Psychological Assessment
2001, Vol. 13, No. 4, 531-542

1040-3590/01/$5.00 DOI: 10.1037//1040-3590.13.4.531

Psychopathy and Ethnicity: Structural, Item, and Test Generalizability of the Psychopathy Checklist—Revised (PCL–R) in Caucasian and African American Participants

David J. Cooke
Glasgow Caledonian University and Douglas Inch Centre

David S. Kosson
Finch University of Health Sciences/
The Chicago Medical School

Christine Michie
Glasgow Caledonian University

The Psychopathy Checklist—Revised (PCL–R) is an important measure in both applied and research settings. Evidence for its validity is mostly derived from male Caucasian participants. PCL–R ratings of 359 Caucasian and 356 African American participants were compared using confirmatory factor analysis (CFA) and item response theory (IRT) analyses. Previous research has indicated that 13 items of the PCL–R can be described by a 3-factor hierarchical model. This model was replicated in this sample. No cross-group difference in factor structure could be found using CFA; the structure of psychopathy is the same in both groups. IRT methods indicated significant but small differences in the performance of 5 of the 20 PCL–R items. No significant differential test functioning was found, indicating that the item differences canceled each other out. It is concluded that the PCL–R can be used, in an unbiased way, with African American participants.

Hare's (1991) Psychopathy Checklist—Revised (PCL–R) is currently the instrument of choice for measuring psychopathy (Conoley & Impara, 1995). Its validity is supported by evidence from both experimental and applied domains (e.g., Cooke, Forth, & Hare, 1998; Hare, Cooke, & Hart, 1999). Within the applied domain, psychopathy is linked to poor treatment response (Loesel, 1998; Rice, Harris, & Cormier, 1992; Seto & Barbaree, 1999). Psychopathy is also recognized as an important predictor of criminal behavior in general and violent behavior in particular (Hart & Hare, 1997; Hemphill, Hare, & Wong, 1998; Salekin, Rogers, & Sewell, 1996). As a consequence, the PCL–R and the Screening Version (Hart, Cox, & Hare, 1995) are important components of many risk assessment procedures, including the HCR–20 (Webster, Douglas, Eaves, & Hart, 1997), the Violence Risk Assessment Guide (Quinsey, Harris, Rice, & Cormier, 1998), and the decision-tree procedure derived from the MacArthur violence study (Steadman et al., 2000). PCL–R results can influence important decisions about individual liberties, including granting of parole from prison, detention under dangerous offender legislation, access to treatment, and indeed, death sentence adjudications (Lyon & Ogloff, 2000). Thus, it is vital that the PCL–R is applied fairly in such contexts (American Educational Research Association & American Psychological Association, 1999).

Whereas the populations in which the PCL–R is used are culturally and ethnically diverse, the bulk of the evidence attesting to the validity of the PCL–R is based on Caucasian male offenders.[1] In fact, prior studies have raised questions about the construct validity of psychopathy in non-Caucasian samples. Although Kosson, Smith, and Newman (1990) reported similar correlations between psychopathy and criminal activity for Caucasian and African American offenders, the pattern of correlations between psychopathy and personality scores was different, and an exploratory factor analysis suggested differences in underlying factor structure. Moreover, some of the laboratory deficits observed in Caucasian offenders do not generalize to African American offenders (Doninger & Kosson, 2001; Newman & Schmitt, 1998;

David J. Cooke, Department of Psychology, Glasgow Caledonian University and Douglas Inch Centre, Glasgow, Scotland, United Kingdom; David S. Kosson, Department of Psychology, Finch University of Health Sciences/The Chicago Medical School; Christine Michie, Department of Psychology, Glasgow Caledonian University.

Christine Michie received support from U. K. Economic and Social Research Council Grant L133222704 while carrying out these analyses. This research was also supported in part by National Institute of Mental Health Grant MH57714 to David Kosson. We thank Allan Durndell and Stephen Hart for their comments on an earlier draft of the article.

Correspondence concerning this article should be addressed to David J. Cooke, Douglas Inch Centre, 2 Woodside Terrace, Glasgow, G3 7UY United Kingdom. Electronic mail may be sent to djcooke@rgardens.u-net.com.

[1] The term *ethnic* is currently used to refer to differences in culture, ancestry, and race, whereas the term *race* is typically used more narrowly to denote differences identified through genetic testing. For this reason, and because psychopathy research addressing cultural variation has relied on self-identification rather than biological classification, we use the terms *ethnic* and *ethnicity* rather than *racial* and *race* (see Okazaki & Sue, 1995, for a full discussion of the issue).

Newman, Schmitt, & Voss, 1997; Vitale & Newman, 1998). By contrast, some psychological deficits do generalize across ethnic groups (Kosson, 1998; Kosson, Suchy, Mayer, & Libby, 2001). Studies of adolescents have also detected no differences between Caucasian and African American samples in the correlates of psychopathy (Brandt, Kennedy, Patrick, & Curtin, 1997; Forth & Mailloux, 2000; Myers, Burket, & Harris, 1995). The nature and extent of ethnic bias of psychopathy measures remain to be evaluated with more detailed analysis.

