

Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist-Revised in Capital Sentencing to Assess Risk for Institutional Violence

David DeMatteo
Drexel University

Stephen D. Hart
Simon Fraser University

Kirk Heilbrun
Drexel University

Marcus T. Boccaccini
Sam Houston State University

Mark D. Cunningham
Seattle, Washington

Kevin S. Douglas
Simon Fraser University

Joel A. Dvoskin
University of Arizona College of Medicine

John F. Edens
Texas A&M University

Laura S. Guy
Simon Fraser University

Daniel C. Murrie
University of Virginia







Randy K. Otto
University of South Florida

Ira K. Packer
University of Massachusetts Medical School

Thomas J. Reidy
Monterey, California

Psychopathy as measured by the Hare Psychopathy Checklist—Revised (PCL–R; Hare, 1991, 2003) is related to a range of rule-breaking and antisocial behaviors. Given this association, psychopathy has received considerable attention from researchers and legal professionals over the past several decades. Concerns remain, however, about using PCL–R scores to make precise and accurate predictions in certain contexts, including an individual’s risk for committing serious violence in high-security custodial facilities. After a brief introduction to psychopathy and the PCL–R, we discuss capital sentencing in the United States and then summarize the empirical literature regarding the ability of PCL–R scores to predict violence, with a particular focus on the PCL–R’s ability to predict serious institutional violence. As described, we believe the research demonstrates that the PCL–R cannot precisely or accurately predict

This article was published Online First January 30, 2020.

 David DeMatteo, Department of Psychology & Thomas R. Kline School of Law, Drexel University;  Stephen D. Hart, Department of Psychology, Simon Fraser University;  Kirk Heilbrun, Department of Psychology, Drexel University; Marcus T. Boccaccini, Department of Psychology and Philosophy, Sam Houston State University; Mark D. Cunningham, Private Practice, Seattle, Washington; Kevin S. Douglas, Department of Psychology, Simon Fraser University; Joel A. Dvoskin, Department of Psychiatry, University of Arizona College of Medicine;  John F. Edens, Department of Psychological & Brain Sciences, Texas A&M University; Laura S. Guy, Department of Psychology, Simon Fraser University; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia; Randy K. Otto, Department of Mental Health Law & Policy, University of South Florida;  Ira K. Packer, Department of Psychiatry, University of Massachusetts Medical School;  Thomas J. Reidy, Private Practice, Monterey, California.

The authors are presented in alphabetical order following the third author. The Statement presented in the [Appendix](#) of this article represents the consensus of our views as individual forensic mental health professionals; it does not necessarily reflect the views of this journal’s Editorial Board and publisher, the American Psychological Association, or any other agencies or organizations with which we are affiliated or for which we work. As any statement derived by consensus reflects compromise in the choice of language, each member comprising the Group of Concerned Forensic Mental Health Professionals may have preferred manner of expressing the findings and opinions that differs from that in the Statement in nuance or detail. We thank Kellie Wiltsie for providing research assistance for this article.

Correspondence concerning this article should be addressed to David DeMatteo, Department of Psychology, Drexel University, 3141 Chestnut Street, Stratton Suite 119, Philadelphia, PA 19104. E-mail: david.dematteo@drexel.edu

an individual's risk for committing serious violence in high-security custodial facilities. Finally, we present a Statement of Concerned Experts that summarizes our findings and opinions, concluding the PCL-R cannot and should not be used to make predictions that an individual will engage in serious institutional violence with any reasonable degree of precision or accuracy, especially when making high-stakes decisions about legal issues such as capital sentencing.

Keywords: psychopathy, Psychopathy Checklist—Revised, violence risk, institutional violence, capital sentencing

Supplemental materials: <http://dx.doi.org/10.1037/law0000223.supp>

There is an essential tension that underlies the study of psychopathy in forensic mental health. On one hand, it continues to garner considerable attention in the scientific and professional literature, and there is a large body of work indicating that the construct is related to a broad range of adverse behavioral outcomes, including antisocial and criminal conduct. Much of this literature focuses on the use of the Hare Psychopathy Checklist—Revised (PCL-R; Hare, 1991, 2003), a psychological rating scale that is often identified as the “gold standard” for the assessment of psychopathy. Perhaps the best summary of the literature to date is that symptoms of psychopathy, as assessed using the PCL-R, are associated with general rule-breaking and trouble-making across settings and populations. But serious concerns have been expressed about the limitations of using both research on psychopathy and scores on the PCL-R to make reliable (i.e., consistent) and accurate (i.e., valid) predictions. This is especially true when those predictions concern specific individuals, specific populations, specific antisocial acts, and specific settings or are made with the goal of assisting decisions about specific legal issues.

It may strike some people as confusing or even logically incoherent to conclude that the scientific and professional literature can, simultaneously, support the general usefulness of psychopathy as a construct but fail to support the use of psychopathy rating scales by forensic mental health professionals to make certain predictions of violence. Yet, that is exactly what we—a group of concerned forensic mental health professionals—believe to be true and exactly what motivated us to prepare the Statement of Concerned Experts (“Statement”) presented in this article, which focuses on what we consider to be the inappropriate use of the PCL-R to draw conclusions about an individual's risk for committing serious violence in high-security custodial facilities. We conclude that the literature does not support the use of the PCL-R to predict serious institutional violence. Our interpretation of the research literature is that not only is there an absence of proof it can do so, but that the literature demonstrates it cannot do so precisely or accurately; that is, there is “proof of absence” of such an association. This conclusion has important real-world implications because PCL-R scores are sometimes offered in capital sentencing evaluations to draw conclusions regarding an offender's “future dangerousness” in the sense of risk for serious institutional violence. Not only do PCL-R scores lack probative value with respect to determining risk for serious institutional violence, there is compelling evidence to suggest that characterizing defendants as “psychopaths” has a substantial prejudicial impact that may make jurors more inclined to support the death penalty for them (Kelley, Edens, Mowle, Penson, & Rulseh, 2019). Quite simply, the question of whether or how much to rely on the PCL-R

to assess risk for serious institutional violence may be a matter of life or death.

We begin this article with a brief introduction to the concept of psychopathy and the PCL-R. We then move on to discuss the relevance of risk for serious institutional violence to capital sentencing decisions and summarize what is known and what is not known with respect to the use of the PCL-R to make precise and accurate predictions of serious institutional violence. Finally, we present the full Statement that summarizes the available scientific literature and ends by concluding the PCL-R cannot make predictions that an individual will engage in serious institutional violence with any reasonable degree of precision or accuracy and should not be used for this purpose in capital sentencing evaluations

Psychopathy and the PCL-R

The disorder currently known as psychopathy has been recognized by various names for hundreds of years, but the conceptualization of psychopathy historically included a wide range of poorly defined and conceptually inconsistent traits (see Millon, Simonsen, & Birket-Smith, 1998). However, the publication of several seminal books and articles on psychopathy in the 1940s, including Cleckley's (1941) *The Mask of Sanity* and Karpman's (1946, 1948) description of primary psychopathy, marked a shift in our understanding of the disorder. As conceptualized by Cleckley, Karpman, and others who followed them, *psychopathy* refers to a distinct constellation of interpersonal, affective, and behavioral personality traits that are extreme and maladaptive, including egocentricity, lack of empathy, shallow affect, impulsivity, and a tendency to violate social norms (Hare & Neumann, 2009).

