

# Reliability and Validity of the Psychopathy Checklist-Revised in the Assessment of Risk for Institutional Violence: A Cautionary Note on DeMatteo et al. (2020)

Mark E. Olver  
University of Saskatchewan

Craig S. Neumann  
University of North Texas

Andreas Mokros  
University of Hagen

Eddy Brand  
Ministry of Justice, The Hague, the Netherlands

Carl Gacono  
Asheville, North Carolina

Kent Kiehl  
University of New Mexico and Mind Research Network, a  
Partner with Lovelace Biomedical, Inc., Albuquerque, New  
Mexico

Elizabeth Leon-Mayer  
National University of La Plata

J. Reid Meloy  
University of California, San Diego, and San Diego  
Psychoanalytic Center, San Diego, California

Randall T. Salekin  
University of Alabama

Nicholas Thomson  
Virginia Commonwealth University

Michael Vitacco  
Augusta University

Keira C. Stockdale  
Saskatoon Police Service, Saskatoon, Saskatchewan, Canada,  
and University of Saskatchewan

Robert D. Hare  
University of British Columbia

Arielle Baskin-Sommers  
Yale University

Jorge Folino  
National University of La Plata

Nicola S. Gray  
Swansea University

Raymond Knight  
Brandeis University

Matt Logan  
HALO Forensic Behavioural Specialists, Abbotsford, British  
Columbia, Canada

Sandeep Roy  
University of North Texas

Robert Snowden  
Cardiff University

Scott Tillem  
Yale University

Dahlmyn Yoon  
University of Hagen

A group of 13 authors (GA) shared a statement of concern (SoC) warning against the use of the *Hare Psychopathy Checklist-Revised* (PCL-R; Hare, 1991, 2003) to assess risk for serious institutional violence in U.S. capital sentencing cases (DeMatteo et al., 2020). Notably, the SoC was not confined to capital sentencing issues, but included institutional violence in general. Central to the arguments

---

 Mark E. Olver, Department of Psychology, University of Saskatchewan; Keira C. Stockdale, Saskatoon Police Service, Saskatoon, Saskatchewan, Canada, and Department of Psychology, University of Saskatchewan;  Craig S. Neumann, Department of Psychology, University of North Texas;  Robert D. Hare, Department of Psychology, University of British Columbia;  Andreas Mokros, Department of Psychology, University of Hagen; Arielle Baskin-Sommers, Department of Psychology, Yale University; Eddy Brand,

Ministry of Justice, The Hague, the Netherlands;  Jorge Folino, Department of Psychiatry, School of Medicine, National University of La Plata; Carl Gacono, Independent Practice, Asheville, North Carolina;  Nicola S. Gray, Department of Psychology, Swansea University; Kent Kiehl, Department of Psychology, University of New Mexico, and Mind Research Network, a Partner with Lovelace Biomedical, Inc., Albuquerque, New Mexico;

*continued*

presented in the SoC was that the PCL-R has poor predictive validity for institutional violence and also inadequate field reliability. The GA also identified important issues about the fallibility and inappropriate use of any clinical/forensic assessments, questionable evaluator qualifications, and their effects on capital sentencing decisions. However, as a group of forensic academics, researchers, and clinicians, we are concerned that the PCL-R represents a psycholegal red herring, while the SoC did not sufficiently address critical legislative, systemic, and evaluator/rating issues that affect all forensic assessment tools. We contend that the SoC's literature review was selective and that some of the resultant opinions about uses and misuses of the PCL-R were potentially misleading. We focus our response on the evidence and conclusions proffered by the GA concerning the use of the PCL-R in capital and other cases. We provide new empirical findings regarding the PCL-R's predictive validity and field reliability to further demonstrate its relevance for institutional violence risk assessment and management. We further demonstrate why the argument that group data cannot be relevant for single-case assessments is erroneous. Recommendations to support the ethical and appropriate use of the PCL-R for risk assessment are provided.

*Keywords:* capital sentencing, field reliability, institutional violence, PCL-R, predictive validity

A group of 13 authors (GA) in forensic psychology issued a statement of concern (SoC) warning against the use of the *Hare Psychopathy Checklist-Revised* (PCL-R; Hare, 1991, 2003) to assess risk for institutional violence in US capital sentencing matters (DeMatteo et al., 2020). In our counterstatement, we critically evaluate the arguments presented by the GA and highlight limitations of their literature review. Also, we provide new empirical findings, both meta-analytic, and latent variable- and person-centered modeling results, to help advance research on this topic and to illustrate the merits of the PCL-R in appraising risk for institutional violence. Finally, we provide recommendations regarding competent use of the PCL-R and other forensic instruments.

At the outset, we note that several members of the GA, and of our group, come from countries without the death penalty. Also, many of the coauthors who helped develop our commentary on the target article do not support the death penalty. Of course, there is an enormous literature that debates the logic, legality, ethics, and morality of the death penalty. Our position is that controversial issues, such as capital punishment, should not obscure the importance of scientific research and empirical evidence for addressing all relevant issues. In line with this position, we focus here on the bases for the evidence and conclusions proffered by the GA

concerning the use of the PCL-R in capital cases. We acknowledge the efforts of the members of the GA, but we respectfully disagree with their characterization of the PCL-R and with their conclusions about its utility in forensic matters.

In general, we agree with the gist of the SoC regarding the fallibility and inappropriate use of clinical/forensic assessments, questionable evaluator qualifications, and their effects on capital sentencing decisions. As a group of clinical/forensic academics and researchers, some with extensive experience working in prisons or forensic-psychiatric hospitals, we are, however, concerned that the PCL-R is being singled out for use as a psycholegal red herring that diverts attention from several broader legislative, systemic, and evaluator/rating issues that contribute to the decisions made about clientele in capital and other sentencing contexts. Blaming the PCL-R or related measures does nothing to fix these issues. The conclusions generated by the GA's selective and limited review of the literature could very well lead to confusion for those in the criminal justice system who must navigate challenging psycholegal issues, particularly those that can be addressed with considerable empirical evidence. Our commentary aims to ensure an accurate representation of the scientific record.

 Raymond Knight, Department of Psychology, Brandeis University;  Elizabeth Leon-Mayer, Department of Psychiatry, School of Medicine, National University of La Plata; Matt Logan, HALO Forensic Behavioural Specialists, Abbotsford, British Columbia, Canada; J. Reid Meloy, Department of Psychiatry, University of California, San Diego, and San Diego Psychoanalytic Center, San Diego, California;  Sandeep Roy, Department of Psychology, University of North Texas;  Randall T. Salekin, Department of Psychology, University of Alabama;  Robert Snowden, Department of Psychology, Cardiff University;  Nicholas Thomson, Departments of Surgery and Psychology, Virginia Commonwealth University;  Scott Tillem, Department of Psychology, Yale University; Michael Vitacco, Medical College of Georgia, Augusta University;  Dahlnym Yoon, Department of Psychology, University of Hagen.

Mark E. Olver, Keira C. Stockdale, Craig S. Neumann, Robert D. Hare, and Andreas Mokros carried out the empirical analyses and summative efforts in coordinating the comments of the remaining authors who are listed alphabetically given substantively equal contributions to this counter statement. The views and opinions expressed here are those of the authors and do not

represent the views or opinions of their respective institutions, their psychology advocacy and regulatory bodies, or this journal's editorial board. As with DeMatteo et al. (2020), the language and contents of the document represent a compromise among our group of authors to generate a collective response, and individual authors may have a preferred mode of communicating substantive issues or nuances in opinion that vary from this commentary. We thank Stephen C. P. Wong for helpful comments and resources on substantive matters relating to the PCL-R and its use in violence risk assessment. We also thank the editor, Michael Lamb, for his feedback and the opportunity to participate in this exchange, as well as DeMatteo et al. (2020) for their helpful and constructive comments on a draft of this article. Robert Hare receives royalties from the sale of the PCL-R and its derivatives. Andreas Mokros is the first author of the German version of the Hare Psychopathy Checklist-Revised and receives royalties from its sale.

Correspondence concerning this article should be addressed to Craig S. Neumann, Department of Psychology, University of North Texas, 1155 Union Circle, #311280, Denton, TX 76203. E-mail: [craig.neumann@unt.edu](mailto:craig.neumann@unt.edu)

## Background and Context

The primary argument advanced in the SoC was that the PCL-R should not be used to predict serious institutional violence in *capital sentencing* matters. Yet, it states, “In this article, we are focusing specifically on the use of the PCL-R to *predict serious (i.e., non-trivial) violence in high-security correctional settings*” (DeMatteo et al., 2020, p. 137; emphases added). The purpose of the SoC may be to inform the court that the use of the PCL-R is not warranted in assessments of *institutional and postrelease* violence; however, the arguments in the SoC may have severe and unwarranted implications for criminal justice, including capital sentencing matters.

There are several issues embedded in the GA’s arguments. The first issue concerns evidence for the efficacy of the PCL-R in the prediction of “serious” institutional violence. Despite the GA’s stated focus on this topic, they did not provide a clear operationalization of “serious” beyond calling it “non-trivial” (DeMatteo et al., 2020, p. 137). The problem with this focus is that it diminishes the seriousness of other acts of aggression or institutional misconduct that may not necessarily result in physical injury but nonetheless could cause harm or pose serious safety, security, or management concerns. Such examples include hostage-taking of a staff member, threatening to harm family members of the staff by associates in the community, setting a cell on fire or flooding it, and even throwing feces through a meal slot into the face of correctional officers which can result in the transmission of infectious diseases (e.g., hepatitis). These, and numerous other examples, would not appear to qualify as “serious” in the SoC sense, because they may not directly result in physical harm to the victim. However, a range of injurious acts, including those that cause significant psychological trauma, is perpetrated by persons with elevated psychopathic traits. Such harmful acts are captured by predictions of serious institutional misconducts, general violence/aggression, or a general misconduct category. As such, the PCL-R has important implications for management of offenders in maximum security, and it would seem unwise and unsafe for prison personnel not to be aware of psychopathic propensities. For these reasons, in our commentary, we will consider the evidence for the PCL-R, relative to other tools, in the prediction of all forms of institutional misconduct, including acts of physical aggression.

Second, the SoC underspecifies the use of the term “predict.” The purpose of risk assessment includes risk management and violence prevention, not just a determination of the likelihood of target behaviors (Meloy, 2015). The issue of using a tool to “predict” an outcome is very much different from assessing risk for an unwanted result and then using the assessment data to manage risk to prevent the outcome. We address this issue in our recommendations.

Third, the SoC appears to be critical of the PCL-R, but reference to the *PCL: Screening Version* (PCL: SV; Hart, Cox, & Hare, 1995), which is strongly related to the PCL-R conceptually and empirically (Cooke, Michie, Hart, & Hare, 1999; Guy & Douglas, 2006; Higgs, Tully, & Browne, 2018), is noticeably scant. This is puzzling, given that Guy and Douglas (2006, p. 229) concluded that “the PCL: SV has a robust relationship to the PCL-R at both the global and factor levels, and that this relationship holds across coding methods and rater (in)dependence.” Thus, most meta-analyses do not distinguish between the PCL-R and PCL: SV. Our

point here is that: (a) users of either tool may be confused that the concerns raised by the GA pertain only to the PCL-R and not the PCL: SV; (b) some institutions may use the PCL: SV instead of the PCL-R to assess psychopathy; and (c) excluding one measure or the other could lead to biased meta-analytic parametric estimates. As such, we consider meta-analytic evidence from both tools to be relevant and that this evidence is relative to other purpose-built risk assessment tools.

Notably, the SoC does not address the use of other structured tools to assess risk for institutional violence in capital sentencing hearings. These include the *Historical Clinical Risk-20<sup>V3</sup>* (HCR-20<sup>V3</sup>), the *Sexual Violence Risk-20* (SVR-20), the *Level of Service Inventory—Revised* (LSI-R), the *Violence Risk Scale* (VRS), the *Violence Risk Appraisal Guide* (VRAG), the *Static-99*, and the *Lifestyle Criminality Screening Form* (LCSF). Relatedly, the SoC does not contain any commentary on the use of neuroimaging in these hearings (Aspinwall, Brown, & Tabery, 2012; Farahany, 2016; Rummel, Glenn, & Cox, 2019; Umbach, Berryessa, & Raine, 2015). Further, the members of the GA do not state if it is inadvisable to use these methods to assess risk for “serious” institutional violence, for institutional violence in general, or in capital sentencing proceedings.

A fourth and central issue in the SoC pertains to the use of the PCL-R for capital sentencing. The GA notes that US States that accept the death penalty differ on the admissibility of “future dangerousness” in capital sentencing. 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 (Bright, 2015). The use of an instrument in this context is different from the use of a tool for the broader purpose of assessing risk for institutional violence in different settings. The GA does not provide a clear opinion on whether or not the PCL-R should be used to assess risk in a more general context of institutional outcomes. However, several states with the death penalty indicate that future dangerousness refers not only to prison violence but also to violence in society (e.g., *Lawlor v. Commonwealth*, 738 S.E.2d 847, 2013). In such jurisdictions, including Texas, the likelihood of postrelease violence is relevant to *evaluations* of future dangerousness, even if the chances of release are minimal or nil.

Personal views about the death penalty aside, we do not support the use of any *single* tool to make categorical “predictions” about an outcome, “serious” institutional violence, or otherwise. We do, however, support the comprehensive assessment of risk for institutional violence, incorporating the PCL scales as one of several appropriate measures, if only to address a personality propensity relevant to violent behaviors. This approach is much different from the use of only one instrument or technology to make life or death decisions in a legal case.

Ultimately, research should focus on determining the optimal ways of combining various assessments to maximize predictive accuracy for specific decisions and to avoid contamination of multiple assessment biases (Grove & Meehl, 1996). Importantly, the SoC does not present a viable alternative to the use of the PCL-R, although the court likely will request information from experts about the continued dangerousness of the offender. Nor does it appear to express concerns about the equally problematic introduction of expert conclusions of *low risk* based on questions about an offender’s age, education, past criminality, employment

history, and so forth (see Heilbrun, Fairfax-Columbo, Wagage, & Brogan, 2017, p. 118),<sup>1</sup> which may generate false negatives in the absence of a formalized comprehensive risk appraisal employing measures with known psychometric properties.