One possible explanation for the differences in factor analysis and laboratory findings across ethnicity is that the psychopathy syndrome is different in African American than in Caucasian samples. Thus, African Americans and Caucasians with high PCL–R scores are characterized by some similar but some different underlying mechanisms. For example, it could be argued that African American psychopaths share with Caucasian psychopaths a propensity for violent and nonviolent criminal activity, criminal versatility, and cognitive deficits associated with left-hemisphere activation (Kosson, 1998; Kosson et al., 1990), but that African American psychopaths are not characterized by the same appraisal and response modulation deficits (Doninger & Kosson, 2001; Newman & Schmitt, 1998; Vitale & Newman, 1998).

Alternatively, psychopathy may be more difficult to measure in African American samples. In particular, it could be argued that the PCL–R is less effective at discriminating between African American offenders or that, the way the PCL–R is typically used by Caucasian examiners, PCL–R ratings are ethnically biased. Ethnic or racial biases found in any psychological instrument can have ethical, practical, and legal implications (Gottfredson, 1994; Okazaki & Sue, 1995). Given the importance of the decisions that are made using the PCL–R, it is imperative to assess whether the PCL–R is biased.

Cross-group differences in test performance can have substantive implications as well as measurement implications: They can lead to hypotheses regarding the development of a disorder and the form that it takes. For example, Cooke and Michie (Cooke & Michie, 1999; Cooke, Michie, & Clark, 2000) found that certain traits associated with psychopathy (e.g., glibness/superficial charm and grandiose sense of self-worth) were only apparent in extreme cases in United Kingdom samples. These findings are consistent with the perspective that the presentation of personality disorders is not immutable but may be influenced by sociocultural factors (Cooke, 1998; Draguns, 1986; Paris, 1998). Indeed, Hare (1998) suggested that such factors are important influences on the manifestations of psychopathy.

To the extent that ethnic and racial differences affect not only genetic variation but also socialization and acculturation, they may affect the expression of psychopathy. Exploring these differences can enhance our understanding of putative risk factors for the disorder. However, whether these differences affect the assessment of psychopathy using the PCL–R remains to be evaluated using contemporary analytic tools.

Assessing Bias

Bias can take several forms. Here we are concerned with construct bias, that is, whether the construct measured is identical across groups. A range of statistical techniques has been used to detect test and item bias or differential item functioning (see Van de Vijver & Leung, 1997, for a detailed review). Traditionally, cross-cultural studies have compared classical test theory (CTT) indices, including Cronbach's alpha and corrected item–total correlations across samples. Factor-analytic approaches have also been applied to assess the cross-cultural/cross-group equivalence of the latent structures underpinning scales (e.g., Barrett & Eysenck, 1984). Because these approaches can demonstrate structural identity but not scalar equivalence, invariance of factor structures is a necessary but not sufficient condition for ensuring cross-group equivalence (Bijnen & Poortinga, 1988; Tanzer, 1995). An example from the physical sciences may illustrate the point: The Fahrenheit and Centigrade scales both measure the same underlying construct—temperature—but they are not metrically equivalent because they do not have the same zero points and their scale intervals are different. To ensure equivalence of measurement across groups, it is necessary to demonstrate not only that the same construct is being measured but also that the metric on which it is measured is the same (Van de Vijver & Leung, 1997). In the absence of metric equivalence, it is meaningless to compare, for example, prevalence estimates based on a cutoff.

Item response theory (IRT) provides a powerful set of modeling techniques for analyzing items and tests (Embretson, 1996; Santor & Ramsay, 1999). These are more appropriate than factor-analytic or CTT approaches for examining cross-group equivalence. One particular advantage of IRT approaches is that they can provide measurement on an identical scale across groups. Group differences in test scores may occur because of measurement bias, group differences, or a combination of these two factors. IRT methods provide a formal psychometric model that allows the groups to be matched on the underlying latent trait. By focusing on latent variables rather than manifest variables, one can distinguish between measurement bias and true group differences. Indeed, Meredith and Millsap (1992) argued that it is not possible to fully assess bias using manifest variables and that is it essential to model manifest variables using latent variables.

In brief, IRT models specify the relationship between item or test scores and the underlying latent trait (θ) that is postulated to underpin item or test scores. Graphical methods can be used to map the probability of an item or test score against θ. With regard to item characteristic curves (ICCs), two characteristics are relevant: their slope and the position of their maximum point of inflection. Items with larger slopes are more discriminating; they have higher saturation of trait-relevant variance. The position of the maximum point of inflection (as measured by the threshold parameter) reflects the extremity or difficulty of the item. For example, in previous studies it has been demonstrated that the PCL–R item *irresponsibility* is usually rated positive at average levels of θ, whereas the PCL–R item *glibness/superficial charm* is generally rated positive only when higher levels of the underlying trait are present (Cooke & Michie, 1997, 1999). If African American and Caucasian samples differ only in uniform differential item functioning (DIF) for some items, this would indicate that all of the same PCL–R items can be used with the two groups but that some items will be particularly useful for individuals in these groups at different overall levels of psychopathy. In contrast, evidence for nonuniform DIF for an item would indicate that this item is more closely related to the underlying construct of psychopathy in one group than in the other.