Since the 1980s, the construct of psychopathy often has been operationalized using instruments developed by Robert Hare and colleagues, including the Psychopathy Checklist (PCL; Hare, 1980), later revised and eventually commercially published as the Hare Psychopathy Checklist—Revised (PCL-R; Hare, 1991, 2003). The Hare scales (as they are sometimes referred to) appear to reflect the interpersonal and affective characteristics of the disorder highlighted by Cleckley (1941), Karpman (1946, 1948), and others better than do many other commonly used psychological tests and diagnostic criteria. We focus on the PCL-R, as it is the Hare scale that is most widely researched and most commonly used in practice by forensic mental health professionals around the world, and also because it formed the basis for the development of other rating scales, among them the Screening Version and Youth

Version of the PCL-R (PCL:SV and PCL:YV, respectively; Hart, Cox, & Hare, 1995; Forth, Kosson, & Hare, 2003).

The PCL-R is a 20-item symptom construct rating scale for the assessment of psychopathy in adult correctional offenders and forensic mental health patients (Hare, 2003). Each of the 20 items reflects a different (putative) feature or characteristic of psychopathy. Standard administration of the PCL-R includes a semistructured interview and a review of collateral records, and on the basis of this information evaluators rate the lifetime presence of each feature using a 3-point ordinal scale (briefly, 0 = *item does not apply to the individual*, 1 = *item applies to a certain extent*, 2 = *item applies*). Scores on individual items can be summed to form various composites, the most commonly used of which is total score, reflecting the unit-weighted sum of all 20 items. Total score ranges from 0 to 40, with higher scores indicating higher levels of psychopathy. Scores of 30 and higher are frequently considered diagnostic of psychopathy, although the PCL-R manual (Hare, 2003) makes clear that this is a cutoff of convenience. Although the general descriptive utility of this particular cutoff is supported by research (e.g., Hare, 1991), it is nevertheless arbitrary. There is no good theoretical or empirical basis for assuming that psychopathy forms a natural taxon; indeed, most research tends to support the view that psychopathy is most parsimoniously and usefully conceptualized in dimensional rather than categorical terms (e.g., Edens, Marcus, Lilienfeld, & Poythress, 2006). Also, there is remarkable heterogeneity—both potential and actual—among those at or above the PCL-R cutoff score of 30 and higher (e.g., Balsis, Busch, Wilfong, Newman, & Edens, 2017; see also Mokros et al., 2015; Poythress et al., 2010).

One strength of PCL-R scores is their high level of interrater reliability (i.e., agreement among independent evaluators with respect to PCL-R scores) reported in professional manuals and in many published research reports. Many studies have found that well-trained evaluators in controlled research contexts produce scores with high levels of interrater reliability and, consequently, a small standard error of measurement (“margin of error”) with respect to the expected level of disagreement between raters (see DeMatteo, Murrie, Edens, & Lankford, 2019, for a review). The PCL-R manual (Hare, 2003) reports the following intraclass correlation coefficients (ICCs): the pooled ICC for male criminal offenders was .86 for a single rating (ICC_1) and .92 for the average of two ratings (ICC_2); ICC_1 was .88 and ICC_2 was .93 for the male forensic psychiatric patients; and ICC_1 was .94 and ICC_2 was .97 for the female criminal offenders, with these values suggesting acceptable reliability (Nunnally & Bernstein, 1994).

But interrater reliability is a property of scores obtained for a particular sample of people and in a particular context; it is not a stable property of the test itself that necessarily generalizes across samples or contexts. Research conducted over the past 10 to 15 years raises concerns about the interrater reliability of PCL-R scores made in psycholegal contexts, and several caselaw reviews have examined the interrater reliability of PCL-R scores in court cases. For example, in their review of United States sexually violent predator (SVP) cases involving use of the PCL-R, DeMatteo et al. (2014a) identified 29 cases in which the same offender was assessed with the PCL-R by two evaluators. In those 29 cases, the ICC_1 was .58, and only 41% of

the score differences were within one standard error of measurement (*SEM*). Further, scores by prosecutor-retained experts were significantly higher than the scores produced by defense-retained experts; prosecution experts reported PCL-R scores of 30 or above in nearly 50% of the cases, compared with less than 10% of the same cases appraised by defense experts. In a caselaw survey that included 102 criminal cases from Canada, the single-rater ICC was .59 for all cases, with an ICC of .66 for cases involving a sexual offense and an ICC .46 for nonsexual offense cases (Edens, Cox, Smith, DeMatteo, & Sörman, 2015).

From a practical perspective, it is useful to note that if the ICC for PCL-R ratings in adversarial legal proceedings do in fact approximate .60 as suggested above, then the corresponding 95% confidence interval around an average PCL score would fall between the 11th and 89th percentiles (Edens & Boccaccini, 2017). This analysis ignores certain important qualifiers, such as the fact that the PCL-R normative data are not normally distributed and that reliability estimates are not constant across the range of possible test score results (i.e., they tend to decrease the further away an obtained score is from the mean), which may further reduce the expected agreement among raters (Cooke & Michie, 2010).

Taken together, these results reveal two things. First, there is a tendency for examiners in adversarial settings to disagree with each other to an extent that is much greater than would be expected based on the ICC values reported in the PCL-R professional manual (Hare, 2003). Second, there is a tendency for prosecution-retained evaluators to report higher PCL-R scores than do defense-retained evaluators in evaluations of the same person, made around the same time, and even when made on the same information base. This tendency for some experts to drift from more objective findings to ratings that better support the party that retained them has been termed *adversarial allegiance* (Murrie & Boccaccini, 2015). Adversarial allegiance has been examined in both field studies and controlled research. In the first field study to examine adversarial allegiance, Murrie, Boccaccini, Johnson, and Janke (2008) collected PCL-R scores assigned by petitioner-retained¹ and respondent-retained psychologists in 23 SVP cases in Texas; these cases permitted the examination of PCL-R scores that opposing evaluators assigned to the same offender. There was a large difference between PCL-R scores assigned by petitioner-retained and respondent-retained evaluators (Cohen’s $d = 1.03$) that reflected a low level of interrater agreement across raters ($ICC = .39$). In 14 of the 23 cases (61%), there was a difference of more than 6.0 points between the two PCL-R total scores; given the *SEM* of roughly 3.0 points for PCL-R scores, differences of this magnitude should occur by chance in less than 5% of cases. In each case, the petitioner-retained evaluator assigned a higher score than the respondent-retained evaluator. A follow-up study that included 35 SVP cases revealed similar allegiance effects in PCL-R scoring (Murrie et al., 2009).