In our view, the GA cites literature that it believed provides a “proof of absence” regarding the usefulness of the PCL-R to assess risk for institutional violence. It focuses on two sets of psychometric properties of the PCL-R: (a) its predictive validity for “serious” institutional violence and (b) its field reliability. We review their arguments and the literature they cited, and provide a synopsis of key findings relevant to these arguments.

### Predictive Accuracy of the PCL Scales for Institutional Violence

The SoC states that the PCL-R lacks “precision or accuracy” in predicting serious institutional violence (DeMatteo et al., 2020, p. 4). To support this contention, the GA reviewed a set of four meta-analyses that have examined the association between scores on the PCL instruments and institutional misconducts: Guy, Edens, Anthony, and Douglas (2005), Walters (2003a, 2003b), Leistico, Salekin, DeCoster, and Rogers (2008), and Campbell, French, and Gendreau (2009). First, the GA cites the meta-analysis by Guy et al. (2005) as one argument for poor predictive validity, focusing on the prediction of institutional, physical aggression. Second, the document reviews Walters (2003a, 2003b) but dismisses this pair of articles because they examined only general violence but did not examine “serious” institutional violence as a separate outcome. Third, the GA cites Campbell et al. (2009) as showing that various risk tools had better predictive accuracy for general violent recidivism than did the predictive accuracy of the PCL-R for institutional violence. Fourth, it cites Leistico et al. (2008) as showing a weak association between the PCL-R and violence. And fifth, the GA cites several individual studies (Camp, Skeem, Barchard, Lilienfeld, & Poythress, 2013; Hogan & Olver, 2016; McDermott, Edens, Quanbeck, Busse, & Scott, 2008; Morrissey et al., 2007; Walters & Mandell, 2007), published since the most recent meta-analysis that suggest a weak association between PCL-R scores and institutional violence.

There are several issues with the GA’s critique and review of the evidence. First, the SoC does not provide any discussion regarding a threshold of acceptable predictive accuracy or guidelines for interpretation. Nor does it include a definition of “precise” or identify what forensic assessment instruments happen to have achieved the threshold of “precise” in the prediction of this outcome. We argue that “precision” is an equivocal concept that varies widely in the measurement of psychological constructs or in risk assessment; it is vague and does not provide a useful threshold. Yet a further concern is that the SoC does not define “accuracy.” Borrowing from Morrison’s (2011) description of forensic trace evidence, the psychological term *reliability* would match the notion of *precision*, whereas the psychological concept of *validity* would be synonymous with *accuracy*. As such, in this response we use the thresholds based on Cohen (1992) and Rice and Harris (2005):  $r_{pb} = .10$ , area under the curve [AUC] = .56, and  $d = .20$ , are small effects;  $r_{pb} = .24$ , AUC = .64, and  $d = .50$ , are medium effects; and  $r_{pb} = .37$ , AUC = .71, and  $d = .80$ , are large effects. Even with these guidelines in mind, the GA did not identify what level of accuracy is desirable for a measure to be useful in

assessing risk for institutional violence or in capital sentencing. Table 1 provides a meta-meta-analysis of PCL measures in the prediction of institutional outcomes; this includes the most recent Hogan and Ennis study (2010), not cited in the SoC, and Edens and Campbell (2007), reflecting youth samples with variants of the PCL measures, thus adding to the robustness of the meta-analytic effects.

In reviewing the evidence, we must consider methodologies and context. For example, it is important to note that Guy et al. (2005) used point biserial correlation ( $r_{pb}$ ) as the measure of effect size (ES;  $r_{pb} = .17$ ), which is attenuated by low base rates (Babchishin & Helmus, 2016). Physical violence in institutions is less common than other forms of aggression, which means that most attempts to “predict” it will be wrong (i.e., false positives). Therefore, the  $r_{pb} = .17$  is a small-to-low moderate effect, partway between .10 and .24. Guy et al. (2005) did not report the base rate of physical violence in their meta-analysis, so a direct conversion to AUC or  $d$ , adjusting for base rates, cannot be done. The most conservative estimate would be  $d = 0.35$  (assuming 50% base rate), or about 1/3 of a standard deviation.<sup>2</sup> That means that there is an almost 3-point difference in PCL-R scores between people who commit acts of physical violence and those who do not. This effect size is more accurately captured as small to medium, and is not trivial, and certainly not “negligible,” as stated in the SoC (DeMatteo et al., 2020, p. 137). The SoC does not indicate that the  $r_{pb}$  was .26 for verbal/destruction and .23 for general aggression. These both are relatively higher base rate outcomes, so naturally, the  $r$  will be higher, and corresponding  $d$ s = 0.52 and 0.46 (without correction for base rate), which is moderate in the magnitude of prediction. All effect sizes were significant ( $p < .001$ ). So, is this good enough? It is unclear, given the SoC does not provide criteria for what is acceptable. It also is worth noting that Guy et al. (2005) examined the PCL-R with the PCL: SV and the PCL, which were subsumed under the common term, “PCL-R.”

Moreover, the SoC did not include discussion of studies by Walters (2003a, 2003b) because the pair of meta-analyses presented did not focus on “serious” institutional violence. We suggest it is unsound to dismiss studies of institutional aggression. Thus, it is important to note that Walters (2003a, 2003b) found the  $r_{pb}$  for institutional violence was = .12 for PCL-R Factor 1, and .22 for Factor 2, and the  $r_{pb}$  for general institutional adjustment was .18 for Factor 1 and .27 for Factor 2. These effect size (ES) values are in line with those reported by Guy et al. (2005).

The GA also cited the Leistico et al. (2008) meta-analysis as providing evidence for weak predictive validity for institutional

<sup>1</sup> Cunningham and Sorensen (2010) argued that a brief list of demographic variables could provide “highly reliable estimates of an improbability of future serious violence” (p. 71). This is in “sharp contrast to the decidedly poor predictive accuracy of assertions of probable future violence in prosecution-sponsored expert testimony at capital sentencing.” Along with low base rates of institutional violence among capital offenders, this points to “an obvious conclusion: except in rare instances, only expert assertions of various degrees of the improbability of future serious prison violence by respective capital defendants are reliable or scientifically supportable” (p. 71). With a very low base rate of violence, the most straightforward, but potentially costly, conclusion is low risk.

<sup>2</sup> Note that  $d$  would be larger the more the base rate differed from 50%. With, say, a base rate of 25%,  $r_{pb} = .17$  would reflect a  $d$  score of 0.40. At a base rate of 10%,  $d$  would equal 0.58.

Table 1  
*Meta-Analysis of PCL Meta-Analyses in the Prediction of Institutional Outcomes*

Study	<i>k</i>	<i>N</i>	Measure	Criterion	Metric	Effect size	<i>d</i>
Walters (2003a)	15	NR	PCL total	Institutional adjustment	<i>r</i>	.27	.54
Walters (2003b)	14	NR	Factor 1	Violent infractions	<i>r</i>	.12	.24
			Factor 2			.22	.44
	16	NR	Factor 1	Institutional adjustment	<i>r</i>	.18	.36
			Factor 2			.27	.54
Guy, Edens, Anthony, and Douglas (2005)	22	3502	PCL total	Physical violence	<i>r</i>	.17	.34
	16	2129	Factor 1			.14	.28
	16	2129	Factor 2			.15	.30
	15	2477	PCL total	Verbal/destruction	<i>r</i>	.26	.52
	9	1073	Factor 1			.20	.40
	9	1073	Factor 2			.24	.48
	31	4483	PCL total	General aggression	<i>r</i>	.23	.46
	22	2786	Factor 1			.15	.30
	22	2786	Factor 2			.20	.40
	38	5381	PCL total	Total/any	<i>r</i>	.29	.58
	25	3219	Factor 1			.21	.42
	25	3219	Factor 2			.27	.54
Edens and Campbell (2007) <sup>a</sup>	10	1001	PCL total	Physical violence	<i>r</i>	.28	.56
		775	Factor 1			.24	.48
		775	Factor 2			.37	.74
	14	1188	PCL total	Aggression	<i>r</i>	.25	.50
		880	Factor 1			.22	.44
		880	Factor 2			.34	.68
	15	1310	PCL total	Total misconducts	<i>r</i>	.24	.48
		1002	Factor 1			.21	.42
		1002	Factor 2			.28	.56
Leistico, Salekin, DeCoster, and Rogers (2008)	45	6137	PCL total	Institutional infractions	<i>d</i>	.53	.53
	30	3898	Factor 1		<i>d</i>	.41	.41
	29	3848	Factor 2		<i>d</i>	.51	.51
Campbell, French, and Gendreau (2009)	5	626	PCL-R	Institutional violence	<i>r</i>	.14	.28
	7	504	PCL: SV		<i>r</i>	.22	.44
Hogan and Ennis (2010)	3	254	PCL-R	Forensic inpatient violence	<i>r</i>	.21	.42
	8	827	PCL: SV		<i>r</i>	.26	.52
	12	1313	PCL combined		<i>r</i>	.26	.52
Meta-meta-analysis	Grand <i>k</i>				<i>r</i>	<i>d</i>	AUC
	4		PCL total	Institutional violence		.23	.45
	2		Factor 1			.19	.38
	2		Factor 2			.26	.52
	2		PCL total	General aggression		.24	.48
	2		Factor 1			.19	.37
	2		Factor 2			.27	.54
	4		PCL total	Any institutional problems		.27	.53
	4		Factor 1			.20	.40
	4		Factor 2			.27	.54

Note. PCL = Psychopathy Checklist; NR = not reported.

<sup>a</sup> Features youth samples assessed with variants of the PCL scales.

violence. Our concern with this conclusion is that [Leistico et al. \(2008\)](#) did not examine predictive validity for serious or general institutional violence, only general institutional problems. Even still, they found a *d* value of 0.53 for PCL total score, 0.41 for Factor 1, and 0.53 for Factor 2; all moderate effects. The ES values were not moderated by setting (i.e., they were consistent between prison and forensic mental health settings), although the ES tended to be higher in Canada and countries outside North America than in the US.

Further, in the SoC, [Campbell et al. \(2009\)](#) is cited as a study that examined prediction of general institutional violence of the Statistical Information on Recidivism (SIR) scale, the VRAG, the HCR-20, the LSI/LSI-R, and the PCL-R and PCL: SV. The *r* for the PCL-R and the PCL: SV was, respectively, .14 and .22. Most

importantly, [Campbell et al. \(2009\)](#) found that the predictive accuracies were not significantly different among any of the instruments, and the confidence intervals overlapped substantially, suggesting that the predictive validity ESs all came from the same population of effect sizes (p. 575). Of note, there were considerably fewer studies examining institutional violence than violent recidivism in the community, so the ESs are less stable. Nevertheless, their meta-analysis showed that the instruments were equivalent in their ability to predict the outcome. In sum, the PCL-R did not fare worse than other tools in the prediction of institutional violence.

We are concerned that the SoC did not provide a full presentation or accurate description of the evidence from these four meta-analyses, all of which generated similar findings and conclusions.

Moreover, the quality of a meta-analysis and the trustworthiness of its conclusions are only as strong as the individual studies used to generate them (Cunliffe et al., 2012; Smith et al., 2018). There is other pertinent literature relevant to the GA's central argument of the PCL-R's predictive validity for institutional violence. For instance, Olver, Stockdale, and Wormith's (2014) meta-analysis of the Level of Service scales showed that the LSI had  $r = .21$  for serious misconduct and  $.24$  for any misconduct. The predictive accuracy values were about moderate in magnitude but consistent with that of the PCL scales for the same type of outcome and also consistent with the Campbell et al. (2009) meta-analysis. Further, Hogan and Ennis (2010) reported the PCL scales ( $r = .26$ ,  $k = 12$ ) and HCR-20 ( $r = .33$ ,  $k = 4$ ) had moderate predictive accuracy for institutional violence and did not significantly differ in their associations with this outcome.

It is also worth discussing the omission of individual studies conducted since the meta-analyses cited in the SoC. In this spirit, we thought it best to be evidence-based and conduct an updated meta-analysis of the prediction of institutional outcomes by the PCL-R and PCL: SV. We focused on: (a) "newer" studies cited in the SoC regarding the predictive properties of the PCL measures (i.e., Camp et al., 2013; Hogan & Olver, 2016; McDermott et al., 2008; Morrissey et al., 2007; Walters & Mandell, 2007); (b) additional studies not cited in the SoC and, to our knowledge, not included in the four previous sets of meta-analyses cited in the SoC (Campbell et al., 2009; Guy et al., 2005; Leistico et al., 2008; Walters, 2003a, 2003b). Most of these were not in Hogan and Ennis (2010), which overlapped with previous meta-analyses; and (c) results of an online literature search of PsycINFO, ProQuest Dissertations and Theses, and Google Scholar using "PCL" and variations on "institutional" or "inpatient" "offending," "recidivism," "misconducts," or "violence." We also examined the reference sections of key works. We converted the ESs to  $d$  via a direct conversion from AUC per Rice and Harris (2005) or from  $r_{pb}$  adjusting for base rates when this information was available. Table 2 provides a synopsis of the new studies, whereas Table 3 contains the results of the updated meta-analysis.

We begin with a brief review of more recent studies cited in the SoC but not included in previous meta-analyses (Camp et al., 2013; Hogan & Olver, 2016; McDermott et al., 2008; Morrissey et al., 2007; Walters & Mandell, 2007). Although the GA presented these studies as illustrations of recent work that repudiates the predictive properties of the PCL-R, the actual findings were more nuanced than as described in the SoC, as scrutiny of Table 2 illustrates.