IRT methods are the methods of choice for detecting DIF or for detecting differential test functioning (DTF) across groups. DIF occurs when an item is more discriminating, more difficult, or more extreme in one group as compared with another. Careful consideration of ICCs can assist in identifying ethnic, gender, or other biases in an item or in a test as a whole. Several features make IRT methods more suitable than CTT methods for examining test biases. First, CTT indices such as Cronbach's alpha and corrected item–total correlations are highly sensitive to variations in the range of test scores across samples (Van de Vijver & Poortinga, 1994). By way of contrast, ICCs are independent of the samples from which they are derived (Mellenbergh, 1996). Second, it is not necessary to obtain representative samples to obtain unbiased estimates of item and test characteristics; nonrepresentative samples may be used (Embretson, 1996).

Third, direct comparison of the performance of items can be made across groups. For example, it is possible to distinguish between item differences in extremity (i.e., differences in the level of the underlying trait at which the inflection point occurs) and differences that reflect differences in the relevance of items across groups (i.e., differences in slopes). Differences in extremity demonstrate *uniform DIF*, indicating that the level of the underlying trait for which the item is useful differs across groups, but the item is still useful for both. Instances of uniform DIF indicate measurement bias but not necessarily true group differences in the composition of the underlying trait. By contrast, differences in slopes demonstrate *nonuniform DIF*, indicating that the discriminating power of the item differs from one sample to another (Holland & Wainer, 1993) and that groups differ in the importance of specific components of a construct. Differences in item performance can provide important information about cross-group differences. Whereas unbiased items define common aspects of the construct, biased items denote cross-group idiosyncrasies (Bontempo, 1993; Holland & Wainer, 1993; Reise, Widaman, & Pugh, 1993; Van de Vijver & Poortinga, 1994).

A fourth advantage of IRT over CTT is that the scale of θ is defined by the items, and when groups are compared, it is possible to ensure that a common metric is used for comparisons. This property of measurement invariance across groups can ensure, for example, that diagnostic cutoffs are equivalent (e.g., Cooke & Michie, 1999; Reise et al., 1993). Constraining items to have identical parameters across groups, or so-called *anchoring*, ensures that responses are underpinned by a latent scale with a common metric. For the anchor items, the same set of parameters is assumed to apply to both groups. This ensures that trait levels and item parameters for the nonanchor items are estimated on the same scale and thus are directly comparable.

Waller, Thompson, and Wenk (2000) emphasized that research on bias should focus on latent rather than observed variables for three reasons. First, differences may be detected in manifest variables when no differences occur on the latent variable. Second, the opposite may also be true: Differences on the latent variable may be masked by a lack of differences on observed variables. Third, although biases may be present at the item level, aggregation across items may result in nonbiased estimates of the underlying trait at the test score level. Cross-group DIF for multiple items may result in *amplification* or *cancelation* of DTF (Raju, Van der Linden, & Fleer, 1995). For example, positive ratings may be obtained by the minority group (as opposed to majority group) at lower levels of θ for some items and at higher levels of θ for other items. Nonetheless, using an explicit IRT model, it is possible to obtain unbiased estimates of the underlying latent trait from scales that contain some biased items.

Previous comparisons of the functioning of the PCL–R items in North American and Scottish samples have demonstrated that the metric of the latent trait underlying PCL–R scores in the two cultures differed (Cooke & Michie, 1999; Cooke et al., 2000). Moreover, IRT analyses have shown that, whereas several of the items demonstrate DIF, DIF was uniform in that the items differed only in the threshold parameter, that is, the level of psychopathy at which the items discriminate among different individuals. However, the items were no more discriminating of the latent trait in one country compared with the other.

To date, there are no published reports of IRT analyses of the properties of the PCL–R items for members of different ethnic groups within a country. Therefore, the present study was designed to provide an evaluation of whether PCL–R items exhibit DIF for African American versus Caucasian inmates within the United States. We examined whether the 20-item PCL–R and the 13-item version of the PCL–R recommended by Cooke and Michie (2001) performed similarly both at the level of the individual item and at the level of the test as a whole. In addition, we conducted an evaluation of the importance of any differences identified using indices previously validated for this purpose.

Method

Participants

The participants were African American and Caucasian adult male inmates drawn from one of two different correctional institutions in the United States. Two hundred and one participants were selected from a federal prison in the southeastern United States. Another 514 inmates were selected from a county jail within a 50-mile radius of a midwestern urban center. In total there were 359 Caucasian and 356 African American participants. Because the correctional institutions were designed to respond to different criminal justice issues, the two groups of inmates had been convicted of different kinds of instant offenses. Whereas inmates in the federal prison were serving time for federal felony convictions, inmates in the county jail had been convicted of either misdemeanors or felonies in the state of Illinois and were serving sentences of no more than 1 year.

Participants from the two sites were relatively similar in demographic characteristics. Nevertheless, as shown in Table 1, federal inmates were significantly older, higher in intelligence, and lower in anxiety or negative affectivity and degree of right-handedness. Federal inmates were also slightly but significantly lower in PCL–R total scores. Because of missing data for some PCL–R items, IRT analyses were based on 514 county jail inmates and 201 federal inmates. The only significant difference across ethnic group was on the Wechsler Adult Intelligence Scale—Revised scores.

