AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychology, Public Policy, and Law

ISSN: 1076-8971

2020, Vol. 26, No. 4, 511–518
http://dx.doi.org/10.1037/law0000285

Death Is Different: Reply to Olver et al. (2020)

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

In our “Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist-Revised [PCL-R] in Capital Sentencing to Assess Risk for Institutional Violence,” DeMatteo et al. (2020) summarized the relevant empirical research and concluded that 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 in the context of capital sentencing decisions. In a solicited commentary, Olver et al. (2020) raised several concerns about our statement and presented new analyses of the research literature. In this reply, we identify crucial points about which Olver et al. disagreed with the statement and, after analyzing their concerns, conclude that their concerns are either (a) based on misunderstanding or mischaracterization of the statement, or (b) irrelevant to the purpose and content of our statement. We also conclude that it is not possible to properly evaluate the new analyses presented by Olver et al. in the absence of full technical detail that would permit adequate peer review.

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

David DeMatteo, Department of Psychology, and 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, Independent 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 and 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 and Policy, University of South Florida; Ira K. Packer, Department of Psychiatry, University of Massachusetts Medical School; Thomas J. Reidy, Independent Practice, Monterey, California.

The authors are presented in alphabetical order following Kirk Heilbrun. This article represents our views as individual forensic mental health professionals; it does not necessarily reflect the views of the agencies or organizations with which we are affiliated or for which we work.

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

In our previous article, we presented a statement of consensus ("Statement") regarding the use of the Hare Psychopathy Checklist-Revised (PCL-R; Hare, 1991, 2003) to predict serious institutional violence in capital sentencing evaluations (DeMatteo et al., 2020). The Statement comprised two parts—a narrative introduction ("Introduction") and an appendix containing the Statement in its original, declarative form ("Appendix"). The Statement summarized the available evidence as follows: (a) the interrater reliability of PCL-R scores in field settings, particularly in adversarial contexts, may be problematically low; (b) the overall association between PCL-R scores and violence at the group level is moderate in terms of effect size, both in absolute terms and relative to the effect size of other established risk factors for violence; (c) the association between PCL-R scores and violence in institutional settings is small in terms of effect size; and (d) the association between PCL-R scores and serious institutional violence is even smaller. We concluded that one cannot use PCL-R scores in 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 their use for this purpose.

In response ("Counterstatement"), Olver et al. (2020) identified several areas of concern with the Statement, presented new analyses regarding the PCL-R's predictive validity and field reliability, and offered recommendations intended to support the ethical and appropriate use of the PCL-R for assessing risk of institutional violence in capital sentencing contexts.[1] We appreciate the opportunity to reply to the Counterstatement, as dialogue can play an important role in advancing law, policy, and practice. After highlighting areas of agreement between the Statement and Counterstatement, we respond to the concerns identified in the Counterstatement. We conclude that Olver et al.'s concerns are of little consequence given the areas of agreement between the Statement and Counterstatement. Most of the concerns reflect a mischaracterization of the Statement or misunderstanding of the law pertaining to capital sentencing evaluations. Some concerns were based on multiple complex data analyses introduced for the first time in the Counterstatement without complete technical detail that would permit adequate peer review or proper evaluation in this response. Other concerns have little or no relevance to the Statement.

Areas of Agreement

It is important to highlight points made in the Statement that are accepted, agreed with, or repeated in the Counterstatement, as they render much of the rest of the Counterstatement moot.

1. The PCL-R is a psychological test of psychopathic traits; it was not developed to predict violence or assess violence risk.

2. As is true for all psychological tests, PCL-R scores have imperfect interrater reliability and, in field settings, may be susceptible to various sources of interference that include adversarial bias.

3. Research—most of which was conducted after the mid-2000s—indicates that the interrater reliability of PCL-R scores observed in field settings (which falls in the range typically characterized as *good* or *substantial* overall, and in the range typically characterized as *fair* or *moderate* for studies conducted in the United States) is substantially lower than that reported in the PCL-R manual (which, on average, falls in the range typically characterized as *excellent*).

4. PCL-R scores have moderate overall predictive validity with respect to violence across a wide range of settings.

5. Research—including considerable research conducted after the mid-2000s—indicates that the overall predictive validity of PCL-R scores with respect to institutional violence is not large (i.e., is low to moderate) in terms of effect size.