Although the results of field studies suggest the presence of adversarial allegiance in PCL-R scoring, the nature of field

¹ As civil proceedings, SVP hearings use slightly different terminology than criminal proceedings. The petitioner, which is the party seeking civil commitment of the offender, is roughly analogous to the prosecution in criminal proceedings, whereas the respondent is roughly analogous to the defendant in criminal proceedings.

studies does not permit alternative explanations for the observed results to be ruled out. It is possible, for example, that the observed pattern in PCL–R scoring in the field studies may be due to savvy attorneys selecting experts who are most favorable to their perspective on the case. It is also possible that the nature of caselaw reviews, which comprise only published cases, is contributing to the appearance of adversarial allegiance. In other words, contentious cases are more likely to go to trial, whereas the large majority of cases that never went to trial may have involved similar PCL–R scores by prosecution-retained and defense-retained experts. Finally, it is possible that one unreliable PCL–R score may lead the parties to reach a plea bargain instead of proceeding to trial, thereby making the case unavailable for research purposes.

Fortunately, some experimental research (which does not have the same limitations as field studies) has examined adversarial allegiance in PCL–R scoring. Murrie, Boccaccini, Guarnera, and Rufino (2013) recruited more than 100 forensic psychologists and psychiatrists under the guise of performing a forensic consultation. These forensic mental health professionals were (without their awareness) randomly assigned to either a prosecution-allegiance or defense-allegiance group. Participants met for 10 to 15 minutes with an attorney who posed as leading either a public defender service or specialized prosecution unit, and the attorney then requested that the expert score two tests, one of which was the PCL–R, based on extensive offender records. Each participant was scoring the same four case files that spanned from low risk to high risk. As hypothesized, the PCL–R scores assigned by prosecution experts and defense experts showed evidence of adversarial allegiance. On average, prosecution evaluators assigned significantly higher PCL–R scores than did defense evaluators for three of four cases, with effect sizes in the medium to large range (Cohen's *d* of .55 to .85). Follow-up analyses examined how likely it was that a randomly selected prosecution expert and a randomly selected defense expert would assign scores that were so different that they could not be explained by random measurement error. Results revealed that more than 20% of the score pairings for each case reflected a score difference that was more than twice the *SEM* in the PCL–R manual. Further, most large (≥ 2 *SEM*) differences were in the direction of adversarial allegiance, with the prosecution expert assigning higher scores and the defense expert assigning lower scores (Murrie et al., 2013).

Capital Sentencing in the United States and the Issue of Risk for Serious Institutional Violence

Capital sentencing is the process by which criminal offenders are sentenced to death or life in prison after being convicted of a capital offense. The Supreme Court of the United States has provided many substantive and procedural constitutional restrictions on imposing the death penalty. Among other rulings, the Supreme Court has held that the death penalty (a) cannot be mandatorily imposed (*Roberts v. Louisiana*, 1976); (b) can only be imposed for crimes involving death (*Kennedy v. Louisiana*, 2008); (c) cannot be imposed on individuals who were juveniles at the time of the offense (*Roper v. Simmons*, 2005), individuals who are intellectually disabled (*Atkins v. Virginia*, 2002), or individuals who are not competent to be executed (*Ford v. Wainwright*, 1986; *Panetti v.*

Quarterman, 2007); (d) requires a jury to reach findings of fact concerning aggravating factors (*Ring v. Arizona*, 2002); and (e) must be based on individualized consideration of each crime and defendant (*Eddings v. Oklahoma*, 1982; *Lockett v. Ohio*, 1978).

In several death penalty decisions dating back to the reinstatement of capital punishment in 1976, the Supreme Court has held that the sentencing jury must be given guidance in deciding whether death is an appropriate punishment (e.g., *Gregg v. Georgia*, 1976). The guidance provided to sentencing juries takes the form of statutorily defined aggravating factors (which support the imposition of the death penalty) and a nonexhaustive statutory list of mitigating factors (which support the imposition of life in prison). Aggravating factors, which are intended to narrow the class of offenders for whom death is appropriate, pertain to the offense and offender (e.g., murdering certain classes of people, committing murder in the course of a felony, an offender's history of prior violent felonies), whereas mitigating factors can be anything that is relevant to the determination of whether death is an appropriate sentence. One aggravating factor outlined by some states is a capital defendant's risk of future danger (see DeMatteo, Murrie, Anumba, & Keesler, 2011; Fairfax-Columbo & DeMatteo, 2017).

In capital sentencing contexts, future dangerousness is the probability that an individual, absent a penalty of death, will engage in future violent behavior. Currently, of the 29 states that have the death penalty, three states (Oregon, Texas, and Virginia) explicitly require that sentencing juries consider future dangerousness as an aggravating factor, three states (Idaho, Oklahoma, Wyoming) explicitly permit consideration of future dangerousness as an aggravating factor, 12 states (Alabama, Arkansas, California, Colorado, Georgia, Kentucky, Missouri, North Carolina, Pennsylvania, South Carolina, South Dakota, Utah) permit consideration of future dangerousness as a non-statutory aggravating factor, and six states (Florida, Indiana, Kansas, Mississippi, Ohio, Tennessee) prohibit consideration of future dangerousness as an aggravating factor, with the remaining five states (Louisiana, Montana, Nebraska, Nevada, New Hampshire) making no mention of future dangerousness in the death penalty statute. In many death penalty jurisdictions, considering risk for future violence is quite commonplace, with research suggesting that future dangerousness often plays a prominent role in capital sentencing contexts (e.g., Cunningham & Goldstein, 2003; Cunningham & Reidy, 1999; Shapiro, 2009).

When considering the role of future dangerousness in capital sentencing proceedings, it is important to frame the question properly. As noted, at the capital sentencing stage, the jury usually deliberates between sentencing the defendant to death or life in prison; in most cases, release to the community is not an option, at least in the foreseeable future and barring unforeseen circumstances.² Therefore, as noted by a number of researchers and scholars, questions about violence risk in capital cases primarily

² A few states impose capital sentences with the possibility of release or parole in the distant future. As such, in these cases, forensic mental health professionals may be asked to opine about the defendant's risk for violence if and when the offender is released to the community many years in the future. These circumstances are rare, however. Most violence risk assessments in capital sentencing proceedings focus on risk of violence in the prison context, and it is difficult to imagine a scenario in which a violence risk assessment for capital sentencing would address risk of violence in the community in the near future.

involve whether the defendant will be violent while incarcerated in a high-security correctional facility (e.g., Cunningham, 2006, 2008; DeMatteo et al., 2011; Edens, Buffington-Vollum, Keilen, Roskamp, & Anthony, 2005). Many courts have explicitly recognized that violence risk assessments in capital cases are specific to the prison context (e.g., *United States v. Sablan*, 2006), although some have taken a broad and amorphous view of what it means to be a potential “danger to society” (e.g., *Coble v. Texas*, 2010). In this article, we are focusing specifically on the use of the PCL-R to predict serious (i.e., nontrivial) violence in high-security correctional settings. It is generally accepted in the field of forensic mental health that violence risk is—and therefore violence risk assessment must be—context specific (e.g., Conroy & Murrie, 2007; Heilbrun, 1992, 2009).