Materials

The PCL–R consists of 20 items (see Table 2). On the basis of interview and file review, a trained rater determines how closely each individual participant meets the characteristics specified in the item descriptors. Each item is scored on a 3-point scale (0 = *absent*, 1 = *maybe/in some respects*, or 2 = *present*) indicating the degree to which the item applies to the individual. The items include the behavioral, affective, and interpersonal items considered to characterize psychopathic personality disorder (Cleckley, 1976; Hare, 1991).

Table 1
Sample Characteristics for Caucasian (C) and African American (AA) Participants

Characteristic	Sample: Federal prisoners (n = 201)	Sample: County jail prisoners (N = 514)	p
Proportion African American (%)	42	52	*ns*
Age (years)	30.8	26.3	<.001
Estimated WAIS–R IQ	92.8	89.0	<.01
Welsh anxiety	16.8	15.6	*ns*
Handedness	9.8	10.9	<.01
PCL–R score	23.4	25.0	<.01
Original Factor 1 score	10.0	9.9	*ns*
Original Factor 2 score	10.5	12.0	<.001
Arrogant and deceitful interpersonal style	4.9	4.5	*ns*
Deficient affective experience	5.2	5.4	*ns*
Impulsive and irresponsible behavioral style	6.7	6.7	*ns*

	Federal prisoners			County jail prisoners			Both		
	C	AA	p	C	AA	p	C	AA	p
Age (years)	32.0	29.4	<.01	25.8	26.2	*ns*	28.5	27.4	*ns*
Estimated WAIS–R IQ	98.2	85.3	<.001	93.9	85.0	<.001	95.6	85.1	<.001
Welsh anxiety	15.8	18.2	*ns*	15.2	16.0	*ns*	15.4	16.7	*ns*
Handedness	10.2	9.1	*ns*	11.0	10.8	*ns*	10.8	10.4	*ns*
PCL–R score	23.5	23.3	*ns*	24.1	25.7	<.01	24.0	25.0	*ns*
Original Factor 1 score	10.2	9.9	*ns*	9.4	10.3	<.01	9.6	10.2	*ns*
Original Factor 2 score	10.6	10.5	*ns*	11.9	12.0	*ns*	11.6	11.5	*ns*
Arrogant and deceitful interpersonal style	4.9	4.8	*ns*	4.2	4.8	<.01	4.4	4.8	*ns*
Deficient affective experience	5.2	5.1	*ns*	5.2	5.6	<.05	5.2	5.4	*ns*
Impulsive and irresponsible behavioral style	6.7	6.8	*ns*	6.6	6.7	*ns*	6.7	6.7	*ns*

Note. Original Factor 1 score = sum of scores for glibness/superficial charm, grandiose sense of self-worth, pathological lying, conning/manipulative, lack of remorse or guilt, shallow affect, callous lack of empathy, failure to accept responsibility for own actions. Original Factor 2 score = sum of scores for need for stimulation, parasitic lifestyle, poor behavioral controls, early behavior problems, lack of realistic long-term goals, impulsivity, irresponsibility, juvenile delinquency, revocation of conditional release. Arrogant and deceitful interpersonal style = sum of scores for glibness/superficial charm, grandiose sense of self-worth, pathological lying, conning/manipulative. Deficient affective experience = sum of scores for lack of remorse or guilt, shallow affect, callous lack of empathy, failure to accept responsibility for own actions. Impulsive and irresponsible behavioral style = sum of scores for need for stimulation/proneness to boredom, impulsivity, irresponsibility, parasitic lifestyle, lack of realistic long-term goals. WAIS–R = Wechsler Adult Intelligence Scale—Revised; PCL–R = Psychopathy Checklist—Revised.

Interview procedures were relatively similar for the two samples, lasting approximately 60 to 90 min and covering childhood and educational histories, sexual relationships, employment, and criminal histories. For the federal prison sample, files were extensive, and raters focused on the sections of the file dealing with criminal histories, presentence investigations, psychological evaluations, and institution discipline reports. For the county jail sample, files were generally brief, including information about infractions within the jail, gang affiliations, and basic demographic information. In most cases, criminal histories and performance while on pretrial supervision were also available.

Eleven Caucasian raters and 1 African American rater classified the federal inmate participants. The average interrater agreement was .84 for the Caucasian (n = 34) and .84 for the African American (n = 44) participants. Twelve Caucasian and 3 Latina raters classified the county jail inmate participants. The average interrater agreement was .79 for the Caucasian (n = 68) and .78 for the African American (n = 80) participants.

Cooke and Michie (2001) demonstrated that PCL–R ratings are structurally complex and can be represented by a hierarchical model in which 13 PCL–R items form a higher order factor underpinned by three distinct but correlated factors: Arrogant and Deceitful Interpersonal Style, Deficient Affective Experience, and Impulsive and Irresponsible Behavioral Style (Cooke & Michie, 2001). The higher order factor has been shown to be clearly unidimensional in other samples. Because most researchers and clinicians use all 20 items, we examined the IRT parameters for both the 13-item model and for all 20 items.