1. In their psychiatric inpatient sample, Hogan and Olver (2016) found Factor 2 and the Antisocial facet had significant moderate predictive accuracy for institutional aggression (AUCs = .65 and .66), whereas the PCL-R total was .63. They obtained similar findings with a small prospectively assessed sample (Hogan & Olver, 2018).
2. McDermott et al. (2008) found that PCL-R total and Factor 2 scores had significant, moderate predictive validity for aggression toward staff (AUCs = .66), and the same magnitude of prediction for this outcome as the VRAG and HCR-20. AUCs for aggression toward patients and overall were nonsignificant (AUCs = .62 and .58, respectively).
3. Camp et al. (2013) found that the PCL-R total score was a moderate predictor of serious institutional violence (AUC = .65), although it did not function as a predictor of infractions for verbal or physical aggression (AUC = .48). The PCL-R was a better predictor of the most serious violations, and a weaker predictor of less serious ones.
4. In a prospective study, Morrissey and colleagues (2007) reported that the PCL-R and its two factors did not predict any form of aggressive behavior among a small sample ( $N = 51-60$ ) of English intellectually disabled offenders, whereas the HCR-20 had good predictive ability. In an earlier, larger, concurrent study of male intellectually disabled offenders detained in high security in England and Wales ( $N = 202$ ), Morrissey et al. 2005 reported that the PCL-R total, Factor 1, and Factor 2 were significantly correlated with having at least one physically aggressive incident (see Table 2). Notably, the correlation between staff ratings of recent verbal and physical aggression and the PCL-R total, Factor 1, and Factor 2 was, respectively, .45, .40, and .43. The SoC and prior meta-analyses did not cite this large inpatient study.
5. Walters and Mandell (2007) examined the PCL: SV and found it had small nonsignificant effects, comparable in magnitude to Guy et al. (2005) and Campbell et al. (2009), for the prediction of major incident and aggressive incident reports ( $r = .16$  for both) and total incident reports ( $r = .15$ ); AUCs were also computed (see Table 2). Although these effects were not significant, in a series of binomial regression analyses, controlling for age, prior incident reports, and Psychological Inventory of Criminal Thinking Styles score (Walters, 1990), PCL: SV scores significantly incrementally predicted all three sets of institutional outcomes. That is, in a more rigorous set of analyses, the PCL: SV improved predictive outcome.

Thus, in the five "newer" studies that reported "similarly weak effects" discussed by DeMatteo et al. (2020, p. 137), four actually found that the effects were either moderate in magnitude or significant, significant in multivariate analyses controlling for other covariates, and comparable to the ES that other tools yielded (e.g., HCR-20, VRAG). Only Morrissey et al. (2007) found weak nonsignificant effects, but they found significant effect sizes in their larger study Morrissey et al. 2005. The SoC cited one new, small sample German study (Huchzermeier, Bruss, Geiger, Kernbichler, & Aldenhoff, 2008) that provided support for the PCL: SV in the prediction of general institutional misconduct. The sample included 10 inmates with a PCL: SV score of 18 or higher, and nine inmates with a score of 12 or lower. A Mann-Whitney  $U$  test indicated that the high PCL: SV group committed significantly greater misconduct than did the low PCL: SV group ( $U = 14$ , which converts to an AUC of .84).

And so, how does all of this add to the overall picture? As presented in Table 3—an updated summary of meta-analytic find-

Table 2  
 Summary of New PCL-R/PCL: SV Studies Included in Updated Meta-Analysis of Prediction of Institutional Outcomes

Study	N	BR	Sample	Country	Measure	Institutional criterion	Metric	ES	$d^a$					
Abbiati, Palix, Gasser, and Moulin (2019)	52	42%	Prison inmates	Switzerland	PCL-R total	Physical violence	AUC	.78	1.09					
					PCL-R F1			.60	.36					
					PCL-R F2			.82	1.30					
		13%			PCL-R total	Other misconduct	AUC	.65	.55					
					PCL-R F1			.58	.30					
					PCL-R F2			.70	.74					
		37%			PCL-R total	Any misconduct	AUC	.66	.59					
					PCL-R F1			.53	.10					
					PCL-R F2			.76	1.00					
Boccaccini, Turner, Murrie, and Rufino (2012)	38	—	SVP	USA	PCL-R total	Any misconduct max disagreement	AUC	.71	.80					
								Any misconduct minimum disagreement						
Camp et al. (2013)	158	8.9%	Prison inmates	USA	PCL-R total	Proximate serious violence	AUC	.65	.54					
					PCL-R F1			.64	.50					
					PCL-R F2			.61	.40					
					Interpersonal Affective			.67	.62					
					Lifestyle			.57	.25					
					Antisocial			.64	.50					
	83	21.7%	Prison inmates	USA	PCL-R total	Infraction verbal/physical aggression	AUC	.55	.18					
					PCL-R F1			.48	-.07					
					PCL-R F2			.48	-.07					
					Interpersonal Affective			.54	.14					
					Lifestyle			.47	-.10					
					Antisocial			.49	-.03					
Carr, Eggenberger, Crawford, and Rotter (2013)	75	53.3%	Forensic inpatients	USA	PCL: SV total	Incident rate	$r$	.14	.28					
								Serious incidents	$r$					
Endrass et al. (2008)	113	9.3%		Switzerland	PCL-R total	Physical aggression	AUC							
					PCL-R F1			.61	.41					
		27.4%			PCL-R F2	Verbal aggression	AUC	.61	.41					
					PCL-R total			.70	.75					
		25.6%			PCL-R F1			.69	.69					
					PCL-R F2			.67	.62					
Hogan and Olver, 2016	77	30.4%	Forensic inpatients	Canada	PCL-R total	Aggression	AUC	.63	.47					
					PCL-R F1			.60	.37					
					PCL-R F2			.65	.55					
					Interpersonal Affective			.52	.07					
					Lifestyle			.62	.43					
					Antisocial			.63	.47					
Hogan and Olver, 2018	19	52.6%	Forensic inpatients	Canada	PCL-R total	Aggression	AUC	.66	.58					
					PCL-R F1			.76	1.00					
					PCL-R F2			.68	.67					
					Interpersonal Affective			.74	.91					
					Lifestyle			.63	.47					
					Antisocial			.73	.86					
Huchzermeier, Bruss, Geiger, Kernbichler, and Aldenhoff (2008)	19	—	Forensic inpatients	Germany	PCL:SV	Security incidents	AUC	.83	1.36					
								.65	.55					
										.84	1.41			
McDermott, Edens, Quanbeck, Busse, and Scott (2008)	108	28%	Forensic inpatients	USA	PCL-R total	Aggression total	AUC			.58	.29			
					PCL-R F1			.56	.20					
					PCL-R F2			.60	.36					
					Interpersonal Affective			.62	.43					
					Lifestyle			.49	-.04					
		16%				Antisocial	Aggression staff	AUC	.58	.29				
						PCL-R total			.56	.20				
									.66	.59				

Table 2 (continued)

Study	N	BR	Sample	Country	Measure	Institutional criterion	Metric	ES	<i>d</i> <sup>a</sup>
					PCL-R F1			.63	.47
					PCL-R F2			.66	.59
					Interpersonal			.64	.50
					Affective			.55	.18
					Lifestyle			.60	.36
					Antisocial			.64	.50
		22%			PCL-R total	Aggression patients	AUC	.62	.43
					PCL-R F1			.57	.25
					PCL-R F2			.65	.55
					Interpersonal			.65	.55
					Affective			.51	.03
					Lifestyle			.61	.39
					Antisocial			.60	.35
Morrisey et al. (2007)	203	31%	Forensic inpatients with ID	UK	PCL-R total	Physical aggression	<i>r</i>	.18	.40
					PCL-R F1			.05	.11
					PCL-R F2			.26	.58
Morrisey et al. (2007)	60	59.3%	Forensic inpatients with ID	UK	PCL-R total	Interpersonal physical	AUC	.54	.14
					PCL-R F1			.48	-.07
					PCL-R F2			.59	.33
		70%			PCL-R total	Verbal/property	AUC	.49	-.03
					PCL-R F1			.50	.00
					PCL-R F2			.54	.14
Neumann and Baskin-Sommers (2019)	385	46%	Prison inmates	USA	PCL-R total	Violence	AUC	.61	.40
Olver, Azizian, D'Orazio, and Rokop (2019)	119	21.8%	SVP	USA	PCL-R total	Violence	AUC	.64	.50
					PCL-R F1			.52	.07
					PCL-R F2			.65	.55
Vitacco et al. (2009)	152	29%	Forensic inpatients	USA	PCL-R total	Physical	<i>d</i>		.18
					Interpersonal				.03
					Affective				-.08
					Lifestyle				.09
					Antisocial				.47
		53%			PCL-R total	Verbal	<i>d</i>		.44
					Interpersonal				.08
					Affective				.13
					Lifestyle				.48
					Antisocial				.57
					PCL-R total	Any	AUC	.54	.14
					Interpersonal			.50	.00
					Affective			.48	-.07
					Lifestyle			.55	.18
					Antisocial			.64	.50
Walters and Heilbrun (2010)	195	38.5%	Forensic inpatients	USA	Interpersonal	Institutional violence	AUC	.61	.40
					Affective			.59	.32
					Lifestyle			.57	.26
					Antisocial			.63	.47
					Interpersonal	Institutional violence		.53	.10
					Affective			.56	.20
					Lifestyle			.57	.26
					Antisocial			.60	.36
		3.2%			Interpersonal	Severe institutional assaults		.69	.71
					Affective			.71	.80
					Lifestyle			.68	.66
					Antisocial			.78	1.09
Walters and Mandell (2007)	136	11%	Prison inmates	USA	PCL: SV total	Aggressive incidents	AUC	.62	.43
					Interpersonal			.50	.00
					Affective			.63	.47
					Lifestyle			.61	.40
		22.1%			PCL: SV total	Major incidents	AUC	.60	.35
					Interpersonal			.51	.03
					Affective			.56	.21
					Lifestyle			.62	.43

(table continues)

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2 (continued)

Study	N	BR	Sample	Country	Measure	Institutional criterion	Metric	ES	$d^a$
		44.8%			PCL: SV total	Total incidents	AUC	.52	.07
					Interpersonal Affective			.43	-.27
					Lifestyle			.54	.14
								.58	.28

Note. BR = base rates; ES = effect size; PCL-R = Psychopathy Checklist – Revised PCL: SV = Psychopathy Checklist: Screening Version; AUC = area under the curve.

<sup>a</sup>  $d$  values converted from area under the curve (AUC) using Rice and Harris (2005) or computed from  $r$  adjusting for base rate, using the formula provided.

ings—the evidence is clear regarding the predictive validity of the PCL-R for institutional violence at a magnitude that is comparable to findings reported in the meta-analytic literature (Abbiati, Palix, Gasser, & Moulin, 2019; Boccaccini, Turner, Murrie, & Rufino, 2012; Carr, Eggenberger, Crawford, & Rotter, 2013; Endrass, Rossegger, Urbaniok, Laubacher, & Vetter, 2008; Neumann & Baskin-Sommers, 2019; Olver, Azizian, D’Orazio, & Rokop, 2019; Vitacco et al., 2009; Walters & Heilbrun, 2010). Moreover, as the comprehensive perspective in Table 3 shows, the PCL scales have significant predictive associations with all institutional outcomes—serious violence, physical aggression, verbal aggression, general aggression, and general misconducts—at a threshold that is close to moderate in magnitude, and on par with prior meta-analyses, including the results of a meta-meta-analysis. As expected, Factor 2, and its Lifestyle and Antisocial facets, tended to predict better than Factor 1 (Interpersonal and Affective facets), although even for the latter, the predictive outcomes were small but significant.<sup>3</sup>

### Conclusions on PCL-R Predictive Validity for Institutional Violence

We can glean several conclusions from these findings. First, the PCL scales demonstrate predictive validity for institutional violence, including “serious” violence, and do so with robustness (i.e., medium in ES magnitude), comparable with other tools, including those designed to assess risk for violence or different outcomes (see Campbell et al., 2009; Hogan & Ennis, 2010; Olver et al., 2014). As Skeem and Polaschek (in press) have noted, “scores on the PCL-R are strongly associated with scores on purpose-built risk assessment tools—and tend to predict violent recidivism about as strongly as these purpose-built tools.”

Second, the base rate of institutional violence is highly relevant for understanding the significance of the PCL scales. Studies typically find that base rates for serious institutional violence (e.g., severe assaults resulting in death or hospitalization, per Walters & Heilbrun, 2010) are small, though not “trivial,” and general acts of aggression may also be relatively infrequent. For instance, in their study of 1,659 convicted homicide offenders in Texas, with an average time at risk of 22 months, Sorensen and Cunningham (2007, Table 4, p. 550) reported similarly low base rates for *assaultive violations* (8.3%) and *assaults resulting in serious injuries* (2.4%), but considerably higher base rates of *potentially violent acts* (27.3%). They also reported a disproportionately high base rate of assaults (17.1%) among their subsample of death-row sentenced inmates, although fortunately, the base rate was 0% for

assaults resulting in serious injury. Although low frequency events are difficult to predict and can generate false positives, overprojections of risk arguably may also trigger punitive and restrictive measures to manage risk that may paradoxically increase it (Sorensen & Cunningham, 2007). This underscores the need to employ structured validated forensic measures to appraise risk accurately, and to inform humane and effective risk management.

In some cases, persons with elevated psychopathic traits can be managed or can manage themselves at times (Klein Haneveld et al., 2018). Still, persons with high PCL scores are more likely to be violent and to cause problems than people with low PCL scores (Patrick, 2018). Naturally, the tighter the security, the lower the level of violence. Even so, the PCL scales predict institutional violence in tightly controlled (e.g., maximum security) settings. But these considerations, in our view, are a far cry from the “proof of absence” advanced within the SoC (DeMatteo et al., 2020, pp. 134, 143). To further illustrate the link between institutional violence and psychopathy, we provide new analyses of currently unpublished data (Neumann & Baskin-Sommers, 2019) within a modern latent variable modeling framework. These model analyses in combination with our meta-analytic findings strongly challenge the GA’s proof of absence claim.

### An Illustration

Precision, as we suggest, can be grounded in the concept of reliability. In particular, “true” score variance is more readily approximated via latent variable approaches, such as structural equation modeling (SEM), given that error variance is modeled separately from common factor variance (Seara-Cardoso, Queirós, Fernandes, Coutinho, & Neumann, 2019; Yang & Green, 2011). Thus, SEM provides precise estimates of effect sizes, given that true score variance is not confounded with error variance. Moreover, SEMs can be used to model a system of interrelated variables and therefore provide a robust context beyond the simple question of how strongly  $X$  (e.g., PCL-R) is associated with  $Y$  (e.g., violence). At the same time, variable-centered approaches, such as

<sup>3</sup> With respect to Factor 1, we note that there is an increasing literature on its value in predicting violence (Cardona, Berman, Sims-Knight, & Knight, 2020; Storey, Hart, Cooke, & Michie, 2016; Langton, Hogue, Daffern, Mannion, & Howells, 2011; Walters & DeLisi, 2015), instrumental violence (Blais, Solodukhin, & Forth, 2014), and treatment/management responsiveness (Brunner et al., 2019; Sewall & Olver, 2019). As indicated in the section, An Illustration, studies that use structural equation modeling (SEM) indicate that Factor 1 plays an important role in the prediction of violence.