As this review makes clear, future dangerousness may be a relevant consideration in capital sentencing evaluations. The PCL-R has been used to assess psychopathy as a risk factor for future violence in such evaluations (e.g., *Busby v. Stephens*, 2015; *Martinez v. Dretke*, 2004; *United States v. Barnette*, 2000; *United States v. Fell*, 2008). Accordingly, it is essential to evaluate the predictive validity of PCL-R scores, and in particular to evaluate its predictive validity with respect to serious institutional violence. In the following section, we turn to this issue.

Predictive Validity of the PCL-R

As noted previously, a well-developed body of research suggests that psychopathy is related to several outcomes that are of considerable interest to the criminal justice system. PCL-R scores are associated with diverse forms of antisocial and criminal behavior in diverse settings and populations (see Patrick, 2018, for a review). As a result, researchers, clinicians, and legal professionals are attentive to psychopathy in a variety of legal contexts. Research suggests that psychopathy evidence, typically in the form of PCL-R scores, is offered in legal proceedings in the United States, the United Kingdom, and Canada (DeMatteo et al., 2014b; Gagnon, Douglas, & DeMatteo, 2007; Howard, Khalifa, Duggan, & Lumsden, 2012). PCL-R scores may be of use in some psycho-legal evaluations when considered as part of a comprehensive, individualized, and contextualized evaluation. But because of their imperfect interrater reliability (which is, of course, a concern in any evaluation) and variability in their predictive validity across outcomes, settings, and samples (which is a concern with respect to prediction of serious institutional violence), PCL-R scores may lack probative value or, worse, have a prejudicial impact. (For a fuller discussion of the potential prejudicial impact of PCL-R scores, see DeMatteo, Hodges, & Fairfax-Columbo, 2016, and DeMatteo et al., 2019.)

The predictive validity (accuracy) of PCL-R scores with respect to general institutional misconduct has been studied for many years. Some early retrospective studies provided evidence of an association between PCL-R scores and past institutional misconduct. However, to the extent that PCL-R scores could have been biased or contaminated by the violence history of people being evaluated, such research is of little or no value in evaluating predictive accuracy. Later studies specifically examined the ability of PCL measure scores to predict institutional misconduct using true prospective research designs. In meta-analyses of such studies, Walters (2003a) reported a moderate association between

PCL-R Total scores and institutional adjustment, including both violent and nonviolent institutional conduct ($r_w = 0.27$), and small ($r_w = 0.18$) to moderate ($r_w = .27$) associations between PCL-R Factor 1 and 2 scores, respectively, for violent and nonviolent infractions (Walters, 2003b). Still, Walters (2003a, 2003b) did not distinguish between more serious institutional violence and other infractions.

In a large meta-analysis of published and unpublished studies, Guy, Edens, Anthony, and Douglas (2005) coded 273 effect sizes to examine the association between PCL, PCL-R, and PCL:SV scores and institutional misconduct in civil psychiatric, forensic psychiatric, and correctional facilities. Importantly, they were able to specifically analyze the association with more serious institutional violence, in this case, physical violence (i.e., any actual or attempted physical harm). The association between total scores and physical violence was small ($r_w = .17$)—indeed, much smaller than the typical violence risk assessment meta-analytic effect sizes, which are best described as moderate in size ($r_s \cong .30$ – $.35$; see Campbell, French, & Gendreau, 2009; Fazel, Singh, Doll, & Grann, 2012; Singh, Grann, & Fazel, 2011; for a review of risk assessment meta-analyses, see Douglas, 2019).

A few studies published after the Guy et al. (2005) meta-analysis found that PCL measures predict institutional misconduct (e.g., Huchzermeier, Bruss, Geiger, Kernbichler, & Aldenhoff, 2008), but most studies have reported similarly weak effects (e.g., Camp, Skeem, Barchard, Lilienfeld, & Poythress, 2013; Hogan & Olver, 2016; McDermott, Edens, Quanbeck, Busse, & Scott, 2008; Morrissey et al., 2007; Walters & Mandell, 2007). It should also be noted that the rate of serious institutional violence among capital murderers sentenced to death is very low (e.g., Cunningham, Reidy, & Sorensen, 2005; Sorensen & Wrinkle, 1996), which would tend to further reduce the predictive validity of the PCL-R.

Conclusion

Two major findings emerged from our review of the literature, summarized above. First, the interrater reliability of PCL-R scores in field settings, and in particular in adversarial contexts, is problematically low. Second, the overall association between PCL-R scores and violence at the group level is only moderate in terms of effect size, both in absolute terms and relative to the effect size of other established risk factors for violence; the association between PCL-R scores and violence in institutional settings is small in terms of effect size; and the association between PCL-R scores and serious institutional violence is negligible. Our conclusion based on these findings was that one cannot use the PCL-R in the context of capital sentencing evaluations to make predictions that an individual will engage in serious violence in high-security institutional settings with adequate precision or accuracy to justify reliance on the PCL-R scores.

Accordingly, we established a Group of Concerned Forensic Mental Health Professionals and developed a Statement to summarize our findings and opinions in this respect (see the Appendix and the online supplemental materials). Our goal in developing and disseminating the Statement was to educate others concerning the current state of the scientific literature and the appropriate use of the PCL-R when making capital sentencing and other high-stakes decisions. We emphasize that although this Statement focuses on the PCL-R, this is only because it is the instrument most widely

used to assess psychopathy in forensic mental health contexts; all our concerns about relying on the PCL–R to predict whether an individual will commit serious institutional violence apply equally or to an even greater degree to the use of other means of assessing psychopathy for that purpose.