6. As is true for all psychological tests, PCL-R scores alone should not be relied on by evaluators to make clinical or forensic decisions, including predictions of violence, but they may be appropriately incorporated into comprehensive, contextualized, individualized, and prevention-oriented evaluations of violence risk in certain contexts.

7. As PCL-R scores have interrater reliability and predictive validity with respect to serious institutional violence that is equal or superior to that of all other psychological tests of psychopathic traits, limitations of the PCL-R apply to those other tests.

If this is the message that people take away from reading the Statement, then we consider it a success. Everything else is in the Statement is detail; everything else in the Counterstatement is either detail or irrelevant to the central issue of using PCL-R scores to predict serious institutional violence in capital contexts.

Areas of Disagreement: Concerns Raised in the Counterstatement

Exclusive Focus on the PCL-R

The Counterstatement expressed concern that the Statement's exclusive focus on the PCL-R—which was characterized as having "singled out" the PCL-R—was to use the test as a "psycholegal red herring" (Olver et al., 2020, p. 491). The Counterstatement would have preferred that the Statement address legislative, systemic, and practical issues that affect other tests of psychopathic traits (e.g., Screening Version of the PCL-R [PCL:SV]; Hart, Cox, & Hare, 1995) and other forensic assessment issues, tools, and procedures more generally (e.g., HCR-20V3; Douglas, Hart, Webster, & Belfrage, 2013), some of which the Counterstatement noted had received "scant" reference (p. 502).

The reasons for the Statement's exclusive focus on the PCL-R are made clear in the Introduction. First, the PCL-R is often identified as the gold standard for assessing psychopathy in forensic mental health—that is, the most widely researched and most

[1] Olver et al.'s (2020) recommendations go beyond the Statement's scope, but we have attempted to limit our comments to issues directly relevant to assessing risk of serious institutional violence in capital sentencing evaluations.

widely used test. Second, although the PCL-R is not intended to assess violence risk, PCL-R scores are sometimes (historically and currently) offered in capital sentencing evaluations as evidence of an offender's future dangerousness in terms of risk for serious institutional violence. We do not understand how the focus of the Statement, or the reasons for it, can be considered a "red herring."

The Counterstatement raises several issues that are irrelevant, misleading, or distracting with respect to the Statement's focus. The issue of using the PCL-R outside of the capital sentencing context is irrelevant. The fact that research may support the use of test scores for one purpose does not support their use for another, unrelated purpose. Also, the issue of using other psychopathy measures (e.g., PCL:SV) in the capital sentencing context is misleading and distracting. We noted in the Introduction that the PCL-R is the gold standard for assessing psychopathy in forensic mental health contexts, and we stated unambiguously that "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" (DeMatteo et al., 2020, pp. 137–138). The Statement concluded the gold standard was not fit for the purpose of predicting serious institutional violence, so we see no value in discussing measures that it acknowledged are either no better or inferior. Also, noting that other measures of psychopathy are at least as flawed as the PCL-R is a facile argument in support of using the test, in effect damning the PCL-R with faint praise. Finally, other forensic issues, assessments, and procedures (e.g., HCR-20V3) are not relevant to consideration of using PCL-R scores to predict serious institutional violence in capital sentencing evaluations. These are interesting and important issues that Olver et al. (2020) are free to explore, as we have done in some of our publications, but raising these issues does not, in and of itself, constitute a valid criticism of the research summaries or opinions in the Statement.

The Counterstatement mischaracterized the Statement as a wholesale condemnation of the PCL-R. Rather than accepting or rejecting the PCL-R holus-bolus, the Statement considered whether it could and should be used to predict with a reasonable degree of accuracy and precision whether a given person will engage in serious institutional violence if incarcerated rather than executed. This approach is consistent with standards for psychological testing (American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education, 2014), which eschew general claims about whether a test is "valid" in favor of more exacting statements concerning the interpretation of test scores in particular contexts. Although the Statement concluded the evidence base indicates PCL-R scores cannot be used to predict serious institutional violence in capital sentencing evaluations, it described the PCL-R as the gold standard for assessing psychopathy (DeMatteo et al., 2020, p. 134); emphasized the potential relevance of the PCL-R as part of comprehensive, individualized, and contextualized violence risk assessments (p. 137); and acknowledged that the Statement authors have used the PCL-R in their research and practice (p. 140).