Table 3

Updated PCL-R/PCL: SV Meta Analytic Findings of Predictive Validity for Institutional Outcomes Featuring Studies From Post “Mid-2000s” Not Included in Prior Meta-Analyses

Criterion	PCL measure	<i>k</i>	<i>n</i>	<i>d</i>	95% CI	<i>Q</i>	<i>I</i> <sup>2</sup>
Serious violence <sup>a</sup>	Total	2	343	.62**	[.16, 1.08]	0.27	0.00
	Factor 1	2	343	.58*	[.12, 1.04]	0.25	0.00
	Factor 2	2	343	.55*	[.09, 1.05]	0.88	0.00
	Interpersonal	2	343	.65***	[.19, 1.11]	0.86	0.00
	Affective	2	343	.42	[−.04, .88]	0.28	15.70
	Lifestyle	2	343	.55*	[.09, 1.01]	0.75	0.00
	Antisocial	2	343	.46*	[.00, .92]	0.07	68.84
Physical aggression	Total	9	1,350	.39***	[.27, .51]	7.78	0.00
	Factor 1	7	813	.20*	[.04, .36]	4.05	0.00
	Factor 2	7	813	.52***	[.35, .69]	7.04	14.82
	Interpersonal	5	798	.27***	[.11, .44]	5.58	28.27
	Affective	5	798	.15	[−.02, .31]	3.96	0.00
	Lifestyle	5	798	.25**	[.08, .41]	1.62	0.00
	Antisocial	5	798	.38***	[.21, .54]	1.81	0.00
Verbal aggression	Total	4	152	.35**	[.13, .56]	8.00	62.51
	Factor 1	3	256	.26	[−.02, .54]	6.14	67.42
	Factor 2	3	256	.34*	[.06, .62]	2.70	26.00
	Interpersonal	2	235	.03	[−.24, .30]	0.33	0.00
	Affective	2	235	.09	[−.19, .36]	0.26	0.00
	Lifestyle	2	235	.34*	[.07, .62]	2.33	57.03
	Antisocial	2	235	.47***	[.18, .75]	1.30	22.84
Any aggression	Total	11	1,579	.41***	[.29, .53]	9.56	0.00
	Factor 1	8	906	.25**	[.09, .41]	4.42	0.00
	Factor 2	8	907	.55***	[.39, .72]	7.29	3.95
	Interpersonal	8	1,027	.21**	[.06, .35]	6.21	0.00
	Affective	8	1,026	.19**	[.05, .34]	7.85	10.88
	Lifestyle	8	1,028	.29***	[.15, .44]	4.39	0.00
	Antisocial	7	888	.41***	[.26, .56]	0.86	0.00
Any misconduct	Total	5	320	.35**	[.12, .58]	9.38	57.33
Major misconduct	Total	2	211	.40*	[.04, .77]	0.28	0.00

Note. PCL-R = Psychopathy Checklist – revised; PCL: SV = Psychopathy Checklist: Screening Version.

<sup>a</sup> Facet score effect sizes (ES) were averaged to generate Factor 1, 2, and Total score ES estimates owing to the small *k* for this criterion. Averaging Facet score ES when Factor and Total score ES were not reported did not change the substantive findings. We do not employ this procedure for other outcomes owing to sufficient *k*.

\*  $p \leq .05$ . \*\*  $p \leq .01$ . \*\*\*  $p \leq .001$ .

SEM, only provide information about variables because they involve scores (e.g., traits) aggregated across groups of individuals (Neumann, Vitacco, & Mokros, 2016). Person-centered approaches, such as latent profile analysis (LPA), provide information about individuals in terms of such (trait) scores. For instance, LPA has been used to uncover subtypes of individuals with distinct psychopathic trait profiles and how the subtypes differ across critical external correlates (Hare, Neumann, & Mokros, 2018; Mokros et al., 2015; Mokros, Hollerbach, & Eher, 2020; Neumann et al., 2016; Olver, Sewall, Sarty, Lewis, & Wong, 2015), including violent behavior (Krstic et al., 2018). Thus, LPA can be used to obtain information about persons who differ in the PCL-R subtype profile and then determine how they differ in risk. Latent variable- and person-centered approaches used together can provide valuable information about variables and persons, respectively, each offering unique viewpoints on the link between psychopathic propensities and risk for institutional violence.

The data presented here are from 385 male offenders in a maximum-security facility (Neumann & Baskin-Sommers, 2019). Offender mean age was 32.44 ( $SD = 9.83$ ), and 58% of the sample was non-White. The mean number of years at the current facility was 5.70 ( $SD = 6.20$ ). The mean number of previous violent and

nonviolent crimes, respectively, was 2.16 ( $SD = 1.10$ ) and 2.93 ( $SD = 1.75$ ). The mean PCL-R score was 23.49 ( $SD = 6.54$ ) and 18.7% rated at 30 or above. The ICC interrater reliabilities for total and factor scores were .98-.99 (for 17% of the sample). We used the SEM and LPA approaches, as in our previous research, for the current illustration (Krstic et al., 2018). For our SEM, we included several covariates (age, years in the facility, previous violence, youth conduct disorder symptoms) to provide a robust test of the predictive capacity of the PCL-R factors. Also, to highlight the narrowness of the GA’s approach to delineating ‘serious’ institutional violence, we modeled an institutional disciplinary reports (DRs) latent variable (LV) that included violence against persons, security violations, and other institutional DRs.

Model fit for the SEM was adequate ( $CFI = .90$ ,  $RMSEA = .08$ ) and accounted for 35% of the DR LV variance. As can be seen in Figure 1, PCL-R Factor 1 was a significant predictor of the DR LV, along with age and years in the facility. The Factor 1 prediction parameter ( $beta = .45$ ) was larger than the meta-analytic results presented in this commentary, as would be expected when controlling for measurement error. Noteworthy was that Factor 2 was not a significant predictor, which is not surprising, given the antisocial nature of the sample. Finally, all of the DR indicators

Table 4  
 Meta-Analysis of Field Reliability Studies for PCL-R Total Scores

Study	Sample	Country	<i>N</i> pairs	Metric	ES
1. Boccacini et al. (2012) <sup>a</sup>	SVP civil commitment evaluatees	USA	38	ICC <sub>A1</sub>	.44/.52
2. DeMatteo, Edens, Galloway, Cox, Toney Smith, and Formon (2014)	SVP civil commitment evaluatees	USA	29	ICC <sub>A1</sub>	.58
3. Edens, Boccacini, and Johnson (2010) <sup>b</sup>	SVP civil commitment evaluatees	USA	20	ICC <sub>A1</sub> / <i>r</i>	.42/.78
4. Edens, Cox, Smith, DeMatteo, and Sörman (2015)	Archived legal cases; majority DO evaluatees	Canada	225	ICC <sub>A1</sub>	.59
5. Ismail and Looman (2018)	Treatment referred sexual offenders	Canada	178	ICC <sub>A1</sub>	.90
6. Jeandarme et al. (2017)	Belgian NGRI offenders	Belgium	74	ICC <sub>A1</sub>	.42
7. Langton, Barbaree, Harkins, and Peacock (2006) <sup>c</sup>	Treated sexual offenders	Canada	47	<i>r</i> <sub>s</sub>	.81
8. Levenson (2004) <sup>d</sup>	SVP civil commitment evaluatees	USA	69	ICC <sub>A1</sub>	.72
9. Lloyd, Clark, and Forth (2010) <sup>e</sup>	Archived legal cases; DO and LTO evaluatees	Canada	24	ICC <sub>A1</sub>	.71
10. Matsushima (2016)	General federal offenders	Canada	42	ICC <sub>A1</sub>	.85
11. Miller, Kimonis, Otto, Kline, and Wasserman (2012)	SVP civil commitment evaluatees	USA	313	ICC <sub>A1</sub>	.60
12. Rufino et al. (2012)	SVP civil commitment evaluatees	USA	44	ICC <sub>A1</sub>	.33
13. Sturup et al. (2014)	Life sentenced prisoners	Sweden	27	ICC <sub>A1</sub>	.70

Meta-analysis	<i>k</i>	Studies	Metric	ES [95% CI]	<i>Q</i>	<i>I</i> <sup>2</sup>
Overall						
Primary studies	13	1–13	All IRR	.69 [.66, .72] <sup>***</sup>	119.33 <sup>***</sup>	89.94
	12	1–6, 8–12	ICC <sub>A1</sub>	.68 [.64, .71] <sup>***</sup>	117.27 <sup>***</sup>	95.32
Adjustment for potential overlap <sup>f</sup>	11	1, 2, 4–6, 8–12	All IRR	.69 [.67, .74] <sup>***</sup>	108.99 <sup>***</sup>	91.60
	10	1, 2, 4–6, 9–12	ICC <sub>A1</sub>	.68 [.64, .71] <sup>***</sup>	115.15 <sup>***</sup>	92.18
Canada						
Primary studies	4	5, 7, 9, 10	All IRR	.87 [.84, .90] <sup>***</sup>	9.63 <sup>*</sup>	68.85
	3	5, 9, 10	ICC <sub>A1</sub>	.88 [.85, .91] <sup>***</sup>	7.18 <sup>*</sup>	72.16
With outlier included <sup>g</sup> and adjustment for potential overlap <sup>f</sup>	4	4, 5, 7, 10	All IRR	.78 [.75, .82] <sup>***</sup>	63.88 <sup>***</sup>	95.30
	3	4, 5, 10	ICC <sub>A1</sub>	.78 [.74, .82] <sup>***</sup>	63.63 <sup>***</sup>	96.86
USA						
Primary studies	6	1–3, 8, 11, 12	All IRR	.60 [.54, .65] <sup>***</sup>	11.21 <sup>*</sup>	55.41
	6	1–3, 8, 11, 12	ICC <sub>A1</sub>	.59 [.53, .64] <sup>***</sup>	9.88	49.39
Adjustment for potential overlap <sup>f</sup>	5	1–3, 11, 12	All IRR	.58 [.51, .64] <sup>***</sup>	7.65	47.68
	5	1–3, 11, 12	ICC <sub>A1</sub>	.56 [.49, .62] <sup>***</sup>	5.60	28.61
Europe						
Primary studies	2	6, 13	ICC <sub>A1</sub>	.50 [.34, .64] <sup>***</sup>	3.16	68.34

Note. All Psychopathy Checklist – revised (PCL-R) ratings completed by at least two independent evaluators in a field setting. SVP = Sexually Violent Predator; DO = Dangerous Offender; LTO = Long Term Offender; ES = effect size; *r*<sub>s</sub> = Spearman's rank correlation.

<sup>a</sup> We used the midpoint of min (ICC<sub>A1</sub> = .52) versus max (ICC<sub>A1</sub> = .44) disagreement. <sup>b</sup> Value also corrected for range restriction as reported by authors owing to high sample mean and small *SD*. Pearson *r* approximates the ICC<sub>C</sub>, which does not consider the magnitude of score differences between raters (Edens et al., 2010). As such, results are reported exclusively with ICC<sub>A</sub> and all measures of interrater reliability including or substituting with *r* (All IRR). <sup>c</sup> Langton et al. (2006) only reported the *r* (equivalent to ICC<sub>C</sub> as above); this ES is aggregated with corrected Edens et al. (2010) *r* in the "All IRR" ES aggregation. <sup>d</sup> Levenson (2004) reported the average rater ICC<sub>A1</sub> which was corrected by Murrie et al. (2009) to generate an estimated single rater ICC<sub>A1</sub> = .72. This earlier study employed a time frame that overlapped with Miller et al. (2012), and given the potential for overlap in cases we report ES with and without Levenson's (2004) corrected ICC<sub>A1</sub> included. <sup>e</sup> ICC<sub>A1</sub> value obtained by meta-analysis of three values reported for three sets of different pairs of opposing raters (ICC<sub>A1</sub> = .67, *n* = 15; ICC<sub>A1</sub> = .71, *n* = 5; ICC<sub>A1</sub> = .82, *n* = 7) to generate ICC<sub>A1w</sub> = .71, *n* = 24. This study also overlapped in time frame with Edens et al. (2015), and as above, we report ES with and without Lloyd et al. (2010). Because Edens et al. (2015) is an outlier, but Lloyd et al. (2010) is not, the primary analyses employ Lloyd et al. (2010), while the analyses with outlier included employs Edens et al. (2015) in place of Lloyd et al. (2010). See note g below. <sup>f</sup> ES aggregated using the larger of two independent studies conducted by different authors with overlapping timeframes (see notes d and e). <sup>g</sup> Edens, Penson, Ruchensky, Cox, and Smith (2016) included, Lloyd et al. (2010) removed (as above).

\* *p* < .01. \*\*\* *p* < .001.

had strong and significant factor loadings, but the strongest indicator involved violence against persons. As such, it would be a mistake to narrow one's perspective to only violence against persons when thinking about institutional violence. The SEM results highlight the broad risk that psychopathic traits portend.

To examine institutional violence risk among individuals who vary in their psychopathic propensities, we conducted LPA using mean item PCL-R facet scores and then validated the subtypes using violence against persons and security violation DRs. A three-class LPA solution was optimal given a significant Lo-Mendell-Rubin likelihood ratio test (LMR LRT) between the two- and three-class solutions (*p* < .001), a nonsignificant result for the four-class solution (*p* = .18), and trivial difference in the Bayesian

Information Criterion (BIC) between the three- and four-class solutions (1618 vs. 1613, respectively). Moreover, the three-class model had excellent classification accuracy (.89). Figure 2 shows the three-class results, with 47% of the sample evidencing a prototypic psychopathy profile (elevations on all four PCL-R facets), 39% an externalizing profile (elevated F2), and 14% of the subtypes reflecting a general offender profile (low on all facets). The subtypes did not differ in age (*p* = .39) or race (*p* = .07). Also, the prototypic and externalizing subtypes did not differ in years incarcerated (*p* = .35). Figure 3 shows the PCL-R total score by subtype. The prototypic subtype had a mean PCL-R of 28.37 (*SD* = 3.7), well within the 3-point standard error for the conventional cut-off of 30. Figure 4 displays the most telling set of results.

