02	2.23	1.65	.95	[.18, 1.70]
33	91	6.25	2.88	1.63	1.90	.93	[.30, 1.55]
32	88	5.06	3.26	2.71	3.17	.90	[.21, 1.26]
12	47	5.86	5.49	3.13	3.26	.75	[−.07, 1.57]
36	52	4.33	2.88	2.54	2.47	.71	[−.15, 1.57]
13	24	8.56	3.97	6.40	3.85	.55	[−.29, 1.39]
22	31	5.25	2.25	3.87	2.74	.53	[−.29, 1.34]
23	77	4.77	2.92	3.13	2.74	.53	[−.08, 1.13]
4	27	4.00	4.24	2.74	2.94	.40	[−.67, 1.47]
16	46	4.23	3.75	3.61	3.81	.19	[−.48, .81]
7	22	2.25	0.71	1.95	2.90	.17	[−1.27, 1.65]
8	24	1.63	1.30	1.63	1.41	.00	[−.02, .02]
PCL:SV Part 2 Scores							
19	20	10.20	1.48	5.88	2.78	1.69	[.52, 2.83]
35	56	8.57	1.72	4.51	2.73	1.54	[.69, 2.38]
33	90	9.16	2.63	4.07	3.56	1.48	[.80, 2.13]
32	87	10.16	2.27	5.12	3.85	1.40	[.84, 1.96]
18	33	8.25	1.98	4.64	2.86	1.35	[.48, 2.20]
24	57	6.45	3.30	3.76	2.63	1.03	[.29, 1.65]
23	75	8.08	2.54	5.43	2.92	.93	[.29, 1.56]
16	45	8.31	1.80	5.45	3.47	.92	[.25, 1.59]
9	28	6.60	1.34	4.74	2.83	.70	[−.29, 1.68]
13	23	8.38	3.97	6.53	3.09	.68	[−.31, 1.44]
12	47	7.86	2.79	5.91	3.50	.57	[−.24, 1.38]
36	52	5.67	3.44	3.82	3.19	.57	[−.29, 1.43]
31	26	6.33	2.08	4.37	2.69	.50	[−.49, 1.96]
7	22	6.50	0.71	5.40	2.35	.48	[−.99, 1.93]
8	24	6.25	2.82	5.36	2.98	.30	[−.55, 1.15]
17	44	5.78	2.59	4.99	3.00	.27	[−.47, 1.00]
22	31	6.38	3.29	5.89	2.86	.16	[−.65, .97]
4	27	5.50	2.52	5.30	2.84	.07	[−.99, 1.13]

The metaregression analyses indicated statistically significant effects for Facet 4 mean and standard deviation values. Scores from raters who assigned higher Facet 4 scores were stronger predictors of future violence than those who assigned lower scores ($b = .66$, 95% CI = .12, 1.20, $p = .02$), even after controlling for weaker predictive effects at Site 1 ($b = -.93$, 95% CI = −1.50, −.40, $p < .001$) and Site 2 ($b = -.49$, 95% CI = −1.02, .05, $p = .07$) compared with Site 3. There was a similar effect for Facet 4 standard deviation values. Scores from raters with larger Facet 4 standard deviations were stronger predictors of future violence than those from raters with smaller standard deviation values ($b = 1.10$, 95% CI = .12, 2.06, $p = .03$). Moreover, once the rater standard deviation values were included the model, the differences between scores from Sites 1 and 3 ($b = -.42$, 95% CI = −.89, .06, $p = .08$) and Sites 2 and 3 ($b = -.33$, 95% CI = −.89, .22, $p = .24$) were no longer large enough to reach statistical significance.

There was a moderate-sized, positive correlation between the mean Facet 4 score and the standard deviation of Facet 4 scores assigned by raters ($r = .35$). Although this correlation is not large enough to reach statistical significance in this sample of only 18 evaluators ($p = .15$), the moderate size of the correlation indicates that raters who assigned higher scores across cases also had larger standard deviations around those scores, indicating a willingness to assign both high and low scores. It also indicates that raters who tended to assign low scores tended to do so consistently across cases (i.e., low mean, low standard deviation). Predictive effects were lowest for these “consistently low” Facet 4 raters.