Failure to Specify "Serious" Institutional Violence

The Counterstatement expressed concern that the Statement did not define "serious" institutional violence. The definition is not explicit in or identical across all statutory and case law relevant to capital sentencing, so we used the term "serious" in its plain language sense, which is broadly consistent with how it is used in the law and in social science research. "Serious" violence is typically defined as illegal conduct that causes or has the potential to cause grave physical or psychological harm, and especially that which is lethal or life-threatening (e.g., Douglas et al., 2013). For example, the MacArthur Violence Risk Assessment Study defined "serious acts of violence" as battery that resulted in physical injury, sexual assaults, assaultive acts that involved the use of a weapon, or threats made with a weapon in hand (Steadman et al., 1998). As another example, the Model Penal Code (American Law Institute, 1985), defines "serious bodily injury" as that "which creates a substantial risk of death or which causes serious, permanent disfigurement, or protracted loss or impairment of the function of any bodily member or organ" (§ 210.0(3)).

The Counterstatement offered its own definition of serious institutional violence: a "range of injurious acts, including those that cause significant psychological trauma" that included "serious institutional misconducts, general violence/aggression, [and] general misconduct" (Olver et al., 2020, p. 492). They may be interested in whether the PCL-R yields useful data regarding the prediction of this broader "range of injurious acts," but it is largely irrelevant to capital sentencing law in the United States. Institutional infractions deemed "serious" in some correctional systems, such as disruptive behaviors and possessing contraband (e.g., food, pornography, cannabis), are inarguably "security or management concerns" (p. 492) and may even be related to violence risk, but they do not constitute violence per se.[2]

Focus on Prediction

The Counterstatement expressed concern that the Statement "underspecifies" the meaning of prediction, focusing too narrowly on a determination of the likelihood of target behaviors to the exclusion of risk mitigation considerations. The Statement's focus on prediction of serious violence reflects the focus of statutory and case law related to capital sentencing in the United States. In many jurisdictions, the decision for the trier of fact is between life in prison and death, and this decision is required to consider a defendant's future dangerousness potential for serious institutional violence if sentenced to life in prison. But this decision need not consider risk management or mitigation and, therefore, is much more predictive than preventive in nature. Furthermore, as noted in the Statement, PCL-R scores have been and still are being offered and interpreted in capital sentencing evaluations as evidence of an offender's future dangerousness.

The Statement authors are aware of the distinction between prediction and prevention approaches to violence risk assessment, both generally (Heilbrun, 1997) and with respect to psychopathy (Hart, 1998). The Statement clarified that the PCL-R manual states that the PCL-R should not be used predictively or in isolation to make decisions about violence risk in any context (p. 143): "As the

[2] The term "future dangerousness" is a common but potentially misleading shorthand used by courts. Case law in many jurisdictions has not operationally defined several key terms, including "probability," "criminal acts of violence," and "continuing threat to society" (Cunningham et al., 2009). Simply because the law leaves terms ambiguously defined does not empower scientists to be equally obtuse and expansive. As such, although case law may not expressly preclude consideration of a wide range of conduct in determining if a given defendant is a future danger, ethical and responsible conduct by psychologists arguably does.

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)." Also, the Statement emphasized that PCL-R scores may be appropriately used "as part of a comprehensive, individualized, and contextualized evaluation" (p. 137). It appears that the Counterstatement agreed with us regarding this issue.

Role of "Future Dangerousness" in Capital Sentencing Evaluations

The Counterstatement expressed concern that the Statement incompletely or inaccurately summarized capital sentencing law in the United States, and it offered its own summary. A broad discussion of legal issues is beyond the scope of this reply, and we refer interested readers to other sources (e.g., Cunningham, Sorensen, & Reidy, 2009; DeMatteo, Murrie, Anumba, & Keesler, 2011). Briefly and with respect, however, the Counterstatement contains certain inaccuracies regarding United States law.

We offer two examples in which the Counterstatement incorrectly claimed the Statement erred in its legal review. First, the Counterstatement claimed the Statement was wrong about the number of jurisdictions that consider future dangerousness in capital sentencing and stated, "Nine states require it, two permit it, four allow its absence as a mitigating factor, and the remainder varies on the admissibility of evidence about dangerousness" (Olver et al., 2020, p. 492). In support, it cited Bright (2015), but this citation is to teaching materials—several years out of date—from Bright's "Class 3" of his capital punishment course.[3] Bright noted that Oregon treats future dangerousness as a special issue the jury must answer affirmatively to impose the death penalty, but Oregon eliminated this special issue in 2019. The Counterstatement also omits that Bright noted future dangerousness may be considered as a nonstatutory aggravating factor in several states and the federal jurisdiction. Second, the Counterstatement incorrectly claimed that the absence of future dangerousness was a mitigating factor in only four jurisdictions. But the four jurisdictions identified by Bright only reflect those in which this mitigating factor is specified by statute; as the U.S. Supreme Court has held, positive prisoner adjustment can be introduced as a mitigating factor in all capital cases (see *Skipper v. South Carolina*, 1986).