References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Press.
- Atkins v. Virginia, 536 U.S. 304 (2002).
- Balsis, S., Busch, A. J., Wilfong, K. M., Newman, J. W., & Edens, J. F. (2017). A statistical consideration regarding the threshold of the Psychopathy Checklist—Revised. *Journal of Personality Assessment*, *99*, 494–502. <http://dx.doi.org/10.1080/00223891.2017.1281819>
- 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 predator evaluations. *Psychology, Public Policy, and Law*, *14*, 262–283. <http://dx.doi.org/10.1037/a0014523>
- Busby v. Stephens, 2015 WL 1037460 (N. D. Tex. Mar. 10, 2015).
- Camp, J. P., Skeem, J. L., Barchard, K., Lilienfeld, S. O., & Poythress, N. G. (2013). Psychopathic predators? Getting specific about the relation between psychopathy and violence. *Journal of Consulting and Clinical Psychology*, *81*, 467–480. <http://dx.doi.org/10.1037/a0031349>
- Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior*, *36*, 567–590. <http://dx.doi.org/10.1177/0093854809333610>
- Cleckley, H. (1941). *The mask of sanity*. St. Louis, MO: Mosby.
- Coble v. Texas, 330 S. W. 3d 253 (Tx. Ct. App. 2010).
- Conroy, M. E., & Murrie, D. C. (2007). *Forensic assessment of violence risk: A guide for risk assessment and risk management*. Hoboken, NJ: Wiley. <http://dx.doi.org/10.1002/9781118269671>
- Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*, 259–274. <http://dx.doi.org/10.1007/s10979-009-9176-x>
- Cunningham, M. D. (2006). Dangerousness and death: A nexus in search of science and reason. *American Psychologist*, *61*, 828–839. <http://dx.doi.org/10.1037/0003-066X.61.8.828>
- Cunningham, M. D. (2008). Forensic psychology evaluations at capital sentencing. In R. Jackson (Ed.), *Learning forensic assessment* (pp. 211–238). New York, NY: Routledge/Taylor & Francis Group.
- Cunningham, M. D., & Goldstein, A. M. (2003). Sentencing determinations in death penalty cases. In A. M. Goldstein & I. B. Weiner (Eds.), *Handbook of psychology: Vol. 11. Forensic psychology* (pp. 407–436). Hoboken, NJ: Wiley.
- Cunningham, M. D., & Reidy, T. J. (1999). Don't confuse me with the facts: Common errors in violence risk assessment at capital sentencing. *Criminal Justice and Behavior*, *26*, 20–43. <http://dx.doi.org/10.1177/0093854899026001002>
- Cunningham, M. D., Reidy, T. J., & Sorensen, J. R. (2005). Is death row obsolete? A decade of mainstreaming death-sentenced inmates in Missouri. *Behavioral Sciences & the Law*, *23*, 307–320. <http://dx.doi.org/10.1002/bsl.608>
- DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., & Formon, D. (2014a). The role and reliability of the Psychopathy Checklist—Revised in U.S. sexually violent predator evaluations: A case law survey. *Law and Human Behavior*, *38*, 248–255. <http://dx.doi.org/10.1037/lhb0000059>
- DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., Koller, J. P., & Bersoff, B. (2014b). Investigating the role of the Psychopathy Checklist—Revised in United States case law. *Psychology, Public Policy, and Law*, *20*, 96–107. <http://dx.doi.org/10.1037/a0035452>
- DeMatteo, D., Hodges, H., & Fairfax-Columbo, J. (2016). An examination of whether Psychopathy Checklist–Revised (PCL–R) evidence satisfies the relevance/prejudice admissibility standard. In B. H. Bornstein & M. K. Miller (Eds.), *Advances in psychology and law* (Vol. 2, pp. 205–239). New York, NY: Springer. http://dx.doi.org/10.1007/978-3-319-43083-6_7
- DeMatteo, D., Murrie, D. C., Anumba, N. M., & Keesler, M. E. (2011). *Forensic mental health assessments in death penalty cases*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195385809.001.0001>
- DeMatteo, D., Murrie, D. C., Edens, J. F., & Lankford, C. (2019). Psychopathy in the courts. In M. DeLisi (Ed.), *Routledge international handbook of psychopathy and crime* (pp. 645–664). New York, NY: Routledge/Taylor & Francis Group.
- Douglas, K. S. (2019). Evaluating and managing risk for violence using structured professional judgment. In A. Day, C. Hollin, & D. Polaschek (Eds.), *Handbook of correctional psychology* (pp. 427–445). Hoboken, NJ: Wiley. <http://dx.doi.org/10.1002/9781119139980.ch26>
- Eddings v. Oklahoma, 455 U.S. 104 (1982).
- Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment “out of the lab” and into “the real world”: Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment*, *29*, 599–610. <http://dx.doi.org/10.1037/pas0000475>
- 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>
- Edens, J. F., Buffington-Vollum, J. K., Keilen, A., Roskamp, P., & Anthony, C. (2005). Predictions of future dangerousness in capital murder trials: Is it time to “disinvent the wheel”? *Law and Human Behavior*, *29*, 55–86. <http://dx.doi.org/10.1007/s10979-005-1399-x>
- Edens, J. F., Cox, J., Smith, S. T., DeMatteo, D., & Sörman, K. (2015). How reliable are Psychopathy Checklist–Revised scores in Canadian criminal trials? A case law review. *Psychological Assessment*, *27*, 447–456. <http://dx.doi.org/10.1037/pas0000048>
- Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G., Jr. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology*, *115*, 131–144. <http://dx.doi.org/10.1037/0021-843X.115.1.131>
- Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review*, *81*, 417–480.
- Fairfax-Columbo, J., & DeMatteo, D. (2017). Reducing the dangers of future dangerousness testimony: Applying the Federal Rules of Evidence to capital sentencing. *The William and Mary Bill of Rights Journal*, *25*, 1047–1072.
- Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ: British Medical Journal*, *345*, e4692. <http://dx.doi.org/10.1136/bmj.e4692>
- Ford v. Wainwright, 477 U.S. 399 (1986).
- Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version*. Toronto, Canada: Multi-Health Systems.
- Gagnon, N., Douglas, K., & DeMatteo, D. (2007, June). *The introduction of the Psychopathy Checklist—Revised in Canadian courts: Uses and misuses*. Paper presented at the 7th Annual Conference of the International Association of Forensic Mental Health Services, Montreal, Quebec, Canada.
- Gregg v. Georgia, 428 U.S. 153 (1976).