The metaregression analyses provided no evidence of other significant effects for evaluator mean ($p = .47$ to .99) or standard deviation ($p = .29$ to .86) values.

Do scores from some raters need to be interpreted differently than those from others? Although our meta-analytic findings show that knowing whether a rater tends to assign high or low

PCL:SV scores does not consistently tell us whether the scores from that rater predict future violence, it may be that a rater's scoring tendencies have implications for score interpretation. For example, the specific score that indicates a high level of risk may be much lower when it comes from a rater who assigns relatively low scores compared to a rater who assigns relatively high scores.

Table 5 lists sensitivity, specificity, and overall hit rate values for PCL:SV Part 1 cut scores for four raters. We used Part 1 scores to examine this issue because of the larger rater difference effect for Part 1 than Part 2 scores. Part 1 scores from each of these four raters were predictive of future violence—with Cohen's d values greater than 1.00—but the mean PCL:SV Part 1 score they assigned across participants ranged from 1.12 to 4.15. The cut score findings in Table 5 show that the raters' scoring tendencies do have implications for score interpretation. This is perhaps most evident in the cut score sensitivity values, which indicate the proportion of future-violent patients scored by the rater who also scored at or above the cut score. For a cut score of ≥ 4 , sensitivity values ranged from .00 (Rater 31) to 1.00 (Rater 19). In other words, Rater 31—who assigned low scores across cases—never assigned a score of 4 or higher to a patient who committed an act of future violence. On the other hand, Rater 19—who assigned high scores across cases—never assigned a score below 4 to a patient who committed an act of future violence.

The pattern was similar for other cut scores, with sensitivity for any given cut score (with the exception of 1) being lower for low scoring raters than high scoring raters. The pattern was in the opposite direction for specificity values, which indicate the proportion of nonviolent patients who scored below the cut score. The

specificity value for any given cut score was higher for low scoring raters than high scoring raters (see Table 5).

Discussion

Although there was a statistically significant amount of variance in PCL:SV scores attributable to evaluators in the MacArthur study, these differences were smaller ($ICC = .03$ to $.14$) than those researchers have documented for PCL-R scores assigned in field settings ($ICC = .13$ to $.32$; Boccaccini et al., 2014), but similar to those for other standardized measures (e.g., $.03$ to $.16$ for IQ scores; McDermott et al., 2014). These findings suggest that there will always be some variance attributable to clinicians or raters in PCL measure scoring, but that large differences are not inevitable.

Rater difference effects in the MacArthur study were smaller for Part 2 scores than for Part 1 scores, corresponding with findings from both field and research studies documenting stronger rater-agreement for PCL-R and PCL:SV Factor/Part 2 than Factor/Part 1 scores (Edens, Boccaccini, & Johnson, 2010; Miller et al., 2012; Pedersen, Kunz, Rasmussen, & Elsass, 2010). As rater-agreement increases, there is less room for rater differences to influence scores.

Training and Scoring Differences

One possible explanation for the smaller rater difference effects in the MacArthur study is that each rater completed the same training program before being allowed to score patients for the study. This training was provided by PCL:SV authors (Robert Hare and Stephen Hart), and included didactic, practice scoring, and reliability check components (see Skeem & Mulvey, 2001). Each of these training activities should facilitate standardized administration and scoring, and reduce idiosyncratic scoring tendencies among raters. In field settings, clinician differences in PCL-R scoring were smaller among the subset of clinicians who reported having completed a PCL-R training workshop provided by a recognized training expert (e.g., Robert Hare, Stephen Hart, Adelle Forth), and also showed the expected pattern of smaller clinician differences for Factor 2 scores ($ICC = .13$) than Factor 1 scores ($ICC = .17$; Boccaccini et al., 2014). Together, these findings from field and research (i.e., MacArthur) settings converge to provide support for completing PCL measure training, but this support is entirely indirect; none of these studies examined rater-agreement or clinician differences both before and after training.