Failure to Define and Identify an "Acceptable" Threshold for Precision and Accuracy

The Counterstatement expressed concern that the Statement did not define the terms *precision* and *accuracy* or identify a threshold that could be used to determine whether the precision and accuracy of PCL-R scores to predict serious institutional violence in the context of capital sentencing evaluations are "acceptable" or "good enough." Given the context in which the terms appear in the Statement, it is clear that we defined precision and accuracy in a manner that is consistent with their usage in social science; that is, as synonyms for *reliability* and *validity*, respectively. Of note, the Counterstatement defined the terms in the same way.

Furthermore, we did not specify a threshold for determining what is acceptable precision or accuracy, because no such threshold exists. Rather, the judgments of acceptability of precision and accuracy depend on context, including the indices of precision or accuracy being considered and the costs of various types of decision error. Let us take accuracy (more specifically, in the current context, predictive validity) as an example. Given the extremely high stakes involved in capital sentencing evaluations, we believe it is reasonable to expect that the threshold for overall accuracy of predictions of serious institutional violence made using the PCL-R should be a large effect size. But even if the use of test scores yields large effect sizes, the specific pattern of errors (or various "error rates") may indicate that those scores lack sufficient precision or accuracy to make decisions in a given case. (For a readable discussion of this issue, see Pogrow, 2019.) Furthermore, even large effect sizes can yield poor positive predictive power if the outcome has a low base rate, which would potentially result in the execution of individuals erroneously predicted to be violent if not put to death.

Later in the Counterstatement, this concern is revisited: "[S]ince when did less than 'perfect' reliability become the threshold for an unacceptable margin or rater error? Do all other measures have 'perfect' reliability?" (Olver et al., 2020, p. 18). First, to our knowledge, "perfect reliability" has never been a required threshold for anything; the Statement never said this. Second, to our knowledge, no psychological measure has "perfect" reliability; the Statement never said this, either. If Olver et al. are highlighting that all other tests of psychopathic traits are no better than the PCL-R in terms of interrater reliability (and predictive validity), we agree—indeed, we said so in the final sentence of the Introduction (DeMatteo et al., 2020, p. 138).

Incomplete or Inaccurate Description of Research on Predictive Validity

The Counterstatement expressed concern that the Statement did not provide a full or accurate summary of the evidence concerning the PCL-R's predictive validity. In support, it presented new analyses that included a metameta-analysis, a meta-analysis, and an "illustration" that includes reanalysis of data from one study using a combination of structural equation modeling and latent profile analysis. Our primary objection to the metameta-analysis and meta-analysis is that they are based in part on consideration of tests other than the PCL-R (i.e., the PCL:SV) and on studies that used a much broader criterion than serious institutional violence. They are, therefore, irrelevant to the research summary in the Statement. It is hardly surprising that analyses of data that included studies using different tests and a different outcome criterion yielded a different finding. We refrain from making other comments about their analyses because a fair and proper review of their work would require presentation of detail regarding the

[3] This material is neither peer-reviewed nor drawn from a law review article, and it does not reflect recent developments in U.S. death penalty law.

analyses in a manner that is consistent with journal article reporting standards.[4]

Incomplete or Inaccurate Description of Research on Field Reliability

The Counterstatement expressed concern that the Statement did not provide a complete or accurate summary of research regarding the PCL-R's interrater reliability in field settings. In support, it presented another new meta-analysis. As with the previous new analyses, we are not able to comment in detail on the new meta-analysis of field reliability due to incomplete detail in the Counterstatement. That said, we note that the findings presented in Table 4 of the Counterstatement indicated that the average interrater reliability of PCL-R total scores in field settings was .68 (ICC_{A1}). According to some commonly used interpretive guidelines, this level of interrater reliability may be characterized as *good* (Cicchetti, 1994), *moderate* (Koo & Li, 2016), or *substantial* (Landis & Koch, 1977) – in every case, below the ranges considered *excellent* (Cicchetti, 1994), *good* or *excellent* (Koo & Li, 2016), and *almost perfect* (Landis & Koch, 1977).