- Guy, L. S., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting and Clinical Psychology, 73*, 1056–1064. <http://dx.doi.org/10.1037/0022-006X.73.6.1056>
- Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences, 1*, 111–119. [http://dx.doi.org/10.1016/0191-8869\(80\)90028-8](http://dx.doi.org/10.1016/0191-8869(80)90028-8)
- Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised manual*. North Tonawanda, NY: Multi-Health Systems.
- Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised manual* (2nd ed.). North Tonawanda, NY: Multi-Health Systems.
- Hare, R. D., & Neumann, C. S. (2009). Psychopathy: Assessment and forensic implications. *Canadian Journal of Psychiatry, 54*, 791–802. <http://dx.doi.org/10.1177/070674370905401202>
- Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences & the Law, 31*, 81–102. <http://dx.doi.org/10.1002/bsl.2049>
- Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare PCL: Screening version*. North Tonawanda, NY: Multi-Health Systems.
- Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the ‘margins of error’ of group v. individual predictions of violence. *The British Journal of Psychiatry, 190*(S49), S60–S65. <http://dx.doi.org/10.1192/bjp.190.5.s60>
- Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257–272. <http://dx.doi.org/10.1007/BF01044769>
- Heilbrun, K. (2009). *Evaluation for risk of violence in adults*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/med:psych/9780195369816.001.0001>
- Hogan, N. R., & Olver, M. E. (2016). Assessing risk for aggression in forensic psychiatric inpatients: An examination of five measures. *Law and Human Behavior, 40*, 233–243. <http://dx.doi.org/10.1037/lhb0000179>
- Howard, R., Khalifa, N., Duggan, C., & Lumsden, J. (2012). Are patients deemed ‘dangerous and severely personality disordered’ different from other personality disordered patients detained in forensic settings? *Criminal Behaviour and Mental Health, 22*, 65–78. <http://dx.doi.org/10.1002/cbm.827>
- Huchzermeier, C., Bruss, E., Geiger, F., Kernbichler, A., & Aldenhoff, J. (2008). Predictive validity of the psychopathy checklist: Screening version for intramural behaviour in violent offenders—a prospective study at a secure psychiatric hospital in Germany. *The Canadian Journal of Psychiatry / La Revue canadienne de psychiatrie, 53*, 384–391. <http://dx.doi.org/10.1177/070674370805300608>
- Karpman, B. (1946). A yardstick for measuring psychopathy. *Federal Probation, 10*, 26–31.
- Karpman, B. (1948). The myth of the psychopathic personality. *The American Journal of Psychiatry, 104*, 523–534. <http://dx.doi.org/10.1176/ajp.104.9.523>
- Kelley, S. E., Edens, J. F., Mowle, E. N., Penson, B. N., & Rulseh, A. (2019). Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies. *Journal of Clinical Psychology, 75*, 627–643. <http://dx.doi.org/10.1002/jclp.22726>
- Kennedy v. Louisiana, 554 U.S. 407 (2008).
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. <http://dx.doi.org/10.1016/j.jcm.2016.02.012>
- Leistico, A. M. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28–45. <http://dx.doi.org/10.1007/s10979-007-9096-6>
- Lockett v. Ohio, 438 U.S. 586 (1978).
- Martinez v. Dretke, 99 Fed. Appx. 538 (5th Cir. 2004).
- McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., & Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable- and person-focused analyses. *Law and Human Behavior, 32*, 325–338. <http://dx.doi.org/10.1007/s10979-007-9094-8>
- 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>
- Millon, T., Simonsen, E., & Birket-Smith, M. (1998). Historical conceptions of psychopathy in the United States and Europe. In T. Millon & E. Simonsen (Eds.), *Psychopathy: Antisocial, criminal, and violent behavior* (pp. 3–31). New York, NY: Guilford Press.
- Mokros, A., Hare, R. D., Neumann, C. S., Santtila, P., Habermeyer, E., & Nitschke, J. (2015). Variants of psychopathy in adult male offenders: A latent profile analysis. *Journal of Abnormal Psychology, 124*, 372–386. <http://dx.doi.org/10.1037/abn0000042>
- Morrissey, C., Hogue, T., Mooney, P., Allen, C., Johnston, S., Hollin, C., . . . Taylor, J. L. (2007). Predictive validity of the PCL-R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry & Psychology, 18*, 1–15. <http://dx.doi.org/10.1080/08990220601116345>
- Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among expert witnesses. *Annual Review of Law and Social Science, 11*, 37–55. <http://dx.doi.org/10.1146/annurev-lawsocsci-120814-121714>
- Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24*, 1889–1897. <http://dx.doi.org/10.1177/0956797613481812>
- Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior, 32*, 352–362. <http://dx.doi.org/10.1007/s10979-007-9097-5>
- Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*, 19–53. <http://dx.doi.org/10.1037/a0014897>
- Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
- Panetti v. Quarterman, 551 U.S. 930 (2007).
- Patrick, C. J. (Ed.), (2018). *Handbook of psychopathy* (2nd ed.). New York, NY: Guilford Press.
- Poythress, N. G., Edens, J. F., Skeem, J. L., Lilienfeld, S. O., Douglas, K. S., Frick, P. J., . . . Wang, T. (2010). Identifying subtypes among offenders with antisocial personality disorder: A cluster-analytic study. *Journal of Abnormal Psychology, 119*, 389–400. <http://dx.doi.org/10.1037/a0018611>
- Ring v. Arizona, 536 U.S. 584 (2002).
- Roberts v. Louisiana, 428 U.S. 325 (1976).
- Roper v. Simmons, 543 U.S. 551 (2005).
- Shapiro, M. (2009). An overdose of dangerousness: How “future dangerousness” catches the least culpable capital defendants and undermines the rationale for the executions it supports. *American Journal of Criminal Law, 35*, 145–200. <http://dx.doi.org/10.2139/ssrn.1402362>
- Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*, 499–513. <http://dx.doi.org/10.1016/j.cpr.2010.11.009>
- Sorensen, J. R., & Wrinkle, R. D. (1996). No hope for parole: Disciplinary infractions among death-sentenced and life-without-parole inmates.

- Criminal Justice and Behavior*, 23, 542–552. <http://dx.doi.org/10.1177/0093854896023004002>
- 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>
- United States v. Barnette, 211 F. 3d 803 (4th Cir. 2000).
- United States v. Fell, 531 F. 3d 197 (2nd Cir. 2008).
- United States v. Sablan, 555 F. Supp. 2d 1177 (D. Colo, 2006).
- Walters, G. D. (2003a). Predicting criminal justice outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences & the Law*, 21, 89–102. <http://dx.doi.org/10.1002/bsl.519>
- Walters, G. D. (2003b). Predicting institutional adjustment and recidivism with the psychopathy checklist factor scores: A meta-analysis. *Law and Human Behavior*, 27, 541–558. <http://dx.doi.org/10.1023/A:1025490207678>
- Walters, G. D., & Mandell, W. (2007). Incremental validity of the psychological inventory of criminal thinking styles and psychopathy checklist: Screening version in predicting disciplinary outcome. *Law and Human Behavior*, 31, 141–157. <http://dx.doi.org/10.1007/s10979-006-9051-y>
- World Health Organization. (1992). *ICD-10: International statistical classification of diseases and related health problems* (10th rev.). Geneva, Switzerland: Author.
- Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *The International Journal of Forensic Mental Health*, 16, 227–238. <http://dx.doi.org/10.1080/14999013.2017.1317302>

Appendix

Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist—Revised (PCL–R) in Capital Sentencing to Assess Risk for Institutional Violence

We, a group of concerned forensic mental health professionals comprising the individuals listed in Attachment A, state the following:

1. It is our consensus opinion that the Hare Psychopathy Checklist—Revised (PCL–R), a quantitative psychological test (Hare, 1991, 2003), is not generally accepted in the field of forensic mental health as a reliable and valid means of predicting serious institutional violence, that is, of estimating or determining the likelihood that a person will commit such violence in the future.
2. Our qualifications and the foundation of our consensus opinion are set out herein.

Qualifications

3. We are forensic scientists who have helped to develop, validate, and test the PCL–R in both laboratory and real-world settings and are familiar with research and practice related to the PCL–R.
4. We are active as researchers or practitioners in the field of forensic mental health. We have played prominent roles in that field as members of scientific and professional associations or the editorial boards of leading scientific and professional journals. We have conducted research on the evaluation of the PCL–R and presented the findings of our research in the form of articles in peer-reviewed journals, books and book chapters, and conference presentations.

Many of us have conducted training workshops on the clinical-forensic use of the PCL–R. Many of us have used the PCL–R in the course of our practice as forensic mental health professionals, and some of us have been qualified to give expert testimony about or based on the PCL–R before courts throughout the United States.

5. None of us has an actual, potential, or perceived conflict of interest with respect to the PCL–R by which we would gain commercially or in some other way from offering the specific opinions herein or that would otherwise compromise our neutrality or objectivity.

The Nature of Quantitative Psychological Tests

6. Psychological tests are, most generally, evaluative devices or procedures intended to provide information relevant to some target construct that is either a real object (i.e., a part of the natural world) or an ideal object (i.e., a linguistic, inferential, or theoretical concept). Some (but not all) psychological tests are quantitative in nature, relying on numeric algorithms to generate scores or decisions that measure (i.e., gauge, represent, or predict) the target construct.
7. In contemporary practice, quantitative psychological tests are developed and evaluated using psychometric theory, which is a set of concepts, principles, and statistical procedures designed specifically for that purpose.