Overview of the IRT Model

Samejima's graded model is an appropriate model for rating data of this type as the assumptions of the model match the structure of PCL–R data (e.g., Cooke & Michie, 1997; Cooke, Michie, Hart, & Hare, 1998). The probability of earning each possible score on an item varies in relation to

Table 2
Model Parameters for PCL–R Items for Caucasian (C) and African American (AA) Participants

	13-item test					20-item test				
		b_1		b_2			b_1		b_2	
Item	*a*	C	AA	C	AA	*a*	C	AA	C	AA
1. Glibness/Superficial charm	**1.3**	**−1.3**	**−1.3**	**0.6**	**0.6**	**1.1**	**−1.4**	**−1.4**	**0.6**	**0.6**
2. Grandiose sense of self-worth	**1.5**	**−1.5**	**−1.5**	**0.3**	**0.3**	**1.3**	**−1.6**	**−1.6**	**0.4**	**0.4**
3. Need for stimulation	1.1	−3.0	−2.2	−0.8	0.3	1.2	−2.9	−2.1	−0.8	0.3
4. Pathological lying	**1.2**	**−1.7**	**−1.7**	**0.6**	**0.6**	**1.2**	**−1.7**	**−1.7**	**0.6**	**0.6**
5. Conning/manipulative	**1.2**	**−1.4**	**−1.4**	**0.6**	**0.6**	**1.2**	**−1.3**	**−1.3**	**0.6**	**0.6**
6. Lack of remorse or guilt	**1.6**	**−2.3**	**−2.3**	**−0.3**	**−0.3**	**1.6**	**−2.4**	**−2.4**	**−0.3**	**−0.3**
7. Shallow affect	**1.5**	**−1.4**	**−1.4**	**0.6**	**0.6**	**1.3**	**−1.5**	**−1.5**	**0.6**	**0.6**
8. Callous/lack of empathy	**2.1**	**−1.4**	**−1.4**	**0.3**	**0.3**	**2.2**	**−1.4**	**−1.4**	**0.3**	**0.3**
9. Parasitic lifestyle	1.0	−1.0	−1.6	1.7	1.0	1.1	−1.0	−1.6	1.6	1.0
10. Poor behavioral controls						**1.1**	**−1.8**	**−1.8**	**0.0**	**0.0**
11. Promiscuous sexual behavior						**0.9**	**−0.8**	**−0.8**	**0.7**	**0.7**
12. Early behavioral problems						**0.7**	**−0.8**	**−0.8**	**1.1**	**1.1**
13. Lack of long-term goals	1.1	−1.9	−2.3	0.2	0.0	1.1	−1.9	−2.4	0.2	0.0
14. Impulsivity	1.1	−2.9	−2.5	−0.5	0.2	1.2	−2.7	−2.3	−0.5	0.2
15. Irresponsibility	0.9	−3.7	−4.5	−1.0	−0.3	1.0	−3.4	−4.2	−0.9	−0.3
16. Failure to accept responsibility	**1.0**	**−2.9**	**−2.9**	**−0.5**	**−0.5**	**0.9**	**−3.0**	**−3.0**	**−0.5**	**−0.5**
17. Many short-term marriages						**0.5**	**0.0**	**0.0**	**1.8**	**1.8**
18. Juvenile delinquency						**0.6**	**−1.2**	**−1.2**	**1.3**	**1.3**
19. Revocation of release						**0.6**	**−3.7**	**−3.7**	**−2.4**	**−2.4**
20. Criminal versatility						**0.8**	**−1.8**	**−1.8**	**0.4**	**0.4**

Note. Items equated across samples are presented in bold. PCL–R = Psychopathy Checklist—Revised; *a* = slope parameter; b_1 and b_2 = difficulty parameters.

degree of the latent trait; this can be illustrated by a curve or trace line. The curves for "0" and "2" responses are symmetric logistic functions. In short, as the degree of psychopathy increases, the probability of earning a 2 on any given item increases, and the probability of earning a 0 on this item decreases at an equivalent rate. The curve for the "1" response is found by subtraction: The total probability of all three responses at any level of the trait must be unity. The shape and position of the curves in relation to the trait can be summarized by the values of three parameters, *a*, b_1, and b_2 (Thissen, 1991). These curves are illustrated in Figure 1.

The point on each curve at which the probability of earning a particular score (0 or 2) is .50 is called the *point of inflection*; for this type of curve this is the maximum value of the slope. The slopes at the point of inflection for the probability of being given a score of 0 or a score of 2 on an item—$P(0)$ and $P(2)$ respectively—are of the same magnitude but opposite in direction; this is given by the parameter *a*. The *a* parameter measures the discriminating power of the item (Holland & Wainer, 1993). The larger the value of *a*, the steeper the slope; items with large *a* parameters are comparatively highly saturated in trait-relevant variance (Waller et al., 2000).

The positions of the points of inflection are given by the parameters b_1 for $P(0)$ and b_2 for $P(2)$. The b_i parameters (b_1 and b_2) therefore provide measures of item difficulty or extremity or frequency of a behavior, attitude, or trait. Increases in the value of b_i move the curve to the right, increasing the level of extremity of the trait at which the item discriminates between individuals low versus high in the latent trait.