More importantly, the descriptive labels are largely irrelevant in the context of making sense of the reliability of an individual PCL-R score from one examiner in a given case. As two of us have detailed elsewhere (Edens & Boccaccini, 2017, Table 1), an ICC value of .70 for a test would produce a 95% confidence interval around an average test score (50th percentile) that ranges from approximately the 14th percentile to the 86th percentile (assuming a normally distributed normative sample). As noted more than a quarter of a century ago, confidence intervals expand quite dramatically as ICC values drop below .90 (Nunnally & Bernstein, 1994).

The average interrater reliability noted in the Counterstatement also varied across regions, ranging from a high of .78 in studies from Canada to .56 in studies from the United States to a low of .50 in studies from Europe. Focusing on the United States, which is the only one of these regions with capital punishment, the level of interrater reliability would be characterized as *fair* by Cicchetti (1994) or *moderate* by Koo and Li (2016) and Landis and Koch (1977). This is substantially lower than the interrater reliability of PCL-R total scores reported in the test manual (ICC_1 = .87; Hare, 2003, p. 65). The Counterstatement concluded these findings demonstrate that "good field reliability with the PCL scales can and does happen" (Olver et al., 2020, p. 21), but those same findings also indicate that poor field reliability also happens—and more frequently than does good field reliability, particularly in the United States. According to Edens and Boccaccini (2017, Table 1), an ICC value of .55, which is nearly identical to the .56 value cited in the Counterstatement for U.S. studies, would produce a 95% confidence interval around an average test score that ranges from approximately the 9th percentile to the 91st percentile. Such confidence interval ranges are more informative for legal professionals to consider rather than the somewhat arbitrary labels of *fair*, *moderate*, or *good*.

"Mid-2000s" Psychometric Decline

The Counterstatement expressed concern that the Statement inaccurately claimed "that since the mid-2000s there was a sudden dropping off point that is almost taxonic in nature, where all the predictive validity and interrater reliability data began to turn up null findings that repudiated past efforts" (Olver et al., 2020, p. 22). We agree that this would have been a concern—if, in fact, the Statement had said this. However, the Statement did not refer to a "sudden dropping off point," characterize the change in the research literature as "almost taxonic in nature," or claim that all research on the PCL-R's field interrater reliability and predictive validity since the mid-2000s had yielded "null findings that repudiated past efforts."

The Statement gave a much more nuanced interpretation. First, the Statement described the state of the science up to the mid-2000s as suffering from a lack of research on the PCL-R that was directly relevant to capital sentencing evaluations, which it characterized as an "absence of proof" that PCL-R scores had field interrater reliability and predictive validity sufficiently high to support their use to predict serious institutional violence in capital sentencing contexts (DeMatteo et al., 2020, p. 134). Second, the Statement described the state of the science since the mid-2000s as reflecting an increase in directly relevant research, which confirmed that the PCL-R had important limitations in these respects and failed to support its use to predict serious institutional violence in capital sentencing contexts; the Statement characterized this as "proof of absence" (p. 134).

Is PCL-R Field Reliability Invariably and Inexorably Poor?

The Counterstatement expressed concern that the Statement viewed the fact that "research shows that high interrater reliability occurs with trained raters using high quality information" as "unexpected or undesirable" (Olver et al., 2020, p. 503). To be clear, the Statement did not state that field interrater reliability is "invariably" or "inexorably" poor.[5] Rather, the Statement noted that research conducted over the past 10 to 15 years indicates 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" (DeMatteo et al., 2020, p. 142).

Is Adversarial Allegiance a Problem That Uniquely Affects the PCL Scales?

The Counterstatement expressed concern that the Statement implied that adversarial allegiance effects are both unique to the PCL-R and "inevitable" (Olver et al., 2020, p. 504). The Statement neither said nor implied this. The Statement authors are very familiar with the literature on adversarial allegiance and have discussed its implications with respect to forensic evaluations elsewhere (e.g., DeMatteo, Murrie, Edens, & Lankford, 2019; Murrie & Boccaccini, 2015; Murrie, Boccaccini, Guarnera, & Rufino, 2013; Murrie et al., 2009). The argument that other forensic tests and procedures may be just as susceptible or even more susceptible to adversarial allegiance than the PCL-R does not undermine the Statement's research summary or conclusions.