(Appendix continues)

8. Two primary concepts in psychometric theory are reliability and validity. In this context, reliability is potential freedom from measurement error and reflects the degree to which test scores or decisions may be precise, replicable, stable, and consistent; and validity is potential meaningfulness of measurement and reflects the degree to which test scores or decisions may be logically or empirically coherent with, representative of, or predictive of the target construct. Reliability limits validity: Test scores or decisions may be high in reliability and low in validity (e.g., precise measures of the wrong thing) but cannot be high in validity unless they are also high in reliability.
9. The steps in developing a quantitative psychological test typically include: derivation, or selection of its format and content; initial validation (also known as construction), or administration of the test in one or more data sets with the goal of exploring the reliability and validity of test scores or decisions and refining the test's format and content; and cross-validation (also known as calibration), or confirmation of the reliability and validity of test scores or decisions made using the final version of the test in one or more new data sets.
10. In forensic mental health practice, quantitative psychological test scores and decisions are expected to have a high level of reliability and validity, due to the important potential consequence of forensic decisions. The decision to use a quantitative psychological test therefore requires evaluators to conclude that the test scores or decisions are likely to have both high reliability and high validity in the case at hand. This conclusion requires two things: First, there is a body of research that provides strong direct or indirect support of the test's reliability and validity for similar purposes, in similar contexts, and for people with similar background; and second, the evaluators have sufficient expertise (i.e., training, supervision, and experience) in the use of the test to ensure they can accurately and appropriately administer, score, and interpret the test. Use of a quantitative psychological test in the absence of supporting research or sufficient expertise is contrary to standards of practice in forensic mental health.

The PCL-R

11. The PCL-R is a specific type of quantitative psychological test known as a symptom construct rating scale. It is

designed to assess features of a construct known as psychopathic personality disorder in correctional and forensic mental health settings. There is active debate in the scientific community concerning the nature of the construct of psychopathic personality disorder and how best to measure it. It is not included as a distinct diagnostic category in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013)* or in the tenth edition of the *International Statistical Classification of Diseases and Related Health Problems (ICD-10; World Health Organization, 1992)*. There is active debate concerning the degree to which the nature of the construct of psychopathic personality disorder and way in which it is measured using the PCL-R relate to antisocial personality disorder as defined and diagnosed according to *DSM-5* and dissocial personality disorder as defined and diagnosed in *ICD-10*.

12. The PCL-R comprises 20 individual items, presented in Attachment B. Each item is defined in detail in the test manual. Trained evaluators use judgment to rate each feature on a 3-point scale (briefly, 0 = *absent*, 1 = *partially present*, 2 = *present*) based on all available clinical data, including an interview with and observation of the person, interviews with collateral informants, and case history information.
13. Scores on the individual PCL-R items are summed to yield facet, factor, and total scores. Total scores, comprising all 20 items, are relied on most heavily as a global measure of the construct in research and practice. The PCL-R test manual suggests that total scores of 30 and higher (out of a maximum possible 40 points) are generally considered indicative of psychopathic personality disorder.

Reliability of PCL-R Scores in Forensic Mental Health Practice

14. Because the PCL-R is a symptom construct rating scale, PCL-R scores rely heavily on the judgment of evaluators. For this reason, a specific facet of reliability known as interrater reliability—that is, measurement precision related to agreement between evaluators with respect to test scores—is an issue of paramount importance. In particular, it is critical to understand how this interrater reliability impacts the expected disagreement between two independent evaluators, rating the same person at the same time on the basis of the same information, with respect to the PCL-R total scores they obtain; for the sake of simplicity, we will refer to this expected disagreement as the “margin of error” of PCL-R total scores.

(Appendix continues)

15. Prior to the mid-2000s, the available research evidence indicated that, overall, the interrater reliability of PCL-R scores was moderate in magnitude. But the research base at that time had two important limitations:
- Most studies were conducted for the purpose of research or in research settings, in which the PCL-R was administered by specially trained research assistants under conditions of anonymity; there was an absence of studies on interrater reliability conducted in the context of forensic mental health practice or in applied settings (i.e., “field settings”), in which the PCL-R was administered by health care professionals as part of routine clinical or forensic practice.
 - Most studies used statistical methods of older rather than more contemporary psychometric theory (i.e., Classical Test Theory as opposed to Generalizability Theory and Modern Test Theory).
16. Since the mid-2000s, several studies on the interrater reliability of PCL-R scores were conducted in the context of forensic mental health practice or in applied settings, or used methods of contemporary psychometric theory. These studies have yielded two new and important findings.
17. The first new and important finding is that the interrater reliability of PCL-R scores is often substantially lower when the test is evaluated in the context of forensic mental health practice or in applied settings than it is when evaluated for research purposes or in research settings. The interrater reliability of PCL-R is typically indexed using intraclass correlation coefficients (ICCs). There are actually many different specific types of ICCs, all of which reflect the agreement between evaluators under different conditions or assumptions. ICCs have a theoretical range from -1 (*perfect disagreement among two or more evaluators*) to 0 (*chance levels of agreement*) to $+1$ (*perfect agreement*). Prior to the mid-2000s, the ICCs reported for agreement between independent evaluators working in research contexts were typically summarized as falling in the range of .80 to .90 (e.g., Hare, 1991, 2003), which may be characterized according to various interpretive guidelines as “good” but not “excellent” (Koo & Li, 2016). Since that time, however, studies in field settings reported ICCs that were much lower, falling in the range of .40 to .70 (e.g., Boccaccini, Turner, & Murrie, 2008; Edens, Boccaccini, & Johnson, 2010; Sturup et al., 2014), which may be characterized as “poor” to “moderate” (Koo & Li, 2016). The relatively low interrater reliability observed in field settings can be attributed in part to the limited quality and quantity of information on which evaluators relied, as well as to the limited training, supervision, and experience of those evaluators; although there is further evidence that it may also be due to the adverse impact of adversarial proceedings on the judgment of evaluators (DeMatteo et al., 2014b; Edens et al., 2015; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Murrie et al., 2013; Murrie et al., 2008; Murrie, Boccaccini, Turner, Meeks, Woods, & Tussey, 2009). This phenomenon has been referred to as “adversarial bias” or “allegiance bias” and may be considered a special case of what is referred to more generally in forensic decision making as “confirmatory bias” (Zapf & Dror, 2017).
18. The second new and important finding is that, for a given estimate of the interrater reliability of PCL-R scores, the expected disagreement between evaluators or “margin of error” is substantially larger than was estimated previously. For example, prior to the mid-2000s, the expected disagreement for PCL-R total scores was estimated to be ± 3 points (out of a total of 40 points) in 68% of cases, and ± 6 points in 95% of cases (e.g., Hare, 1991, 2003). Put simply, the total scores of two independent evaluators were expected to be within 3 points of each other most of the time, and within 6 points almost all the time. But since that time, more precise calculations based on contemporary psychometric theory indicate the margin of error—even assuming the same level of interrater reliability, that is, .85—is actually ± 3 points in only 68% of cases, but ± 9 points in 95% of cases (e.g., Cooke & Michie, 2010). Additional analyses indicate that even this is an overly optimistic estimate of the margin of error, for two reasons (Cooke & Michie, 2010). First, it assumes that the interrater reliability of PCL-R total scores is about .85, whereas in field settings the interrater reliability may be considerably lower. Second, it is an estimate of the margin of error around the center of the distribution of PCL-R scores (i.e., about 20 points out of 40); however, the margin of error in fact becomes asymmetric and increases as scores approach the extremes or “tails” of the distribution (i.e., <10 and >30). This means the margin of error is larger at or around the score typically used to define psychopathic personality disorder, which is 30 points or higher out of 40. Thus, if one assumes that the interrater reliability of PCL-R scores is .80 (i.e., only slightly lower than the value of .85 assumed in the PCL-R manual), and assuming the evaluator reported a PCL-R total score of 30 points out of 40, then the total score obtained by independent evaluators would be expected to fall somewhere between 24 and 33 points out of 40 in 68% of cases, and between 19 and 36 points in 95% of cases (Cooke & Michie, 2010). In sum, the consequence of this large margin of error is considerable—and possibly even grave—uncertainty about the accuracy of a PCL-R total score obtained by a given evaluator. For example, if an evaluator administers the PCL-R and obtains a total score of 30, then one out of three evaluators who independently readministered the PCL-R would obtain scores less than or equal to 23 or, alternatively, greater than or equal to 34. This is true even assuming the interrater reliability for PCL-R total scores is good (i.e., .80), the evaluators all have the same level of training and experience, and the assessments were conducted at the same time and on the basis of the same information.