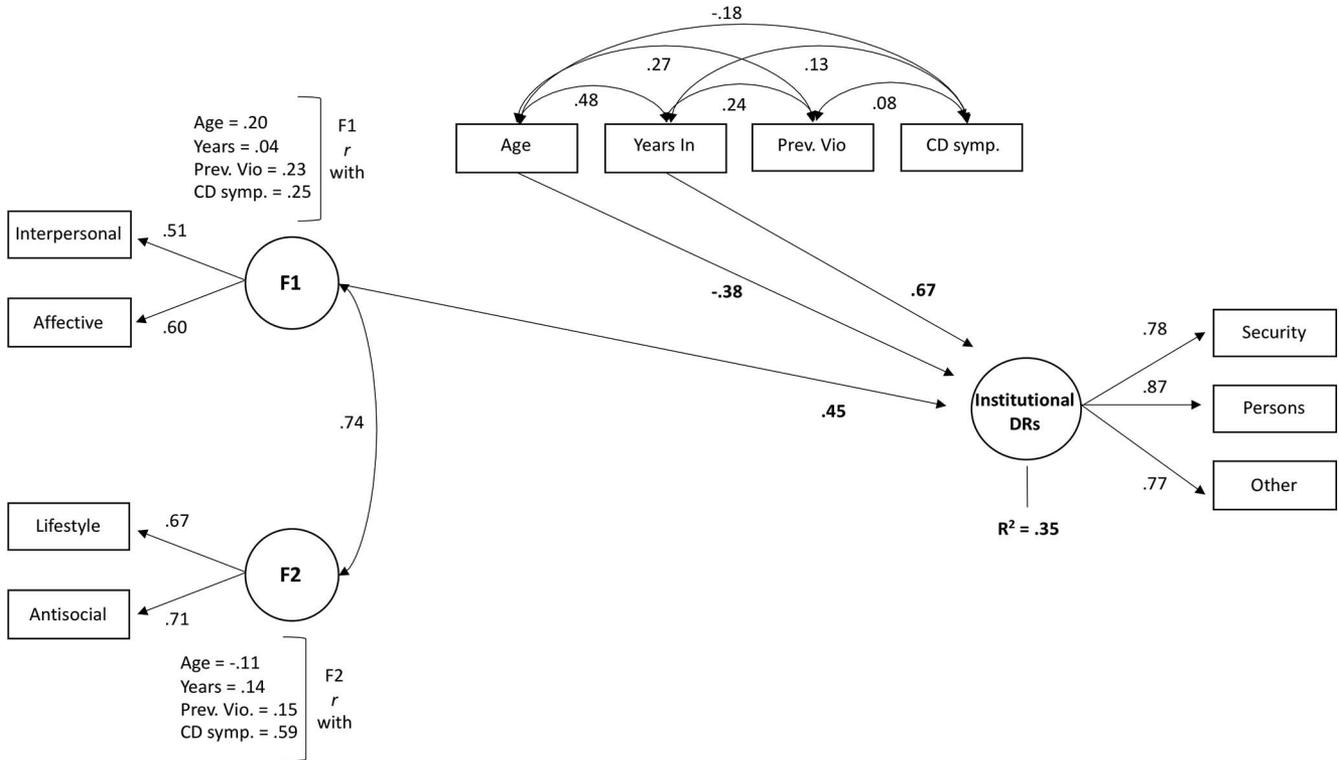


Figure 1. Structural equation modeling results: Factor 1 traits predicting institutional risk.

Concerning violence against persons, both the prototypic and externalizing subtypes had significantly more DRs than the general offenders, but the prototypic produced the stronger effect size ( $d = .63$ ) compared with the externalizing subtype (.51). Also, only the prototypic subtype differed from the general offender subtype for security DRs, signifying the broad risk of prototypically psychopathic individuals.<sup>4</sup> Finally, a synthesis of the SEM and LPA results indicates that it is Factor 1 traits that differentiated the externalizing from prototypic variants and augmented risk for institutional violence. These results clearly challenge the statements written in the SoC regarding a “proof of absence.”

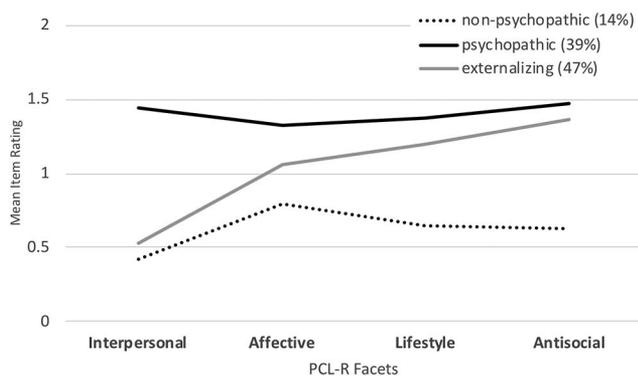


Figure 2. Latent profile analysis results: Psychopathy Checklist-Revised (PCL-R) subtypes as a function of mean item facet score.

### Field Reliability of the PCL Scales

The SoC also did not define the threshold for acceptable reliability of a structured forensic assessment measure to be employed in high stakes psycholegal contexts. DeMatteo et al. (2020) shared concerns that PCL-R scores have the potential for a lack of “probative value or, worse, have a prejudicial impact” that is in part “due to their imperfect interrater reliability (which is, of course, a concern in any evaluation)” (p. 137). But since when did less than “perfect” reliability become the threshold for an unacceptable margin of rater error? Do all other measures have “perfect” reliability? Is the PCL-R or its derivatives any less “perfect”? As a side note, the SoC highlights the Koo and Li (2016) intraclass correlation coefficient (ICC) interpretation guidelines, the most conservative, above that of other well-established guidelines, such as Landis and Koch (1977); Cicchetti and Sparrow (1981), and Fleiss, Levin, and Paik (2003). Koo and Li define .50 to .74 as “moderate,” .75 to .90 as “good,” and .91 to 1.0 as “excellent.” Earlier guidelines tended to define “excellent” as .75 and higher, “good” as .60 to .74, “substantial” as .60-.80, and “fair to good” as .40 to .74.

The GA cited field reliability research to demonstrate that interrater reliability (IRR) is often weak, particularly for the inter-

<sup>4</sup> When selecting cases at or above 30 on the PCL-R total score, versus those below, the elevated cases had significantly more DRs against persons ( $p < .006$ ), but not so for security DRs ( $p = .07$ ), thus attesting to the strength of using PCL-R facet profiles to assess individuals for institutional risk.

personal and affective features of the PCL scales, when completed in adversarial contexts by independent raters (Boccaccini, Turner, & Murrie, 2008, 2014; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Murrie et al., 2009). But this is not always the case, and there are uncited studies or nuanced findings that demonstrated explicitly, or the potential for, strong interrater agreement for the PCL-R in field settings.

To examine the SoC's assertions empirically, we conducted a fixed-effects meta-analysis of PCL-R total scores of published and unpublished field reliability studies that featured two or more PCL-R ratings completed by independent evaluators. We excluded studies that featured evaluations completed by trained student raters (e.g., graduate student rating ICCs from Rufino et al., 2012), field rating simulations (e.g., Blais, Forth, & Hare, 2017; Däderman & Hellström, 2018), or ratings from archival documents under structured conditions in a research setting, many of which report good to excellent interrater reliability (i.e.,  $ICC \geq .75$ ; Cicchetti & Sparrow, 1981; Fleiss et al., 2003; Harris, Rice, & Cormier, 2013).<sup>5</sup> We obtained 13 independent evaluations, most of which reported the intraclass correlation coefficient, absolute agreement single rater ( $ICC_{A1}$ ). We culled studies from (a) a review of the SoC sources, (b) reviews of PCL-R reliability (e.g., Däderman & Hellström, 2018), and (c) an online literature search of PsycINFO, ProQuest Dissertations and Theses, and Google Scholar featuring the search terms "PCL" and "field reliability."

Given that a thorough analysis including the PCL-R factor scores and moderators that affect rater agreement is beyond the scope of our commentary, we limited the meta-analysis to the interrater agreement on the overall sampling of cases in the study. It is noteworthy that the ICC values, or other indexes of rater agreement, were often lower than when other moderators, such as rater training (e.g., Boccaccini et al., 2014), or ratings completed for the same legal side (e.g., Murrie et al., 2009), were considered. When a study with a larger sample (e.g., Boccaccini et al., 2012; Edens, Cox, Smith, DeMatteo, & Sörman, 2015; Rufino et al., 2012) subsumed the same cases of a smaller sample (e.g., Boccaccini et al., 2008; Edens, Penson, Ruchensky, Cox, & Smith, 2016; Murrie et al., 2009), the study with the larger sample (which usually had lower IRR) was employed. Moreover, in one study with range restriction of preselected cases (i.e., all scores above 25; Edens, Boccaccini, & Johnson, 2010), ESs were aggregated with

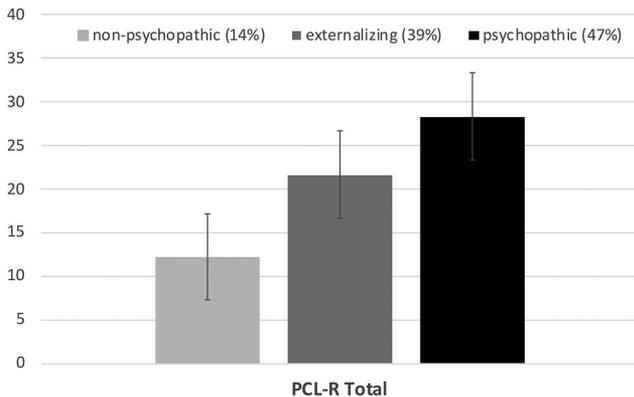


Figure 3. Psychopathy Checklist-Revised (PCL-R) total score as a function of subtype.

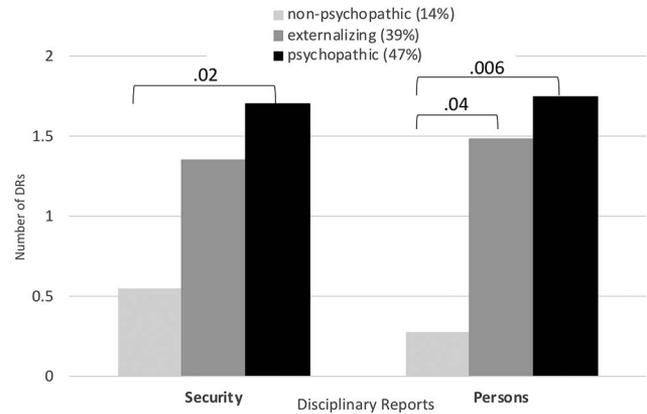


Figure 4. Disciplinary reports against persons and security violations as a function of Psychopathy Checklist-Revised (PCL-R) subtype.

and without the correction for attenuation. Some studies, although completed by independent groups of authors (Edens et al., 2015; Miller et al., 2012), overlapped in the timeframe with a smaller earlier study (Levenson, 2004; Lloyd, Clark, & Forth, 2010) and obtained IRR cases from either the same or a similar catchment source. In instances such as these, where overlap in cases was likely but the degree of overlap could not be established, we aggregated effect sizes with and without inclusion of the smaller earlier study. ES heterogeneity was examined via the  $Q$  statistic and  $I^2$ , the latter of which quantifies ES variability according to thresholds of 25% (small), 50% (medium) and 75% (large; Higgins, Thompson, Deeks, & Altman, 2003). Finally, we identified an ES as an outlier if it was: (a) extreme in value (highest or lowest in an ES); (b) the  $Q$  statistic was significant and  $I^2$  was large; and (c) the single-study finding accounted for more than 50% of the  $Q$  statistic. Thus, given the high standards for study inclusion and efforts to minimize overlap, the present meta-analytic findings are likely a conservative estimate of the PCL-R's field reliability in criminal justice settings.

As presented in Table 4, all studies were published or reported in the mid-2000s to late 2010s. The overall  $ICC_{A1}$ /All IRR magnitudes varied from .68 to .69 across  $k = 10$  to 13 studies, taking into account potential for overlap and true field reliability investigations that did not employ  $ICC_{A1}$  as the reliability metric. In two studies that used  $r$ , either Pearson, which approximates the ICC

<sup>5</sup> Harris et al. (2013) highlighted that PCL-R scores might be more reliable and valid when obtained from extensive file-reviews alone than from interviews plus file reviews. The reason is that highly psychopathic individuals are skilled in the use of positive impression management (PIM), and may be able to manipulate an interviewer into assigning a lowered score. Gillard and Rogers (2015) reported that male jail detainees with a moderate to high Factor 1 score were more successful at using PIM to conceal antisocial behavior and to reduce their scores on several risk instruments, including the HCR-20; thus, the issue of PIM may extend to other tools. In their large meta-analysis, Leistico et al. (2008, p. 35) reported that the ES predicting antisocial behavior was larger for studies that scored the PCL scales from file information ( $d = 0.60$ ) than for studies that used interviews and file data ( $d = 0.52$ ). They advised researchers and clinicians to be cautious in interpreting the "limited predictability of F1 scores . . . which are likely associated with duping the system and escaping documentation of antisocial conduct" (p. 40).

consistency agreement (Edens et al., 2010), or Spearman's (Langton, Barbaree, Harkins, & Peacock, 2006), the resulting ESs were substantively the same. The large  $Q$  and  $I^2$  values indicate substantial ES heterogeneity in the ICCs, which were dubbed "good" or "moderate" by conventional thresholds (Koo & Li, 2016). One obvious result was that the country in which the evaluations were conducted mattered, with  $ICC_{A1}$ /All IRR magnitudes of .78–.88 (Canadian), .56–.60 (U.S.), and .50 (European). This reduced the  $Q$  and  $I^2$  values considerably, whereas most of the heterogeneity in Canadian ES was resolved through removal of an outlier (Edens et al., 2015) that accounted for nearly 90% of the heterogeneity and reduced the  $I^2$  from large to medium. The Canadian ES did not overlap with the other jurisdictions, demonstrating these to be from a different population of ES. It is also noteworthy, however, that all U.S. examinations featured Sexually Violent Predator (SVP) civil commitment samples from one or more of the 21 jurisdictions that employ the statute. This is a psycholegal context that could impact field reliability in different ways, as discussed below. The European ES was lower, falling within "fair to good" or "moderate" by conventional thresholds, although based on only two true field reliability studies, and is likely an unstable estimate of the PCL-R's field reliability in European correctional or forensic mental health contexts.