MULTILOG VI (Thissen, 1991) was used for all IRT analyses. The program uses maximum likelihood methods to estimate item parameters simultaneously in two or more groups. The program allows a variety of constraints to be imposed on the parameters, and generalized likelihood ratio testing (GLRT) can be performed. The equivalence of parameters across groups can be determined by comparing the goodness of fit of a constrained model with the goodness of fit of an unconstrained model. Thus, two models are compared, one in which parameters are constrained to be equivalent across the groups and one in which no such constraints are imposed. If GLRT reveals no significant difference between the models, this confirms that there is no evidence of any differences in the item parameters across the two groups.

Analyses of DTF

We also examined the properties of the PCL–R test as a whole in two ways. Test information functions provide an estimate of the precision of measurement at different points on the underlying latent trait. Information is the IRT equivalent of reliability but has the advantage of being an estimate across the trait rather than being an estimate at the average score. Information measures the accuracy of the trait estimate for a given individual and is asymptotically 1 over the square of the standard error (Nunnally & Bernstein, 1994). Information can be calculated from the item parameter estimates produced by Samejima's graded model.

Because DIF does not always have an impact at the level of the test as a whole, we also examined DTF both graphically and numerically. Test characteristic curves (TCCs) are the test score equivalents of ICCs in which test scores are plotted as a nonlinear function of θ (Lord, 1980). The slope of a TCC describes the extent to which a change in the test score varies with the level of θ. Visual inspection of TCCs can be used to evaluate whether there is differential test functioning for two groups. The impact of metric inequivalence on test scores can be assessed graphically.

To provide a numerical index of DTF, Raju et al. (1995) introduced the root differential test function (rDTF). In essence, this index compares the PCL–R test scores generated for the Caucasian participants estimated using the model parameters for the African American participants with the test scores estimated using the model parameters for the Caucasian participants. The rDTF index expresses the difference in TCCs in the metric of the test.

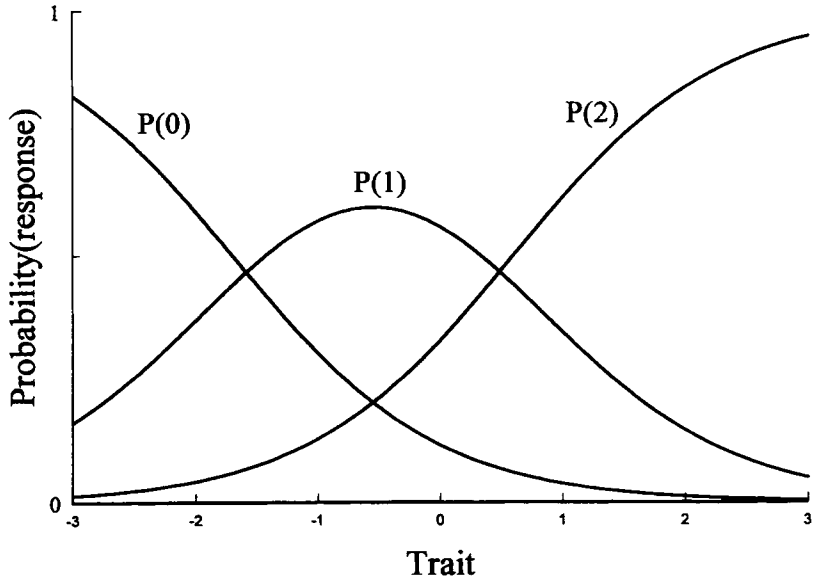

Figure 1. Example of item characteristic curves. P = probability.

Results

Before proceeding to the IRT analyses, we evaluated the structural properties of the PCL–R ratings. We did this for three reasons. First, structural equivalence is a necessary but not a sufficient condition of equivalence across groups. Second, previous research using exploratory factor analysis has been interpreted as indicating differences in the factor structure underlying ratings of Caucasian and African American participants (Kosson, Smith, & Newman, 1990). Third, it was important to ensure that the ratings were sufficiently unidimensional to allow the appropriate application of the chosen IRT procedures. Having carried out confirmatory factor analysis (CFA), we conducted a detailed IRT analysis of the data.

Assumption of Unidimensionality

Although Cooke and Michie (2001) demonstrated that PCL–R ratings are structurally complex and can be represented as a superordinate psychopathy factor underpinned by three distinct but correlated factors, the higher order factor has been shown to be clearly unidimensional in other samples. We began by examining the similarity of the factor structure underlying PCL–R ratings in the two ethic groups using CFA. The quality of fit was determined using multiple measures of fit, because all measures have limitations and there are no agreed methods for absolutely determining goodness of fit (Kline, 1998; MacCallum & Austin, 2000). To provide a broad estimate of the quality of fit, we used indices that provided information about different aspects of fit (i.e., absolute fit, fit adjusted for model parsimony, and fit relative to a null model). The overall fit of the model was first assessed using the chi-square test. It is well recognized that this test is sensitive to sample size and will generally be significant in samples even of moderate size. As a consequence, this statistic is not generally interpreted. A range of other indices was used. These indices have been reported to yield estimates that are relatively independent of sample size; each index has its own merits (Kline, 1998). The nonnormed fit index (NNFI) compensates for model complexity, the robust comparative fit index (CFI) is sensitive to model misspecification, and the root mean square error of approximation (RMSEA) is least affected by the estimation method used. Following convention, adequate fit was determined by values of the first two coefficients greater than .90 and RMSEA less than .08 (Byrne, 1994). Using a combination of indices provides a more conservative and reliable evaluation of fit.