(Appendix continues)

Validity of PCL-R With Respect to Prediction of Serious Institutional Violence

19. According to the test manual and the writings of the test developer, the PCL-R was not developed and is not recommended to estimate the likelihood or predict that a person will commit violence in the future, either in the community or in an institution. As the test manual states, "Properly used, the PCL-R provides a reliable and valid assessment of an important clinical construct—psychopathy. **Strictly speaking, that is all it does**" (Hare, 2003, p. 15; emphasis in original).
20. Prior to the mid-2000s, the available research evidence indicated that, overall, PCL-R scores were associated with increased risk for violence in general; but they could not be used, either on their own or in combination with other risk factors, to estimate the likelihood of or predict future institutional violence by an individual with high reliability or validity. There were at least two major reasons for this:
 - a. There was little or no research on the prediction of serious institutional violence using the PCL-R generally, and none at all on the prediction of serious violence in federal prisons in the United States.
 - b. There was no research at all on the prediction of violence using the PCL-R at the individual level, as opposed to the group level.
21. Since the mid-2000s, several studies on prediction of serious institutional violence using the PCL-R have been conducted. These studies have yielded two new and important findings.
22. The first new and important finding is that the predictive validity of PCL-R scores is inadequate to support its use as a tool to assess risk for serious institutional violence. For example, a number of meta-analytic reviews of the literature (e.g., Campbell et al., 2009; Guy et al., 2005; Leistico, Salekin, DeCoster, & Rogers, 2008) have demonstrated that the association between PCL-R total scores and serious institutional violence is limited; and, furthermore, the magnitude of the association tended to be even smaller in studies that were conducted in prisons (as opposed to forensic mental

health facilities) or in the United States (as opposed to other countries).

23. The second new and important finding is that there are significant challenges inferring an individual's likelihood of recidivism from group-level data with a high degree of accuracy and precision. A number of scholars (e.g., Cooke & Michie, 2010; Faigman, Monahan, & Slobogin, 2014; Hart & Cooke, 2013; Hart, Michie, & Cooke, 2007) have discussed the logical, methodological, and statistical barriers to defining and estimating individual-level predictions of violence risk, including predictions of violence using the PCL-R.
24. These two new and important findings concerning the validity of the PCL-R with respect to the prediction of institutional violence are likely due, at least in part, to the limited interrater reliability and substantial margin of error of PCL-R total scores.

Changes Over Time in the Evidence Base Concerning the Interrater Reliability and Predictive Validity of the PCL-R

25. Prior to the mid-2000s, the existing evidence base (i.e., body of peer-reviewed research) concerning the PCL-R was limited in important respects. There was no research supporting either the interrater reliability of the PCL-R in field settings or the predictive validity of the PCL-R with respect to serious institutional violence—that is, there was an "absence of proof" of the PCL-R's reliability and validity in these respects.
26. Since the mid-2000s, the evidence base concerning the PCL-R has expanded greatly. There is now a body of research indicating serious problems with the interrater reliability of the PCL-R in field settings and the predictive validity of the PCL-R with respect to serious institutional violence—that is, there is now "proof of absence" of the PCL-R's reliability and validity in these respects.
27. For these reasons, it is our consensus opinion that PCL-R scores cannot and should not be used to estimate the likelihood or predict that people will commit serious institutional violence. The use of PCL-R scores for such purposes is inconsistent with standards of practice in the field of forensic mental health.

(Appendix continues)

Attachment A

Members of the Group of Concerned Forensic Mental Health Professionals

- Marcus T. Boccaccini
Department of Psychology and Philosophy, Sam Houston State University
- Mark D. Cunningham
Private Practice, Seattle, Washington
- David DeMatteo
Department of Psychology & Thomas R. Kline School of Law, Drexel University
- Kevin S. Douglas
Department of Psychology, Simon Fraser University
- Joel A. Dvoskin
Department of Psychiatry, University of Arizona College of Medicine
- John F. Edens
Department of Psychological & Brain Sciences, Texas A&M University
- Laura S. Guy
Department of Psychology, Simon Fraser University
- Stephen D. Hart
Department of Psychology, Simon Fraser University
- Kirk Heilbrun
Department of Psychology, Drexel University
- Daniel C. Murrie
Institute of Law, Psychiatry, and Public Policy, University of Virginia
- Randy K. Otto
Department of Mental Health Law & Policy, University of South Florida

- Ira K. Packer
Department of Psychiatry, University of Massachusetts Medical School
- Thomas J. Reidy
Private Practice, Monterey, California

Attachment B

Items in the Hare Psychopathy Checklist—Revised

Item
1. Glibness/superficial charm
2. Grandiose sense of self worth
3. Need for stimulation/proneness to boredom
4. Pathological lying
5. Conning/manipulative
6. Lack of remorse or guilt
7. Shallow affect
8. Callous/lack of empathy
9. Parasitic lifestyle
10. Poor behavioral controls
11. Promiscuous sexual behavior
12. Early behavioral problems
13. Lack of realistic, long-term goals
14. Impulsivity
15. Irresponsibility
16. Failure to accept responsibility for own actions
17. Many short-term marital relationships
18. Juvenile delinquency
19. Revocation of conditional release
20. Criminal versatility

Received October 25, 2019
Revision received November 25, 2019
Accepted November 26, 2019 ■