Importantly, our review demonstrates that good field reliability with the PCL scales can and does happen. Moreover, we note that even when field reliability is low, it may be improved. Boccaccini et al. (2014) demonstrated that completion of formal PCL training from an authorized trainer may improve reliability. Specifically, they found that rater disagreement (as opposed to variability in PCL-R scores) accounted for about 32% of the variance in ICC values, but that this value decreased to 20% among raters who reported having received training from an authorized trainer; that is, up to 80% of variability may have been attributable to differences on the trait measured. Similarly, in a well-designed Swedish field simulation, Dåderman and Hellström (2018) examined PCL-R ratings by mental health professionals with formal PCL-R training and who had access to quality information about the patients. Interrater agreement was strong ( $ICC_{A1} = .89, n = 43$ ).<sup>6</sup>

Further, field reliability is slightly to substantially lower for instruments other than the PCL-R, such as the VRAG ( $ICC_{A1} = .66, r = .76$ , corrected for range restriction; Edens et al., 2016), Static-99 ( $ICC_{A1} = .61$ ; Boccaccini, Murrie, Caperton, & Hawes, 2009;  $ICC_{A1} = .62$ ; Murrie et al., 2009), and Minnesota Sex Offender Screening Tool (MnSOST;  $ICC_{A1} = .68$ ; Boccaccini et al., 2009;  $ICC_{A1} = .44$ ; Murrie et al., 2009). Importantly, these are *objective* static actuarial tools that do not require an interview. It is worth noting, however, that these are "high-stakes" evaluation contexts, such as Dangerous Offender (DO; Canada) and SVP hearings, where adversarial allegiance may be most prevalent and where the sampling of cases is not routine or representative.<sup>7</sup> To this end, Boccaccini et al. (2014) found that independent ratings could have good field reliability for the Static-99 in two large routine correctional samples (Texas,  $N = 600, ICC_{A1} = .79$ ; New Jersey,  $N = 135, ICC_{A1} = .88$ ).

Finally, on the topic of reliability, it is worth noting that in the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition (*DSM-5*) field trials, the PCL-R and psychopathy diagnoses had better reliability than Antisocial Personality Disorder (ASPD;  $\kappa = .22$ ; Freedman et al., 2013). Yet, the courts frequently permit

ASPD diagnosis as evidence in psycholegal matters, which is not discussed in the SoC.

## Additional Arguments and Evidence

### The "Mid-2000s" Psychometric Decline?

There is no evidence that since the mid-2000s there was a sudden dropping off point that is almost taxonomic in nature, where all the predictive validity and interrater reliability data began to turn up null findings that repudiated past efforts. Andrews and Bonta (1994) referred to this practice as "knowledge destruction" characterized by a rejection and/or neglect of sound studies with significant effects, and disproportionate weighting of studies with null effects. A thorough and balanced review of the literature hardly supports "proof of absence" and in our view suggests the contrary. The irony is that all the meta-analyses that supposedly provide a "proof of absence," published between 2003 and 2008, were based on the very works of literature accumulated during the period when things were supposedly rosy (i.e., around or before 2005 or whatever "mid-2000s" represents). In contrast, the results of updated meta-analyses (e.g., here and by Hogan & Ennis, 2010) have been consistent in upholding findings from previous studies of PCL predictive validity for institutional violence and other outcomes. Recent field reliability studies also have demonstrated continuity in reports of strong psychometric properties of the PCL scales, as have controlled investigations using quality information sources and well-trained raters (e.g., Blais et al., 2017; Dåderman & Hellström, 2018; Harris et al., 2013; Rufino et al., 2012).

### Is PCL Field Reliability Invariably and Inexorably Poor?

Our review of the research shows that high interrater reliability occurs with trained raters using high quality information. We are at a loss as to why some might view this as unexpected or undesirable. Field research shows that reliability tends to be relatively low when the quality and consistency of information, and the training and qualifications of raters, are unknown. As noted in our review, Boccaccini et al. (2014) found that having received formal PCL training from authorized trainers resulted in a reduction in rater variance.

Moreover, we accept that field reliability often is, but not inexorably, not as high as it is in research contexts, and sometimes it is notably weaker. As noted in our updated meta-analysis of interrater reliability, there are field reliability studies that show good agreement (some quite substantial) for the PCL measures. The SoC does not mention these studies. That field reliability may be lower than research reliability is not unique to the PCL-R; it occurs for other tools, including the Static-99 and MnSOST (Boc-

<sup>6</sup> We excluded this study from the meta-analysis, given that it was a field simulation of the PCL-R's interrater reliability and not a true field evaluation.

<sup>7</sup> For instance, DeMatteo, Edens, Galloway, Cox, Toney Smith, Present Koller, et al. (2014) note "that state trial court cases, which presumably form the bulk of cases involving the PCL-R, are not routinely published in LexisNexis" (p. 99), which is an electronic database frequently accessed for legal research and practice.

caccini et al., 2009; Edens et al., 2016; Murrie et al., 2009). We argue, though, that the problem of weaker field reliability is at least as much an issue of rater training, information quality and consistency, rater drift, and allegiance effects. We can, and should, address these contributions to measurement error, which are not unique to the PCL-R.

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

Adversarial allegiance is a genuine issue (Simon, Ahn, Stenstrom, & Read, 2020), and it may be one mechanism that reduces field reliability, not only for the PCL scales but also for other instruments, including the Static-99 and the MnSOST, each of which was associated with relatively high scores by the prosecuting side. For instance, in a sample of SVP evaluatees, Murrie et al. (2009) found similar discrepancies between opposing sides of upward of three-quarters of an *SD* for the PCL-R ( $d = 0.78$ ) and the MnSOST ( $d = 0.85$ ). Although the Static-99R showed less allegiance effect, still, there was a one third *SD* higher score ( $d = 0.34$ ) for the prosecuting than for the defense side. It is sobering to see that allegiance effects appear to be endemic to adversarial settings, regardless of the measure employed (see footnote 8 for an example of how defense counsel might use the PCL-R to its advantage in capital sentencing).

Fortunately, adversarial allegiance effects are not inevitable; Edens et al. (2016) did not find evidence of it in a Canadian DO study of the PCL-R (which generated PCL-R  $ICC_{A1} = .82$ ,  $n = 36$ , a subset from Edens et al., 2015), for which they suggest “it is possible if not likely that many experts were appointed by the court rather than retained by prosecutors or defense counsel” (p. 1547). As noted above, when adversarial allegiance effects occur, they affect other tools as well (e.g., Boccaccini et al., 2009; Murrie et al., 2009), not just the PCL-R. We suggest that instead of focusing on banning specific instruments whose use has demonstrated adversarial allegiance effects, we should take steps to try to manage or minimize such effects.<sup>8</sup>

### We Can Apply Group Data to the Individual Case

We were surprised by the assertion in the SoC that group data cannot be used to make predictions about individuals. As succinctly summarized Monahan and Skeem (2016), the notion that one can never use group data and apply it to the individual case, given the unwieldy margins of error, is a “canard.” They cite, with appropriate documentation, that group data routinely are used to make probabilistic statements ranging anywhere from the weather (e.g., a 70% chance of precipitation) to inform insurance rates for individual cases by insurance adjusters. Flawless accuracy is hardly required for risk assessments to be informative, regardless of whether the PCL-R is involved. Statisticians have noted in this regard that the “technical statistical arguments against actuarial risk estimation are simply fallacious” (Imrey & Dawid, 2015, p. 40). Instead, if structured measures can reliably distinguish individuals with higher versus lower probabilities of violence, this can be useful for case planning, sentencing, release decisions and efforts at violence prevention. The LPA results presented above clearly show the increased risk for institutional violence among prototypic versus externalizing psychopathy subtypes relative to general offenders.

One type of opinion leveled against the PCL-R as an indicator of high-risk offenders is grounded in the circular argument that there was no sizable association with recidivism risk within so-called high-risk offenders identified by high PCL-R scores in the first place (Coid, Ullrich, & Kallis, 2013). Demonstrated for some time now in psychometrics, reductions in variance lead to pronounced decreases in correlation. Gulliksen (1950, p. 138) wrote 70 years ago that the fact that an equation on validity depends on restrictions in variance “was first derived by Pearson (1903a). It has also been presented by Kelley (1923c); Holzinger (1928); Thurstone (1931a); Thorndike (1947); Crawford and Burnham (1932), and others.” We illustrate the point with the following. Let us assume that the total score on a screening questionnaire for anorexia had a sizable negative association with daily calorie intake in a nonselect community sample. If we used the same questionnaire with a sample of inpatients with anorexia from an eating disorders clinic, this association would likely vanish because there is little variance in both the independent variable (test score) and the dependent variable (daily calorie intake) in this select sample. The differences among the patients would likely be unsystematic. Goodwin and Leech (2006) provide a numerical example. Hence, as Buchanan (2014) convincingly showed, it is no surprise that Coid et al. (2013) found no sizable association with reoffending for a risk assessment instrument (the HCR-20 Version 2; Webster, Douglas, Eaves, & Hart, 1997) within a subsample of highly psychopathic offenders. On the other hand, if a person-centered approach like LPA was employed, there is a good chance of finding gradations of difference across cases and thus demonstrating valid links between IV and DV.

Moreover, an argument raised against the use of the PCL-R as an indicator of risk for individual cases puts the axe to the roots of psychological assessment. If it were true that predictions about the behavior of individuals were fraught with such uncertainty as to be nearly useless, scholastic aptitude tests, vocational counseling, personnel selection based on achievement scores, and many other areas of applied psychology would be a forlorn enterprise. At best, psychological assessment would be a waste of time and resources; at worst, it would be a detriment to society.

In itself, questioning the epistemological foundations of psychological assessment, in general, is not an argument that the claim

<sup>8</sup> Though seldom discussed, *The California Death Penalty Manual, Volume III* (California Attorneys for Criminal Justice, and California Public Defenders Association, 1998) reprinted the 1991 *PCL-R Interview Schedule*. It provided advice on how defense counsel should handle the PCL-R in sentencing hearings. Briefly, counsel should use the Interview Schedule to determine what sort of PCL-R score the defendant (client) might receive and to decide whether or not to have a defense expert complete a formal assessment. “Obviously, if the answers to these types of [interview] questions are damaging, then the mental health [prosecution] expert should not be exposed to the interview contents, the PCL-R should not be given, and the client should be prepared carefully for any prosecution expert who wants to ask the same questions”. “On the other hand, if the interview and other collateral information suggest the client might obtain a favorable score on the PCL-R, then counsel, after careful consultation with the defense expert, might decide to have the defense expert administer the test and *thereby rule out psychopathy*” (p. 108; italics added). Should the prosecution expert give the client a high PCL-R score, the Manual recommends, among other things, that the defense counsel introduce other diagnoses as mitigating factors, and to use Cunningham’s articles to argue that the PCL-R is not valid with ethnic minorities, females, and adolescents.

(we cannot apply group data to individual cases) is incorrect. Still, it shows that the reasoning put forward by its proponents ought to be very sound. Therefore, let us have a look at their argument in detail. In the article spawning the debate, [Hart, Michie, and Cooke \(2007\)](#) applied a formula that would be appropriate for estimating *confidence intervals* in sample data only to the individual case (i.e., inserting  $n = 1$ ) and noticed exceedingly large margins of error. [Hart et al. \(2007\)](#) concluded the following on the use of actuarial risk assessment instruments (ARAI)s:

At worst, they [i.e., the findings] suggest that professionals should avoid using ARAIs altogether, as the predictive accuracy of these tests may be too low to support their use when making high-stakes decisions about individuals. Low predictive accuracy not only makes reliance on ARAIs ethically problematic, but it also means that they may not meet legal standards for the admissibility of expert or scientific evidence. (p. s64)

Several scholars replied to [Hart et al. \(2007\)](#), including [Hanson and Howard \(2010\)](#) and [Harris, Rice, and Quinsey \(2008\)](#). Indeed, [Mossman](#) wrote that the method chosen by [Hart et al. \(2007\)](#) “pile[s] nonsense on top of meaninglessness” ([Mossman & Sellke, 2007](#), p. 561). The criticism did not, however, deter [Cooke and Michie \(2010\)](#) from reiterating the assertion that group data were inapplicable to individual cases, now focusing on the PCL-R and deriving *prediction intervals* instead of confidence intervals. Based on the exceedingly wide intervals that they purportedly found and citing the previous article by [Hart et al. \(2007\)](#), [Cooke and Michie \(2010\)](#) concluded: “Statistical predictions about individuals will always be poor ([Hart et al., 2007](#)).”

[Scurich and John \(2012\)](#) comprehensively critiqued both kinds of assertions (i.e., wide margins of error in confidence intervals and prediction intervals). First, [Scurich and John](#) made clear that “prediction intervals only apply when a continuous random variable can represent the observations. There is no reasonable interpretation of a prediction interval when the outcome is binary, for there is no purpose in creating intervals around discrete random variables” (p. 240). In other words, there is no meaningful interpretation for prediction intervals around the possible outcomes of reoffending (numerical value: 1) and law-abiding behavior (numerical value: 0) which are not continuous variables (unlike the binomial parameter  $\hat{p}$  which conveys the proportion of recidivists in a given sample). As [Scurich and John](#) continue to elucidate, prediction intervals are about the next observed value to be expected (i.e., 0 or 1 in our case—intervals around which would be meaningless), not about the parameters of a distribution (which would be a case for a confidence interval).

Second, [Scurich and John \(2012\)](#) explained why the notion of a confidence interval, as applied to a single case, was a misnomer. Confidence intervals indicate the range based on a sample parameter estimate  $\hat{p}$  in which the true population parameter  $\pi$  will be situated with a given probability. Thus, a confidence interval applies to a *sample*, not to the individual case.