Although the MacArthur study findings highlight potential benefits of standardized training and reliability checks, they also show that training cannot completely prevent rater differences. Rather, these differences will always play some role in PCL scoring. Our findings also show that completing the same training does not ensure that scores from each rater will be equally predictive of patient outcomes. Cohen's d values for predicting future violence ranged from .00 to 1.51 across raters for PCL:SV Part 1 scores, and .07 to 1.69 for Part 2 scores. Although some of the more extreme effects were based on scores from raters conducting a relatively small number of evaluations (e.g., 20–40), effects still ranged noticeably among raters who scored 40 or more participants ($d = .19$ to 1.22 for Part 1, $.27$ to 1.54 for Part 2).

There are, however, several aspects of the MacArthur Study that prevent firm conclusions about the association between PCL train-

Table 5
PCL:SV Part 1 Cut Score Performance for Scores From Four Ratets With Similar Overall Predictive Effects

Cut score/Rater	Sensitivity	Specificity	Hit rate
Score ≥ 1			
31 ($M = 1.12$)	1.00	.44	.50
35 ($M = 1.93$)	1.00	.45	.52
33 ($M = 3.99$)	1.00	.17	.27
19 ($M = 4.15$)	1.00	.13	.35
Score ≥ 2			
31	.67	.83	.81
35	.86	.55	.59
33	.92	.30	.38
19	1.00	.40	.55
Score ≥ 3			
31	.67	.87	.85
35	.86	.67	.70
33	.92	.39	.46
19	1.00	.53	.65
Score ≥ 4			
31	.00	.91	.81
35	.43	.86	.80
33	.83	.52	.56
19	1.0	.60	.70
Score ≥ 5			
31	.00	1.00	.88
35	.29	.92	.84
33	.67	.62	.63
19	.80	.67	.70

Note. Base rates of future violence = .12 (Rater 31), .13 (Rater 35), .13 (Rater 33), and .25 (Rater 19).

ing and rater differences, especially about scoring in the field. On the one hand, the MacArthur study raters completed only one day of training and did not receive individualized feedback as part of the reliability check. Those scoring a PCL measure for other research studies often receive more intensive training and supervision, as do some field evaluators who attend the most rigorous training workshops. On the other hand, even with only a brief training, MacArthur study raters may have been more reliable than evaluators in the field who, over time, may develop more idiosyncratic scoring rules or tendencies. Experience might lead to greater reliability, but it might also lead to less reliability (i.e., rater drift) without ongoing monitoring.

Other reasons it is difficult to directly compare MacArthur study results with field results involve the differences in reliability analyses and reporting. The MacArthur study researchers did not provide an ICC value (i.e., a proportion of variance) for rater agreement, making it difficult to compare the level of rater agreement in the MacArthur study with the level of agreement in other studies. The MacArthur study researchers reported a kappa statistic based on evaluators' scores being within 5 points of one another. This approach treats very large differences (e.g., 12 point difference) identically to smaller ones (e.g., 6 point difference), and considers differences that are more than twice the standard error of measurement (reported to be about 2 points in the PCL:SV manual) to indicate agreement. Although smaller evaluator differences in the MacArthur study than field studies (e.g., Boccaccini et al., 2008, 2014) suggest that rater agreement was stronger in the MacArthur study, the MacArthur study kappa statistic does not allow for a direct comparison with studies reporting ICCs.

Rater Differences in Scoring and Rater Differences in Predictive Validity

The mean Facet 4 score that raters assigned across cases and the standard deviation of their Facet 4 scores did help explain why Facet 4 scores from some raters were more predictive than Facet 4 scores from others. Predictive effects increased as the rater mean and rater standard deviation increased, suggesting that Facet 4 scores from those willing to assign relatively high scores in some (but not all) cases were the best predictors of future violence. These effects highlight the potential impact of range restriction on predictive validity, especially the limited predictive utility of scores from raters who assign consistently low scores across cases.