The 13-item model developed by Cooke and Michie (2001) was fitted to the data from the African American and Caucasian samples independently. Satisfactory fit was achieved for both samples (Byrne, 1994): for Caucasians, $\chi^2(56, N = 311) = 140.4$, $p < .001$, NNFI = .88, CFI = .92, RMSEA = .070; for African Americans, χ^2 (56, $N = 312$) = 146.7, $p < .001$, NNFI = .88, CFI = .92, RMSEA = .072.

Next, we applied a demanding test of cross-group invariance by fitting the 13-item hierarchical factor model to the data from the two samples simultaneously. We used CFA to estimate a baseline model in which all the parameters were allowed to take different values for the two samples. Comparison of this model with a model in which all model parameters (i.e., loadings, variances, disturbance, and errors) were constrained to be equal for the two samples resulted in a nonsignificant change in chi-square: $\Delta\chi^2$ (35) = 27.8, *ns*. This suggested that the model fitted identically in both the African American and the Caucasian samples.

Zinbarg, Barlow, and Brown (1997) indicated that the unidimensionality or coherence of a superordinate measure in a hierarchical model can be determined by estimating the total test variance accounted for by the superordinate factor: The ratio of this value to the observed variance in total scores provides an estimate of general factor saturation (GFS; Zinbarg et al., 1997). Values of GFS over .50 are consistent with a coherent measure, because more than half of the variance of the total scores is accounted for by a single construct. Coherence was achieved in both the African

American and the Caucasian samples (GFS = .74 for both samples).

These analyses demonstrate sufficient GFS to warrant application of IRT methods that assume unidimensionality. The CFA also provides evidence indicating cross-group invariance of factor structures. However, as noted above, such evidence is insufficient to guarantee the cross-group equivalence of the PCL–R (Bijnen & Poortinga 1988; Huang, Church, & Katigbak, 1997; Tanzer, 1995). For this purpose, further analyses are required.

IRT Comparison of African American and Caucasian Samples Using the 13-Item Hierarchical Model

Initial analyses were based on the 13 PCL–R items that fit the hierarchical model. An unconstrained baseline model was generated in which the mean level of the underlying trait and all item parameters were allowed to vary across the two groups. As shown in Table 2, a series of models was estimated and compared with the baseline model. The improvement in the goodness of fit of the constrained model, relative to the unconstrained model, was evaluated using GLRT theory. Under GLRT, the statistic $G^2 = -2(\log \text{likelihood})$ is calculated separately for the unconstrained and constrained models, and the difference between these values, ΔG^2, is distributed as chi-square with degrees of freedom equal to the number of constraints imposed. One advantage of IRT methods is that it is not essential that all items have equivalent parameters to ensure that a common metric underpins scores in different groups. Items that are invariant across groups can act as "anchors" to establish a common metric for θ across groups. (For an account of this method, see Cooke & Michie, 1999; Reise et al., 1993.)

We started by constraining the slope parameters to be equal; equality of slopes across groups indicates that items have similar relevance for defining the underlying trait in each group. As noted above, cross-group differences in slope are more serious in relation to bias than differences in threshold. Imposing the constraints of equal slopes across groups for the 13 items and equal thresholds for 8 items led to a nonsignificant increase in G^2, $\Delta G^2 = 34.9$, $df = 29$, *ns*. The slopes of the different items varied, but there was no variation within items across groups. Thus, the 13 PCL–R items are equally discriminating for African American and Caucasian participants. However, equating all thresholds to be equal did result in a significant increase in G^2, $\Delta G^2 = 144.2$, $df = 39$, $p < .001$. Total difference in b_i parameters was calculated by summing the absolute values of differences in b_1 and differences in b_2 for each item across the two groups after the slopes had been constrained to be equal. Items were added in order of total absolute difference (smallest difference first), one item at a time, until a significant increase in G^2 was found. Parameter values are displayed in Table 2. These results suggest the presence of some *uniform DIF*. Items equated across samples are presented in bold text. The model fits the data well, predicting the observed pattern of responses for each item within 1%.

Given metric equivalence, it is legitimate to compare the overall means on the latent trait, and these were significantly different with the African American prisoners having, on average, higher scores (Caucasian = −0.22, African American = 0.00, $\Delta G^2 = 6.7$, $df = 1$, $p < .01$).

In line with previous North American results, the PCL–R items *callous/lack of empathy*, *lack of guilt and remorse*, and *grandiose sense of self-worth* were the most discriminating items (i.e., largest *a* parameters; Cooke & Michie, 1997). It is also noteworthy that there are no differences in the b_i parameter (i.e., thresholds) for the items that load on the first two factors: Arrogant and Deceitful Interpersonal Style and Deficient Affective Experience. All the differences emerge in the items that load on the Impulsive and Irresponsible Behavioral Style factor. No consistent pattern emerged; some thresholds were lower in African Americans and some were lower in Caucasians. Ratings denoting presence of the disposition occur at lower levels of the underlying trait for African Americans on *parasitic lifestyle* and *lack of long-term goals* and at higher levels on *need for stimulation*, *impulsivity*, and *irresponsibility*.