Instead, according to [Scurich and John \(2012\)](#), one would need to invoke the Bayesian concept of the *credible interval* to gauge how uncertain an individual score is. The use of the Bayesian credible interval, however, necessitates an assumption on plausibility before acknowledging the data of the individual case. In this sense, the credible interval is derived from a posterior distribution that is obtained through: (a) a prior distribution conveying general

knowledge (e.g., concerning the relative rate of recidivists in a given time within a suitable comparison sample); and (b) the discriminatory power of the psychometric instrument in separating recidivists from nonrecidivists (i.e., the likelihood ratio). Ironically, if using a noninformative prior (like the Jeffreys prior), the Bayesian credible interval may look very similar numerically to a frequentist confidence interval, as [Scurich and John](#) pointed out. Using a Jeffreys prior and based on a meta-analysis of recidivism studies, [Mokros, Vohs, and Habermeyer \(2014\)](#) reported a 95%-Bayesian credible interval at a PCL-R score of 25 ranging from 38% to 50%—a margin that is clearly much narrower than the so-called confidence intervals (based on  $n = 1$ ) reported by [Hart et al. \(2007\)](#) for two ARAIs or the so-called prediction intervals provided by [Cooke and Michie \(2010\)](#) for the PCL-R.<sup>9</sup>

### Misuse of the PCL Scales

Any tool can be misused, unfortunately, and we would suggest that this is not a reason to abandon the PCL scales in high stakes psycholegal evaluations. 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](#)). By comparison, sadly, IQ testing has an ignoble history in North America, ranging from the forced sterilization of residential schoolchildren to the deportation of US immigrants. But it has not, and should not, be discontinued, because IQ testing: (a) is a powerful tool that can be used as much for good (e.g., identification of children in need of special services, or supports, such as Binet’s original motivation for development of the Binet-Simon scales); (b) has guidelines for its and other psychoeducational testing’s responsible use that maximizes positive benefits and minimizes harmful effects (e.g., Standards for Educational & Psychological Testing, 2014 [[American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014](#)]); and (c) should never be used in isolation. IQ testing is often coupled with a measure of academic achievement, or even a measure of adaptive functioning, to inform services and accommodations for children and adults. As we highlight in the conclusion, the PCL scales have many useful applications to aid decision-makers and case planning.

<sup>9</sup> [Hart and Cooke \(2013\)](#) repeated the claims from their earlier publication ([Hart et al., 2007](#); [Cooke & Michie, 2010](#)). In the meantime (and regardless of the methodological and conceptual flaws in their argument) the assertions of [Hart et al. \(2007\)](#); [Cooke and Michie \(2010\)](#), and [Hart and Cooke \(2013\)](#) have found their way into legal textbooks and are reiterated in the target article (see statement #23 in the Appendix). The reader who would like to read further on single-case assessments from group data involving Bayesian credible intervals should peruse [Mossman \(2015\)](#) instead.

### Conclusions on Use of the PCL Measures in the Assessment of Risk for Institutional Violence in Psycholegal Evaluations

In summary, our major points of contention with the SoC are as follows:

1. The PCL-R, like other well-established assessment tools, is subject to misuse in clinical/forensic assessments; however, singling it out and discarding it does nothing to solve this problem.
2. Rejection of empirically validated assessment tools for guiding clinical/forensic decisions, whether because of potential misuse or a misguided rejection of using group data to inform individual decisions, is essentially a rejection of science.
3. The SoC's review and synopsis of the predictive validity of the PCL-R for institutional violence were selective, unnuanced, and incomplete. Evidence from meta-analysis and an updated quantitative review demonstrates that the PCL scales have broadly moderate level predictive accuracy for institutional violence that is on par with the level of accuracy of purpose-built risk tools.
4. The SoC's review and synopsis of the field reliability of the PCL-R were similarly selective, unnuanced, and incomplete. Fuller examination of the interrater reliability of PCL-R scores conducted in the field demonstrates that the PCL-R can be a reliable measure of psychopathy.
5. The GA did not provide adaptive solutions for ethical and evidence-informed assessments of risk in capital sentencing and other psycholegal contexts.

What does this all mean? Unfortunately, it appears to us that the PCL-R has become a psycholegal red herring that obscures other legislative, systemic, and evaluator/rating issues contributing to adverse decisions made about clientele in capital sentencing contexts. Blaming the PCL-R or related measures does nothing to fix these fundamental issues. Unfortunately, the SoC did not include guidance on how to address the problem, nor did it provide viable alternatives. Absent of recommendations, readers may walk away, concluding that we should use nothing to assess risk for "serious" institutional violence, especially because the various tools have many common strengths and weaknesses. Instead, we suggest that the PCL-R (and its derivatives) can and should be part of a comprehensive violence risk assessment. We recommend the following for PCL-R users in such evaluations:

1. Do not make life or death recommendations or decisions about an individual based solely on the PCL-R or on any single test or procedure.
2. Exercise extreme caution with harmful or stigmatizing labels such as "psychopath," especially given that: (a) the label "psychopath" is a damaging moniker that can be misconstrued as de facto dangerous, untreatable, or unchangeable; and (b) psychopathy is a dimensional con-

struct, with percentile ranks available to communicate PCL-R scores.

3. 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.
4. The PCL-R should be used with other psychometric measures of risk, need, responsivity, and psychological functioning.
5. PCL-R information should be integrated with data from risk assessment tools to yield comprehensive appraisals of risk to inform risk management and violence prevention efforts.
6. Evaluations and statements of risk should be qualified, contextualized, and informative for decision-makers and those charged with risk management and prevention of violence. This is particularly critical, given the low base rates of serious violence (especially in tightly controlled environments), the dynamic nature of risk (Douglas & Skeem, 2005), and emerging evidence to support the treatability of high PCL scoring individuals (e.g., Caldwell, 2013; Salekin, Worley, & Grimes, 2010; Wong, Gordon, Gu, Lewis, & Olver, 2012).
7. Follow practice guidelines on forensic assessment such as Heilbrun, DeMatteo, Holiday, and LaDuke (2014), Heilbrun (2006), and Dvoskin, Skeem, Novaco, and Douglas (2012) among other authoritative works on violence risk assessment and management. Become familiar with the literature on adversarial allegiance and field reliability, and seek further research, training, consultation, and/or guidance to mitigate its impact.

Finally, we refer readers to the review by Heilbrun et al. (2017) of instruments used in evaluations of risk for violence, including the PCL-R, which concludes that "Risk assessment is relevant in criminal contexts such as capital sentencing, criminal responsibility, and commitment of sexually violent predators" (p. 116) and that the use of specialized measures is strongly indicated, "and even compelled under Daubert" (p. 125). We leave it to the reader to determine how the perspective contained in this quote squares with the content of the SoC and current practice guidelines.

### References

- Abbiati, M., Palix, J., Gasser, J., & Moulin, V. (2019). Predicting physically violent misconduct in prison: A comparison of four risk assessment instruments. *Behavioral Sciences & the Law*, *37*, 61–77. <http://dx.doi.org/10.1002/bsl.2364>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Andrews, D. A., & Bonta, J. (1994–2010). *The psychology of criminal conduct* (1st to 5th ed.). Cincinnati, OH: Anderson Publishing.
- Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of

- psychopaths? *Science*, 337, 846–849. <http://dx.doi.org/10.1126/science.1219569>
- Babchishin, K. M., & Helmus, L. M. (2016). The influence of base rates on correlations: An evaluation of proposed alternative effect sizes with real-world data. *Behavior Research Methods*, 48, 1021–1031. <http://dx.doi.org/10.3758/s13428-015-0627-7>
- 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>
- Blais, J., Solodukhin, E., & Forth, A. E. (2014). A meta-analysis exploring the relationship between psychopathy and instrumental versus reactive violence. *Criminal Justice and Behavior*, 41, 797–821. <http://dx.doi.org/10.1177/0093854813519629>
- Boccaccini, M. T., Murrie, D. C., Caperton, J., & Hawes, S. (2009). Field Validity of the Static-99 and MnSOST-R among sex offenders evaluated for civil commitment as sexually violent predators. *Psychology, Public Policy, and Law*, 15, 278–314. <http://dx.doi.org/10.1037/a0017232>
- Boccaccini, M. T., Murrie, D. C., Rufino, K. A., & Gardner, B. O. (2014). Evaluator differences in Psychopathy Checklist-Revised factor and facet scores. *Law and Human Behavior*, 38, 337–345. <http://dx.doi.org/10.1037/lhb0000069>
- 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>
- Boccaccini, M. T., Turner, D. B., Murrie, D. C., & Rufino, K. A. (2012). Do PCL-R scores from state or defense experts best predict future misconduct among civilly committed sex offenders? *Law and Human Behavior*, 36, 159–169. <http://dx.doi.org/10.1037/h0093949>
- Bright, S. B. (2015). *Capital punishment: Race, poverty, and disadvantage*. Retrieved from <http://campuspress.yale.edu/capitalpunishment/class-3-proportionality-aggravating-circumstances-and-future-dangerousness/>
- Brunner, F., Neumann, I., Yoon, D., Rettenberger, M., Stück, E., & Briken, P. (2019). Determinants of dropout from correctional offender treatment. *Frontiers in Psychiatry*, 10, 142. <http://dx.doi.org/10.3389/fpsy.2019.00142>
- Buchanan, A. (2014). Predicting violent offences by released prisoners. *The British Journal of Psychiatry*, 204, 240. <http://dx.doi.org/10.1192/bjp.204.3.240>
- Caldwell, M. (2013). Treatment of adolescents with psychopathic features. In K. Kiehl & W. P. Sinnott-Armstrong (Eds.), *Handbook of psychopathy and law* (pp. 201–228). New York, NY: Oxford University Press
- California Attorneys for Criminal Justice, and California Public Defenders Association. (1998). *The California death penalty defense manual*, Vol. III. Los Angeles, CA: Author.
- 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>
- Cardona, N., Berman, A. K., Sims-Knight, J. E., & Knight, R. A. (2020). Covariates of the severity of aggression in sexual crimes: Psychopathy and borderline characteristics. *Sexual Abuse*, 32, 154–178. <http://dx.doi.org/10.1177/1079063218807485>
- Carr, W. A., Eggenberger, M., Crawford, L., & Rotter, M. (2013). Prediction of institutional misconduct among civil psychiatric patients: Evaluating the role of correctional adaptations. *Criminal Justice and Behavior*, 40, 541–550. <http://dx.doi.org/10.1177/0093854812456645>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. <http://dx.doi.org/10.1037/0033-2909.112.1.155>
- Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86, 127–137.
- Coid, J. W., Ullrich, S., & Kallis, C. (2013). Predicting future violence among individuals with psychopathy. *The British Journal of Psychiatry*, 203, 387–388. <http://dx.doi.org/10.1192/bjp.bp.112.118471>
- 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>
- Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the Screening Version of the Hare Psychopathy Checklist—Revised (PCL: SV): An item response theory analysis. *Psychological Assessment*, 11, 3–13. <http://dx.doi.org/10.1037/1040-3590.11.1.3>
- Cunliffe, T. B., Gacono, C. B., Meloy, J. R., Smith, J. M., Taylor, E. E., & Landry, D. (2012). Psychopathy and the Rorschach: A response to Wood et al (2010). *Archives of Assessment Psychology*, 2, 1–31.
- 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.
- Dåderman, A. M., & Hellström, A. (2018). Interrater reliability of Psychopathy Checklist-Revised: Results on multiple analysis levels for a sample of patients undergoing forensic psychiatric evaluation. *Criminal Justice and Behavior*, 45, 234–263. <http://dx.doi.org/10.1177/0093854817747647>
- DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Toney Smith, S., & Formon, D. (2014). 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., Toney Smith, S., Present Koller, J., & Bersoff, B. (2014). 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., 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>
- Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law*, 11, 347–383. <http://dx.doi.org/10.1037/1076-8971.11.3.347>
- Dvoskin, J. A., Skeem, J. L., Novaco, R. W., & Douglas, K. S. (2012). *Using social science to reduce violent offending*. New York, NY: Oxford University Press.
- 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., & Campbell, J. S. (2007). Identifying youths at risk for institutional misconduct: A meta-analytic investigation of the Psychopathy Checklist measures. *Psychological Services*, 4, 13–27. <http://dx.doi.org/10.1037/1541-1559.4.1.13>
- 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., Penson, B. N., Ruchensky, J. R., Cox, J., & Smith, S. T. (2016). Interrater reliability of Violence Risk Appraisal Guide scores