[4] The authors introduced complex new analyses in response to the Statement without including the detail necessary to permit proper peer review.

[5] For a broader discussion of the psychometric properties of assessment tools in field studies, we refer readers to a special issue of *Psychological Assessment* (Volume 29, Issue 6, 2017) edited by John F. Edens and Marcus T. Boccaccini.

Applying Group Data to the Individual Case

The Counterstatement expressed concern that the Statement asserted group data cannot be used to make predictions about individuals. But the Statement made no such bald assertion; rather, it said, "there are significant challenges inferring an individual's likelihood of recidivism from group-level data with a high degree of accuracy and precision" (DeMatteo et al., 2020, p. 143). The Statement's assertion in this respect paraphrased Faigman, Monahan, and Slobogin (2014): "In terms of scientific inference, reasoning from the group to an individual case presents considerable challenges" (p. 420). Put simply, of course one can make predictions about individuals based on group data; the question is to what extent one can do so with precision and accuracy.

For example, "The global mean height of adult men born in 1996 is 171 centimetres (cm), or 5 foot and 7.5 inches" (Roser, Appel, & Ritchie, 2019). The group data are certainly educative, but do not at all guarantee that a randomly selected male born in 1996 will be 171 cm tall. The precision and accuracy of predictions based on the group data will be affected not only by probabilistic (ludic or aleatory) uncertainty, as reflected in the confidence interval for the mean, but also by any bias in the original sampling procedures, the manner in which the survey data from various countries around the world were weighted to make them more representative of the global population, the specific methods used to calculate the mean and its confidence interval, and the other characteristics of the randomly selected male born in 1996 (as height is systematically related to many factors other than gender and age). The Statement authors are familiar with the extensive and vigorous debate about technical aspects of the "G2i" issue, which cannot be summarized here due to space limitations. But we assume all Statement and Counterstatement authors would agree with Faigman et al. (2014) that drawing inferences about, say, an individual offender's risk for serious institutional violence from group data is neither straightforward nor uncomplicated.

Conflating the Misuse and Psychometric Properties of the PCL-R

The Counterstatement expressed concern that the Statement conflated the improper use of the PCL-R with its psychometric properties (Olver et al., 2020, p. 493):

> Attributing poor and unethical use of an instrument to its psychometric properties only serves to fuel "pseudo-debates" and "apparent controversies" (Smith, Gacono, Fontan, Cunliffe, & Andronikof, 2020). In such instances, failure to consider the context of the discussion of issues can serve to create plausible-sounding arguments (e.g., straw person arguments) that, in actuality, are conceptually flawed (Smith et al., 2020).[6]

We agree that the issue of appropriate use of a test is distinct from the issue of psychometric properties, insofar as a test may be used unethically or irresponsibly even when the scores it yields have good reliability and validity for some particular purpose. This is precisely why the Statement focused so narrowly on problems associated with the PCL-R's interrater reliability and predictive validity, rather than its unethical or irresponsible use more broadly. The Statement summarized research that raises concerns about the limited interrater reliability of PCL-R scores in field settings, and research indicating that even if PCL-R scores have good interrater reliability, their predictive validity with respect to serious institutional violence is limited (i.e., by "limited," we mean limited to a degree that made it problematic to make predictions of serious institutional violence with a degree of precision and accuracy to support their use in capital sentencing evaluations). This is crucial because just as the reliability of test scores sets an upper bound on their validity, as discussed in the Statement's Appendix (DeMatteo et al., 2020, p. 141), limitations in both the reliability and validity of test scores constrain the degree to which one can rely on test scores to make decisions in high-stakes evaluations—and it does not get more high stakes than life versus death. The cost of decision errors is not the same across different uses of a test.

Rejection of Empirically Validated Assessment Tools

The Counterstatement expressed concern that the Statement was a "[r]ejection of empirically validated tools for guiding clinical/forensic decisions, whether because of potential misuse or a misguided rejection of using group data to inform individual decisions" that was "essentially a rejection of science" (Olver et al., 2020, p. 506). We make two points in response. First, the Statement does not reject empirically validated tools for guiding clinical/forensic evaluations. Quite the opposite; the Statement is an attempt to ensure that tools for guiding clinical/forensic decisions (here, the PCL-R) are empirically validated to an extent that they are fit for their intended purpose (here, predicting serious institutional violence in capital sentencing evaluations). Second, the Statement does not reject science. Again, quite the opposite; the Statement is an attempt to ensure that the science used as the foundation for developing and interpreting psychological tests is good science—that it generates trustworthy (precise, accurate) and useful (practically and legally relevant) data.