But the mean score that raters assigned across cases and the standard deviation of their scores did not consistently explain a significant amount of variability in predictive effects. For example, there was no evidence of a significant association between rater scoring tendencies and predictive effects for the scores described in the PCL:SV manual (i.e., Total, Part 1, and Part 2). Indeed, there were instances in which scores from two raters with very similar Part 1 and 2 scoring tendencies across cases differed markedly with respect to predictive validity (e.g., Rater 4 and Rater 18). And there were other instances in which scores from raters with very different scoring tendencies both led to strong predictive effects (e.g., Rater 8 and Rater 25). In other words, it was possible for scores from one rater to be more accurate than scores from another (with respect to predicting future violence), even though the scores they assigned across cases tended to be the same.

These findings of similar predictive effects, but different scoring tendencies, have implications for score interpretation. As we documented in Table 5, sensitivity values for any Part 1 cut score were lower for low scoring raters than high scoring raters (assuming similar levels of overall predictive validity for each rater). In other words, the score from any rater that indicates high risk depends on the rater's scoring tendencies. A cut score of ≥ 4 accurately identified all of the patients who committed acts of future violence assessed by Rater 19 (i.e., all scored 4 or higher), but none of the patients who committed acts of future violence assessed by Rater 31 (i.e., all scored 3 or lower). In other words, a Part 1 score of 4 meant something very different when it was assigned by Rater 19 than when it was assigned by Rater 31.

Of course, we selected the four raters in Table 5 to highlight the possibility that scores from some raters may need to be interpreted differently than scores from other raters. We knew that scores from each of these raters were predictive of future violence (i.e., strong *d* values) and that some assigned much higher scores across cases than others. The pattern of lower sensitivity values for lower scoring raters would not have held, or would have been less consistent, if we included raters whose scores were weak predictors of future violence. At the same time, field research reveals instances when these types of scoring differences have real-world implications for score interpretation. Specifically, scores from some evaluators are clearly less predictive than scores from other evaluators (Murrie et al., 2012), and the PCL-R score that indicates high-risk among offenders evaluated for SVP commitment in Texas is about 5 points lower when it comes from a defense evaluator as opposed to a state evaluator (Boccaccini et al., 2012).

With respect to practice implications, these findings suggest that clinicians should at least consider the extent to which the scores they assign across cases are consistent with those in the normative samples described in the instrument's manual. If an instrument manual or relevant research study reports cut score statistics, the clinician should only rely on those statistics if his or her overall pattern of scores across cases is generally consistent with those in the normative sample. This does not necessarily mean that a clinician who assigns notably higher or lower scores across cases is assigning scores that are too low or too high. It could be that a clinician is scoring a select subset of offenders or patients with especially high or low levels of psychopathy. In these instances, the clinician should consider the extent to which those "select" offenders or patients are appropriately represented in the sample used to develop cut scores. If they are adequately represented, it may be reasonable to use the cut score statistics.

Ideally, clinicians would also know whether or not the scores they assign have similar levels of predictive validity compared with those used to calculate cut score statistics. In most field settings, clinicians do not have access to this information. In these situations, clinicians may benefit from considering the extent to which the scores they assign vary across cases. We found some evidence that predictive effects were stronger for raters who assigned a wider range of scores across cases. Although we only found evidence for this effect for Facet 4 scores, this finding is consistent with the well-known attenuating effect of range restriction on effect-size estimates (Bobko, Roth, & Bobko, 2001). Clinicians who assign very similar scores across cases are either evaluating offenders with remarkably similar levels of psycho-

pathic traits, or they are failing to adequately differentiate those with high and low levels of psychopathic traits.

Limitations

Although the MacArthur study dataset allowed for an examination of the generalizability of rater/clinician difference effects beyond Texas SVP evaluators, differences in study design limit the ability to draw firm conclusions about why there were smaller rater difference effects in the MacArthur study. Patients or offenders undergoing real-world forensic evaluations—which have real world consequences—may behave or respond to interview questions differently than those participating in a voluntary research study. It is also possible that the evaluation context may affect clinicians' and raters' interview styles and scoring tendencies. For example, some field evaluators may be more strongly influenced by pull of adversarial allegiance (see Murrie, Boccaccini, Guarnera, & Rufino, 2013) than others, which could lead to scoring tendencies that generalize across cases. Or, it may be that there is something unique to the PCL-R (as opposed to the PCL:SV) that allows for greater idiosyncratic clinician influences on scoring. Although each PCL:SV item corresponds with at least one PCL-R item, the scoring instructions are not identical and it may be that there is less room for subjectivity when scoring the PCL:SV.