- provided in Canadian criminal proceedings. *Psychological Assessment*, 28, 1543–1549. <http://dx.doi.org/10.1037/pas0000278>
- Endrass, J., Rossegger, A., Urbaniok, F., Laubacher, A., & Vetter, S. (2008). Predicting violent infractions in a Swiss state penitentiary: A replication study of the PCL-R in a population of sex and violent offenders. *BMC Psychiatry*, 8, 74. <http://dx.doi.org/10.1186/1471-244X-8-74>
- Farahany, N. A. (2016). Neuroscience and behavioral genetics in U.S. criminal law: An empirical analysis. *Journal of Law and the Biosciences*, 2, 485–509. <http://dx.doi.org/10.1093/jlb/lsv059>
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York, NY: Wiley. <http://dx.doi.org/10.1002/0471445428>
- Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., . . . Yager, J. (2013). The initial field trials of *DSM-5*: New blooms and old thorns. *The American Journal of Psychiatry*, 170, 1–5. <http://dx.doi.org/10.1176/appi.ajp.2012.12091189>
- Gillard, N. D., & Rogers, R. (2015). Denial of risk: The effects of positive impression management on risk assessments for psychopathic and non-psychopathic offenders. *International Journal of Law and Psychiatry*, 42–43, 106–113. <http://dx.doi.org/10.1016/j.ijlp.2015.08.014>
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *Journal of Experimental Education*, 74, 249–266. <http://dx.doi.org/10.3200/JEXE.74.3.249-266>
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2, 293–323. <http://dx.doi.org/10.1037/1076-8971.2.2.293>
- Gulliksen, H. (1950). *Theory of mental tests*. Hoboken, NJ: Wiley. <http://dx.doi.org/10.1037/13240-000>
- Guy, L. S., & Douglas, K. S. (2006). Examining the utility of the PCL:SV as a screening measure using competing factor models of psychopathy. *Psychological Assessment*, 18, 225–230. <http://dx.doi.org/10.1037/1040-3590.18.2.225>
- 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>
- Hanson, R. K., & Howard, P. D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior*, 34, 275–281. <http://dx.doi.org/10.1007/s10979-010-9227-3>
- Hare, R. D. (1991). *Manual for the Revised Psychopathy Checklist*. Toronto, Canada: Multi-Health Systems.
- Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised* (2nd ed.). Toronto, Canada: Multi-Health Systems.
- Hare, R. D., Neumann, C. S., & Mokros, A. (2018). The PCL-R assessment of psychopathy: Development, structural properties, and new directions. In C. Patrick (Ed.), *Handbook of psychopathy* (2nd ed., pp. 39–79). New York, NY: Guilford Press.
- Harris, G. T., Rice, M. E., & Cormier, C. A. (2013). Research and clinical scoring of the Psychopathy Checklist can show good agreement. *Criminal Justice and Behavior*, 40, 1349–1362. <http://dx.doi.org/10.1177/0093854813492959>
- Harris, G. T., Rice, M. E., & Quinsey, V. L. (2008). Shall evidence-based risk assessment be abandoned? *The British Journal of Psychiatry*, 192, 154. <http://dx.doi.org/10.1192/bjp.192.2.154>
- 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 Psychopathy Checklist: Screening version*. Toronto, Canada: 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, s60–s65. <http://dx.doi.org/10.1192/bjp.190.5.s60>
- Heilbrun, K. (2006). *Principles of forensic mental health assessment*. New York, NY: Kluwer.
- Heilbrun, K., DeMatteo, D., Brooks Holiday, S., & LaDuke, C. (Eds.). (2014). *Forensic mental health assessment: A casebook* (2nd ed.). New York, NY: Oxford. <http://dx.doi.org/10.1093/med:psych/9780199941551.001.0001>
- Heilbrun, K., Fairfax-Columbo, J., Wagage, S., & Brogan, L. (2017). Risk assessment for future offending: The values and limits of expert evidence at sentencing. *Court Review*, 53, 116–125. <https://digitalcommons.unl.edu/ajacourtreview/608>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557–560. <http://dx.doi.org/10.1136/bmj.327.7414.557>
- Higgs, T., Tully, R. J., & Browne, K. D. (2018). Psychometric properties in forensic application of the Screening Version of the Psychopathy Checklist. *International Journal of Offender Therapy and Comparative Criminology*, 62, 1869–1887. <http://dx.doi.org/10.1177/0306624X17719289>
- Hogan, N., & Ennis, L. (2010). Assessing risk for forensic psychiatric inpatient violence: A meta-analysis. *Open Access Journal of Forensic Psychology*, 2, 137–147. <https://psycnet.apa.org/doi/10.1037/lhb0000179>
- 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>
- Hogan, N. R., & Olver, M. E. (2018). A prospective examination of the predictive validity of five structured instruments for inpatient violence in a secure forensic hospital. *The International Journal of Forensic Mental Health*, 17, 122–132. <http://dx.doi.org/10.1080/14999013.2018.1431339>
- 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>
- Imrey, P. B., & Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy*, 2, 1–18. <http://dx.doi.org/10.1080/2330443X.2015.1029338>
- Ismail, G., & Looman, J. (2018). Field Inter-Rater Reliability of the Psychopathy Checklist-Revised. *International Journal of Offender Therapy and Comparative Criminology*, 62, 468–481. <http://dx.doi.org/10.1177/0306624X16652452>
- Jeandarme, I., Edens, J. F., Habets, P., Bruckers, L., Oei, K., & Bogaerts, S. (2017). PCL-R field validity in prison and hospital settings. *Law and Human Behavior*, 41, 29–43. <http://dx.doi.org/10.1037/lhb0000222>
- Klein Haneveld, E., Neumann, C. S., Smid, W., Wever, E., & Kamphuis, J. H. (2018). Treatment responsiveness of replicated psychopathy profiles. *Law and Human Behavior*, 42, 484–495. <http://dx.doi.org/10.1037/lhb0000305>
- 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>
- Krstic, S., Neumann, C. S., Roy, S., Robertson, C. A., Knight, R. A., & Hare, R. D. (2018). Using latent variable- and person-centered approaches to examine the role of psychopathic traits in sex offenders. *Personality Disorders: Theory, Research, and Treatment*, 9, 207–216. <http://dx.doi.org/10.1037/per0000249>

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. <https://www.jstor.org/stable/2529310>. <http://dx.doi.org/10.2307/2529310>
- Langton, C. M., Barbaree, H. E., Harkins, L., & Peacock, E. J. (2006). Sex offenders' response to treatment and its association with recidivism as a function of psychopathy. *Sexual Abuse*, *18*, 99–120. <http://dx.doi.org/10.1177/107906320601800107>
- Langton, C. M., Hogue, T. E., Daffern, M., Mannion, A., & Howells, K. (2011). Personality traits as predictors of inpatient aggression in a high-security forensic psychiatric setting: Prospective evaluation of the PCL-R and IPDE dimension ratings. *International Journal of Offender Therapy and Comparative Criminology*, *55*, 392–415. <http://dx.doi.org/10.1177/0306624X10370828>
- Lawlor v. Commonwealth*, 738 S. E. 2d 847. (2013). *Supreme Court of Virginia*.
- 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>
- Levenson, J. S. (2004). Reliability of sexually violent predator civil commitment criteria in Florida. *Law and Human Behavior*, *28*, 357–368. <http://dx.doi.org/10.1023/B:LAHU.0000039330.22347.ad>
- Lloyd, C. D., Clark, H. J., & Forth, A. E. (2010). Psychopathy, expert testimony, and indeterminate sentences: Exploring the relationship between Psychopathy Checklist—Revised testimony and trial outcome in Canada. *Legal and Criminological Psychology*, *15*, 323–339. <http://dx.doi.org/10.1348/135532509X468432>
- Matsushima, Y. (2016). *The inter-rater reliability of the Psychopathy Checklist-Revised in practical field settings*. Unpublished Master's Thesis, Southern Illinois University Carbondale.
- 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>
- Meloy, J. R. (2015). Threat assessment: Scholars, operators, our past, our future. *Journal of Threat Assessment and Management*, *2*, 231–242. <http://dx.doi.org/10.1037/tam0000054>
- 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>
- 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>
- Mokros, A., Hollerbach, P. S., & Eher, R. (2020). Offender subtypes based on psychopathic traits: Results from factor-mixture modeling. *European Journal of Psychological Assessment*. Advance online publication. <http://dx.doi.org/10.1027/1015-5759/a000582>
- Mokros, A., Vohs, K., & Habermeyer, E. (2014). Psychopathy and violent reoffending in German-speaking countries: A meta-analysis. *European Journal of Psychological Assessment*, *30*, 117–129. <http://dx.doi.org/10.1027/1015-5759/a000178>
- Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*, 489–513. <http://dx.doi.org/10.1146/annurev-clinpsy-021815-092945>
- Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*, 91–98. <http://dx.doi.org/10.1016/j.scijus.2011.03.002>
- 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>
- Morrissey, C., Hogue, T. E., Mooney, P., Lindsay, W. R., Steptoe, L., Taylor, J., & Johnston, S. (2005). Applicability, reliability, and validity of the Psychopathy Checklist-Revised in offenders with intellectual disabilities: Some initial findings. *The International Journal of Forensic Mental Health*, *4*, 207–220. <http://dx.doi.org/10.1080/14999013.2005.10471225>
- Mossman, D. (2015). From group data to useful probabilities: The relevance of actuarial risk assessment in individual instances. *Journal of the American Academy of Psychiatry and the Law*, *43*, 93–102. <http://dx.doi.org/10.2139/ssrn.2372101>
- Mossman, D., & Sellke, T. (2007). Avoiding errors about 'margins of error.' *The British Journal of Psychiatry*, *191*, 561. <http://dx.doi.org/10.1192/bjp.191.6.561>
- 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>
- Neumann, C. S., & Baskin-Sommers, A. (2019). *PCL-R prediction of institutional violence: Cheshire sample*. Unpublished raw data.
- Neumann, C. S., Vitacco, M. J., & Mokros, A. S. (2016). Using both variable-centered and person-centered approaches to understanding psychopathic personality. In C. B. Gacono (Ed.), *The clinical and forensic assessment of psychopathy: A practitioner's guide* (2nd ed., pp. 14–31). New York, NY: Routledge.
- Olver, M. E., Azizian, A., D'Orazio, D., & Rokop, J. (2019, November 8). *Institutional behaviors of high psychopathy men in a California SVP program*. Paper presented at the 38th Annual research and treatment conference, Association for the Treatment of Sexual Abusers, Atlanta, GA.
- Olver, M. E., Sewall, L. A., Sarty, G. E., Lewis, K., & Wong, S. C. P. (2015). A cluster analytic examination and external validation of psychopathic offender subtypes in a multisite sample of Canadian federal offenders. *Journal of Abnormal Psychology*, *124*, 355–371. <http://dx.doi.org/10.1037/abn0000038>
- Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*, 156–176. <http://dx.doi.org/10.1037/a0035080>
- Patrick, C. J. (Ed.). (2018). *Handbook of psychopathy* (2nd ed.). New York, NY: Guilford Press.
- Rommel, R. J., Glenn, A. L., & Cox, J. (2019). Biological evidence regarding psychopathy does not affect mock jury sentencing. *Journal of Personality Disorders*, *33*, 164–184. [http://dx.doi.org/10.1521/pedi\\_2018\\_32\\_337](http://dx.doi.org/10.1521/pedi_2018_32_337)
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, *29*, 615–620. <http://dx.doi.org/10.1007/s10979-005-6832-7>
- Rufino, K. A., Boccaccini, M. T., Hawes, S. W., & Murrie, D. C. (2012). When experts disagreed, who was correct? A comparison of PCL-R scores from independent raters and opposing forensic experts. *Law and Human Behavior*, *36*, 527–537. <http://dx.doi.org/10.1037/h0093988>
- Salekin, R. T., Worley, C., & Grimes, R. D. (2010). Treatment of psychopathy: A review and brief introduction to the mental model approach for psychopathy. *Behavioral Sciences & the Law*, *28*, 235–266. <http://dx.doi.org/10.1002/bsl.928>
- Scurich, N., & John, R. S. (2012). A Bayesian approach to the group versus individual prediction controversy in actuarial risk assessment. *Law and Human Behavior*, *36*, 237–246. <http://dx.doi.org/10.1037/h0093973>
- Seara-Cardoso, A., Queirós, A., Fernandes, E., Coutinho, J., & Neumann, C. (2019). Psychometric properties and construct validity of the short version of the Self-Report Psychopathy Scale in a Southern European

- sample. *Journal of Personality Assessment*. Advance online publication. <http://dx.doi.org/10.1080/00223891.2019.1617297>
- Sewall, L. A., & Olver, M. E. (2019). Psychopathy and treatment outcome: Results from a sexual violence reduction program. *Personality Disorders: Theory, Research, and Treatment*, *10*, 59–69. <http://dx.doi.org/10.1037/per0000297>
- Simon, D., Ahn, M., Stenstrom, D. M., & Read, S. J. (2020). The adversarial mindset. *Psychology, Public Policy, and Law*. Advance online publication. <http://dx.doi.org/10.1037/law0000226>
- Skeem, J. L., & Polaschek, D. L. (in press). High risk, not hopeless: Correctional intervention for people at risk for violence. *Marquette Law Review*.
- 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*. Advance online publication.
- Smith, J. M., Gacono, C. B., Fontan, P., Taylor, E. E., Cunliffe, T. B., & Andronikof, A. (2018). A scientific critique of Rorschach research: Revisiting Exner's *Issues and Methods in Rorschach Research* (1995). *Rorschachiana*, *39*, 180–203. <http://dx.doi.org/10.1027/1192-5604/a000102>
- Sorensen, J. R., & Cunningham, M. D. (2007). Operationalizing risk: The influence of measurement choice on the prevalence and correlates of prison violence among incarcerated murderers. *Journal of Criminal Justice*, *35*, 546–555. <http://dx.doi.org/10.1016/j.jcrimjus.2007.07.007>
- Storey, J. E., Hart, S. D., Cooke, D. J., & Michie, C. (2016). Psychometric properties of the Hare Psychopathy Checklist-Revised (PCL-R) in a representative sample of Canadian federal offenders. *Law and Human Behavior*, *40*, 136–146. <http://dx.doi.org/10.1037/lhb0000174>
- 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>
- Umbach, R., Berryessa, C., & Raine, A. (2015). Brain imaging research on psychopathy: Implications for punishment, prediction, and treatment in youth and adults. *Journal of Criminal Justice*, *43*, 295–306. <http://dx.doi.org/10.1016/j.jcrimjus.2015.04.003>
- Vitacco, M. J., Van Rybroek, G. J., Rogstad, J. E., Yahr, L. E., Tomony, J. D., & Saewert, E. (2009). Predicting short-term institutional aggression in forensic patients: A multi-trait method for understanding subtypes of aggression. *Law and Human Behavior*, *33*, 308–319. <http://dx.doi.org/10.1007/s10979-008-9155-7>
- Walters, G. D. (1990). *The criminal lifestyle: Patterns of serious criminal conduct*. Newbury Park, CA: Sage. <http://dx.doi.org/10.4135/9781483325569>
- 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., & DeLisi, M. (2013). Antisocial cognition and crime continuity: Cognitive mediation of the past crime-future crime relationship. *Journal of Criminal Justice*, *41*, 135–140. <http://dx.doi.org/10.1016/j.jcrimjus.2012.12.004>
- Walters, G. D., & Heilbrun, K. (2010). Violence risk assessment and facet 4 of the Psychopathy Checklist: Predicting institutional and community aggression in two forensic samples. *Assessment*, *17*, 259–268. <http://dx.doi.org/10.1177/1073191109356685>
- 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>
- Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR 20: Assessing risk for violence (Version 2)*. Burnaby, Canada: Mental Health, Law and Policy Institute.
- Wong, S. C. P., Gordon, A., Gu, D., Lewis, K., & Olver, M. E. (2012). The effectiveness of violence reduction treatment for psychopathic offenders: Empirical evidence and a treatment model. *The International Journal of Forensic Mental Health*, *11*, 336–349. <http://dx.doi.org/10.1080/14999013.2012.746760>
- Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*, 377–392. <https://psycnet.apa.org/doi/10.1177/0734282911406668>. <http://dx.doi.org/10.1177/0734282911406668>

Received March 12, 2020

Revision received April 26, 2020

Accepted April 27, 2020 ■