What Is the Alternative?

The Counterstatement expressed concern that the Statement did not provide "a viable alternative to the use of the PCL-R" (Olver et al., 2020, p. 492). We did not identify a different psychological test that, in our view, could predict serious institutional violence with sufficient precision and accuracy to justify or support its use in capital sentencing evaluations because we do not believe any test is fit for this purpose at present. But we recommended a viable alternative to the use of PCL-R scores to predict serious institutional violence—that is, we recommended that PCL-R scores could be used "as part of a comprehensive, individualized, and contextualized evaluation" (DeMatteo et al., 2020, p. 137).[7] First, as Olver et al. (2020) apparently agreed with us, we are not sure

[6] Note that the Smith et al. (2020) reference cited by Olver et al. (2020) relates to the use of the Rorschach Inkblot Test, not the PCL-R.

[7] Of course, the precision and accuracy of predictions of serious institutional violence made using such an evaluation would itself be a matter for debate. We refer readers to Cunningham et al. (2009) and other research concluding that with a very low base rate of serious prison violence among capital offenders and the dynamic responses of corrections staff, any mental health methodology predicting a probability of serious prison violence by a capital defendant will have a very high error rate (e.g., Cunningham, Reidy, & Sorensen, 2016; Cunningham & Sorensen, 2010; Cunningham, Sorensen, Vigen, & Woods, 2011; Edens, Buffington-Vollum, Keilen, Roskamp, & Anthony, 2005; Reidy, Sorensen, & Cunningham, 2012, 2013; Sorensen & Cunningham, 2010).

why they identified this as a concern. Second, we fail to see how this concern, even if true, would undermine the validity of our research summary and conclusions about the PCL-R. Third, even if there were no viable alternative to the PCL-R, that would not mean that the PCL-R is *ipso facto* "acceptable" or "good enough" as a predictor of serious institutional violence. As noted by Reynolds (2016), "Vetting tests for application in context . . . may occasionally lead us to the decision not to test as well" (p. 415).

Conflict of Interest

We agree with many of the recommendations made in the Counterstatement, some of which simply reiterated what we said in the Statement. But the Counterstatement makes one recommendation that we find odd. The third recommendation states, "An authorized PCL-R/PCL:SV trainer should train all evaluators to a high standard emphasizing that proper scoring requires the unbiased use of extensive, high-quality information" (Olver et al., 2020, p. 506). Why does the Counterstatement specify that trainers should be "authorized"? Those who deliver training are ethically and legally obligated to ensure that they have the knowledge, skills, and experience to do so. Although it is not uncommon for psychological test developers to develop programs to accredit or certify trainers, it is highly unusual to state that all training must be done only by people accredited or certified by a specific entity. Hare and colleagues have recently acknowledged that individuals going through this training program produced reliability estimates that "did not meet the standard recommended for criminal cases" (Blais, Forth, & Hare, 2017, p. 762).

Conclusion

The Statement by DeMatteo et al. (2020) was narrowly focused on the question of whether the PCL-R should be used to make predictions of serious institutional violence in the context of capital sentencing evaluations. Based on a review of the relevant literature, we concluded that one cannot use PCL-R scores to make such predictions with adequate precision or accuracy to justify their use for this purpose. In their Counterstatement, Olver et al. (2020) raised numerous concerns about the Statement. We suggest that these concerns are either irrelevant to the circumscribed issue (indeed the only issue) addressed in the Statement, or that they reflect a misunderstanding or mischaracterization of the Statement. Even when viewed in its most favorable light, the Counterstatement does not provide meaningful evidence or a convincing rationale to refute the Statement's conclusion that the PCL-R should not be used to make predictions of serious institutional violence in capital contexts.

References

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Law Institute. (1985). *Model penal code*. Retrieved from https://www.legal-tools.org/doc/08d77d/pdf

Blais, J., Forth, A. E., & Hare, R. D. (2017). Examining the interrater reliability of the Hare Psychopathy Checklist-Revised across a large sample of trained raters. *Psychological Assessment, 29,* 762–775. http://dx.doi.org/10.1037/pas0000455

Bright, S. B. (2015). *Capital punishment: Race, poverty, and disadvantage*. New Haven, CT: Yale University. Retrieved from http://campuspress.yale.edu/capitalpunishment/

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Cunningham, M. D., Reidy, T. J., & Sorensen, J. R. (2016). Wasted resources and gratuitous suffering: The failure of a security rationale for death row. *Psychology, Public Policy, and Law, 22,* 185–199. http://dx.doi.org/10.1037/law0000072

Cunningham, M. D., & Sorensen, J. R. (2010). Improbable predictions at capital sentencing: Contrasting prison violence outcomes. *Journal of the American Academy of Psychiatry and the Law, 38,* 61–72.