One limitation that the MacArthur study shares with existing field studies is that participants were not randomly assigned to raters, raising the question of whether rater difference findings might be attributable to some raters being assigned to score more psychopathic patients than others. Although we attempted to control for this possibility in our analyses and there is no reason to suspect that case assignment in the MacArthur study or in field study settings was based on psychopathy, some of the variability we have attributed to raters may be attributable to true differences in psychopathy among the patients assigned to the rater. Of course, the ideal way to study clinician difference effects would be to have each clinician score the same large set of offenders or patients. This type of design is impractical for most studies because of the time it takes to complete each PCL-R or PCL:SV assessment, but it would allow researchers to clearly separate variance attributable to participant differences in psychopathy and variance attributable to clinician scoring tendencies (see Miller et al., 2011). Neither the MacArthur nor the Texas field studies allow for this type of partitioning of variance.

Finally, the MacArthur study dataset provided no information about rater characteristics that might explain variability in scoring (e.g., age, experience, demographic). There is some evidence that evaluator characteristics, such as evaluator agreeableness and prior experience, may help explain PCL-R scoring tendencies (see Miller et al., 2011; Rufino, Boccaccini, Hawes, & Murrie, 2012), but more research about evaluator characteristics is clearly needed.

Conclusion

The large number of patients and raters in the MacArthur study, along with the availability of patient outcomes, allowed for a more detailed examination of rater differences in psychopathy measure scoring than has been possible in prior studies (e.g., Boccaccini et al., 2008, 2014; Miller et al., 2011). Perhaps most importantly, the

MacArthur study allowed for an examination of whether field study findings generalize to a setting in which differences should be least likely to occur—a setting in which all raters completed the same PCL:SV training and passed a series of reliability checks. Although the size of the rater effects was smaller in this research setting than in field settings, the findings suggest that rater differences in PCL scoring will be present to some extent in most (if not all) contexts. With respect to predictive validity, rater differences appear to have important implications for score interpretation. We found that scores from raters with different scoring tendencies could be similarly predictive of future offending, but that the scores from each rater might need to be interpreted differently. Together, these findings suggest that rater differences likely impact psychopathy measure scores assigned by evaluators in both research and clinical practice, but that training regarding reliability and comparing scoring tendencies to norms in instrument manuals may be important mechanisms for reducing and managing these differences.

References

- Ake v. Oklahoma. 470 U.S. 68. (1986).
- Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods*, 4, 46–61. <http://dx.doi.org/10.1177/109442810141003>
- Boccaccini, M. T., Murrie, D. C., Rufino, K. A., & Gardner, B. O. (2014). Evaluator differences in Psychopathy Checklist-Revised factor and facet scores. *Law and Human Behavior*, 38, 337–345. <http://dx.doi.org/10.1037/lbh0000069>
- Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Findings from a statewide sample of sexually violent predators evaluations. *Psychology, Public Policy, and Law*, 14, 262–283. <http://dx.doi.org/10.1037/a0014523>
- Boccaccini, M. T., Turner, D. B., Murrie, D. C., & Rufino, K. A. (2012). Do PCL-R scores from state or defense experts best predict future misconduct among civilly committed sex offenders? *Law and Human Behavior*, 36, 159–169. <http://dx.doi.org/10.1037/h0093949>
- 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 behavioral features? *Behavioral Sciences & the Law*, 28, 106–119. <http://dx.doi.org/10.1002/bsl.918>
- Hare, R. D. (2003). *Psychopathy Checklist—Revised* (2nd ed.): technical manual. Toronto, Ontario, Canada: Multi-Health Systems.
- Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version (PCL:SV)*. Toronto, Ontario, Canada: Multi-Health Systems.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128. <http://dx.doi.org/10.2307/1164588>
- Laurell, J., Belfrage, H., & Hellström, A. (2010). Facets on the psychopathy checklist screening version and instrumental violence in forensic psychiatric patients. *Criminal Behaviour and Mental Health*, 20, 285–294. <http://dx.doi.org/10.1002/cbm.779>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Lloyd, C. D., Clark, H. J., & Forth, A. E. (2010). Psychopathy, expert testimony, and indeterminate sentences. Exploring the relationship between Psychopathy Checklist Revised testimony and trial outcome in Canada. *Legal and Criminological Psychology*, 15, 323–339. <http://dx.doi.org/10.1348/135532509X468432>

- MacArthur Research Network. (2005, September). *Macarthur violence risk assessment data*. Retrieved from <http://www.macarthur.virginia.edu/risk.html>
- McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment*, 26, 207–214. <http://dx.doi.org/10.1037/a0034832>
- 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. <http://dx.doi.org/10.1177/1073191111402460>
- 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. <http://dx.doi.org/10.1037/a0028411>
- Monahan, J., Steadman, H., Silver, E., Appelbaum, P., Robbins, P., Mulvey, E., . . . Banks, S. (2001). *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. New York, NY: New York: Oxford University Press.
- Morey, L. C. (1991). *Personality Assessment Inventory (PAI)*. Odessa, FL: Psychological Assessment Resources Inc.
- Murrie, D. C., Boccaccini, M. T., Caperton, J., & Rufino, K. (2012). Field validity of the Psychopathy Checklist-Revised in sex offender risk assessment. *Psychological Assessment*, 24, 524–529. <http://dx.doi.org/10.1037/a0026015>
- Murrie, D. C., Boccaccini, M. T., Guarnera, L., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, 24, 1889–1897. <http://dx.doi.org/10.1177/0956797613481812>
- Murrie, D. C., Boccaccini, M. T., Zapf, P. A., Warren, J. I., & Henderson, C. E. (2008). Clinician variation in findings of competence to stand trial. *Psychology, Public Policy, and Law*, 14, 177–193. <http://dx.doi.org/10.1037/a0013578>
- 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. <http://dx.doi.org/10.1037/0735-7028.36.5.519>
- Pedersen, L., Kunz, C., Rasmussen, K., & Elsass, P. (2010). Psychopathy as a risk factor for violent recidivism: Investigating the Psychopathy Checklist Screening Version (PCL:SV) and the Comprehensive Assessment of Psychopathic Personality (CAPP) in a forensic psychiatry setting. *The International Journal of Forensic Mental Health*, 9, 308–315. <http://dx.doi.org/10.1080/14999013.2010.526681>
- Rufino, K. A., Boccaccini, M. T., & Guy, L. S. (2011). Scoring subjectivity and item performance on measures used to assess violence risk: The PCL-R and HCR-20 as exemplars. *Assessment*, 18, 453–463. <http://dx.doi.org/10.1177/10731911110378482>
- 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. <http://dx.doi.org/10.1037/h0093988>
- Skeem, J. L., & Mulvey, E. P. (2001). Psychopathy and community violence among civil psychiatric patients: Results from the MacArthur Violence Risk Assessment Study. *Journal of Consulting and Clinical Psychology*, 69, 358–374. <http://dx.doi.org/10.1037/0022-006X.69.3.358>
- 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. <http://dx.doi.org/10.1037/lhb0000063>
- Walters, G. D., Knight, R. A., Grann, M., & Dahle, K. P. (2008). Incremental validity of the Psychopathy Checklist facet scores: Predicting release outcome in six samples. *Journal of Abnormal Psychology*, 117, 396–405. <http://dx.doi.org/10.1037/0021-843X.117.2.396>
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.
- Zapf, P. A., Hubbard, K. L., Cooper, V. G., Wheelles, M. C., & Ronan, K. A. (2004). Have the courts abdicated their responsibility for determination of competency to stand trial to clinicians? *Journal of Forensic Psychology Practice*, 4, 27–44. http://dx.doi.org/10.1300/J158v04n01_02

Received July 3, 2014

Revision received November 4, 2014

Accepted November 4, 2014 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!