Cunningham, M. D., Sorensen, J. R., & Reidy, T. J. (2009). Capital jury decision-making: The limitations of predictions of future violence. *Psychology, Public Policy, and Law, 15,* 223–256. http://dx.doi.org/10.1037/a0017296

Cunningham, M. D., Sorensen, J. R., Vigen, M. P., & Woods, S. O. (2011). Life and death in the Lone Star State: Three decades of violence predictions by capital juries. *Behavioral Sciences & the Law, 29,* 1–22. http://dx.doi.org/10.1002/bsl.963

DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S., . . . Reidy, T. J. (2020). Statement of concerned experts on the use of the Hare Psychopathy Checklist-Revised in capital sentencing to assess risk for institutional violence. *Psychology, Public Policy, and Law, 26,* 133–144. http://dx.doi.org/10.1037/law0000223

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., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20 V3: Professional guidelines for evaluating risk of violence*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

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., 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

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.

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.

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology, 3,* 121–137. http://dx.doi.org/10.1111/j.2044-8333.1998.tb00354.x

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare PCL: Screening version*. North Tonawanda, NY: Multi-Health Systems.

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. *Law and Human Behavior, 21,* 347–359. http://dx.doi.org/10.1023/A:1024851017947

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. http://dx.doi.org/10.2307/2529310

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., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15,* 19–53. http://dx.doi.org/10.1037/a0014897

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Olver, M. E., Stockdale, K. C., Neumann, C. S., Hare, R. D., Mokros, A., Baskin-Sommers, A., . . . Yoon, D. (2020). Reliability and validity of the Psychopathy Checklist-Revised in the assessment of risk for institutional violence: A cautionary note on DeMatteo et al. (2020). *Psychology, Public Policy, and Law, 26,* 490–510. http://dx.doi.org/10.1037/law0000256

Pogrow, S. (2019). How effect size (practical significance) misleads clinical practice: The case for switching to practical benefit to assess applied research findings. *The American Statistician, 73,* 223–234. http://dx.doi.org/10.1080/00031305.2018.1549101

Reidy, T. J., Sorensen, J. R., & Cunningham, M. D. (2012). Community violence to prison assault: A test of the behavioral continuity hypothesis. *Law and Human Behavior, 36,* 356–363. http://dx.doi.org/10.1037/h0093934

Reidy, T. J., Sorensen, J. R., & Cunningham, M. D. (2013). Probability of criminal acts of violence: A test of jury predictive accuracy. *Behavioral Sciences & the Law, 31,* 286–305. http://dx.doi.org/10.1002/bsl.2064

Reynolds, C. R. (2016). Contextualized evidence and empirically based testing and assessment. *Clinical Psychology: Science and Practice, 23,* 410–416. http://dx.doi.org/10.1111/cpsp.12181

Roser, M., Appel, C., & Ritchie, H. (2019). *Human height*. Retrieved from https://ourworldindata.org/human-height

Skipper v. South Carolina, 476 U.S. 1 (1986).

Smith, J. M., Gacono, C. B., Fontan, P., Cunliffe, T. B., & Andronikof, A. (2020). Understanding Rorschach research: Using the Mihura (2019) commentary as a reference. *SIS Journal of Projective Psychology & Mental Health, 27,* 71–82.

Sorensen, J. R., & Cunningham, M. D. (2010). Conviction offense and prison violence: A comparative study of murderers and other offenders. *Crime & Delinquency, 56,* 103–125. http://dx.doi.org/10.1177/0011128707307175

Steadman, H. J., Mulvey, E. P., Monahan, J., Robbins, P. C., Appelbaum, P. S., Grisso, T., . . . Silver, E. (1998). Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. *Archives of General Psychiatry, 55,* 393–401. http://dx.doi.org/10.1001/archpsyc.55.5.393

Received June 10, 2020
Revision received June 25, 2020
Accepted June 25, 2020 ■