In summary, these findings indicate that the 13 item PCL–R does not differ significantly in its ability to discriminate between levels of the underlying trait across African American and Caucasian participants. However, certain items related to the Impulsive and Irresponsible Behavioral Style factor become apparent at different levels of psychopathy in African American compared with Caucasian participants.

Importance of DIF

In large samples, statistically significant DIF may occur even when the effect is of little practical significance (see Kirk, 1996). Unfortunately, the estimation methods used in MULTILOG VI do not yield variance estimates for the IRT parameters, so effect sizes cannot be calculated simply. Zumbo (1999) described an alternative procedure, based on ordinal logistic regression, from which effect sizes for DIF can be obtained. In this procedure the item score is the dependent variable; the total test score is forced into the regression equation on the first step, the ethnicity variable on the second step, and their interaction on the third step. Both uniform and nonuniform DIF can be detected. Zumbo (1999) indicated that for an item to be classified as displaying DIF, it should have $p < .01$ and the Zumbo–Thomas effect size for the full model should be >.13. Examination of the regression equation allows the determination of whether uniform or nonuniform DIF is present.

Zumbo's procedure cannot be applied to cases with missing values. The procedure was carried out, therefore, with smaller samples (Caucasian $n = 216$; African American $n = 222$). Examination of the results[2] of this procedure indicated that, although 4 items showed differences that were significant at the .01 level, the largest Zumbo–Thomas effect size was .054, well below the .13 level regarded as indicating an important effect size. This method indicated that none of the 13 items should be regarded as showing DIF that should give cause for concern.

IRT Comparison of African American and Caucasian Samples Using All 20 PCL–R Items

A comparable analysis was carried out with all 20 PCL–R items. Results are presented in Table 2. Imposing the constraint of equal slopes for all 20 items and equal thresholds for 15 items led to a nonsignificant increase in G^2, $\Delta G^2 = 56.3$, $df = 50$, *ns*. The

[2] Detailed tables of the results can be obtained from David J. Cooke.

same 5 items that could not be constrained to be equal in the 13-item solution could not be constrained to be equal in this analysis. Equating all thresholds resulted in significant increases in G^2, $\Delta G^2 = 170.7$, $df = 60$, $p < .001$. Parameter values are displayed in Table 3. The model fits the data well, predicting the observed pattern of responses for each item within 1%. It is noteworthy that the parameter values for the 13 items were essentially similar whether they were estimated for 13 items alone or all 20 items. This implies that the 20 item PCL–R is sufficiently unidimensional for the appropriate application of the graded model. Unfortunately, the method of Zinbarg et al. (1997) cannot be used to directly test for unidimensionality in this case because all 20 items do not fit a hierarchical model adequately.

Because the 15 anchoring items demonstrated metric equivalence for the two groups, it is legitimate to compare overall means on the latent trait, and these were significantly different with African American prisoners having, on average, higher scores than Caucasians (Caucasian = −0.22, African American = 0.00, $\Delta G^2 = 6.7$, $df = 1$, $p < .01$).

Information Functions

As discussed earlier, test information functions provide an estimate of the precision of measurement at different levels of the underlying trait. Examination of the test information functions for the 13 and 20 item versions (Figure 2) indicated that, at low to moderate levels of the trait, the information levels are high; however, as the trait level approaches the diagnostic cutoff (around $\theta = 1$), there is a sharp decrease in information. Consequently, the precision of measurement is high around average values of the score, then declines sharply around a PCL–R score of 28. The pattern was identical in both ethnic groups.

Differential Test Functioning

Although DIF was detected, this may or may not have an impact at the test level. Aggregation across items with cross-group DIF may result in amplification or cancelation of DTF. Thus summing across items may result in biased or nonbiased estimates of the underlying trait at the test score level. Given that the present analysis demonstrated uniform DIF in which ratings of African American participants denoting the presence of the disposition occurred at lower levels of the underlying trait for two items and at higher levels for three items, cancelation might be expected.

To examine DTF graphically, we plotted PCL–R scores as a nonlinear function of θ. In such TCCs (Lord, 1980), the slope illustrates the extent to which a change in the test score varies with the level of θ. A comparison of TCCs for the 13-item version of the PCL–R indicated that there are no discernible differences in the relations between PCL–R total score and level of the underlying trait across ethnic group (Figure 3). The same pattern was evident for the 20-item version.

To provide a numerical index of DTF, we also calculated the rDTF (Raju et al., 1995) to compare the PCL–R test scores generated for the Caucasian participants based on the model parameters for the African American participants with the test scores estimated using the model parameters for the Caucasian participants. The rDTF index expresses the differences in TCC in the metric of the test. Because the 20-item test is used in clinical and research practice, we estimated the index for the full test. The rDTF = 0.3, *ns* (i.e., < 1% of the maximum score), indicating that, on average, the difference in the PCL–R total score would be 0.5 points, or 1.2% of the maximum PCL–R score, when estimated using the model parameters for the other group.

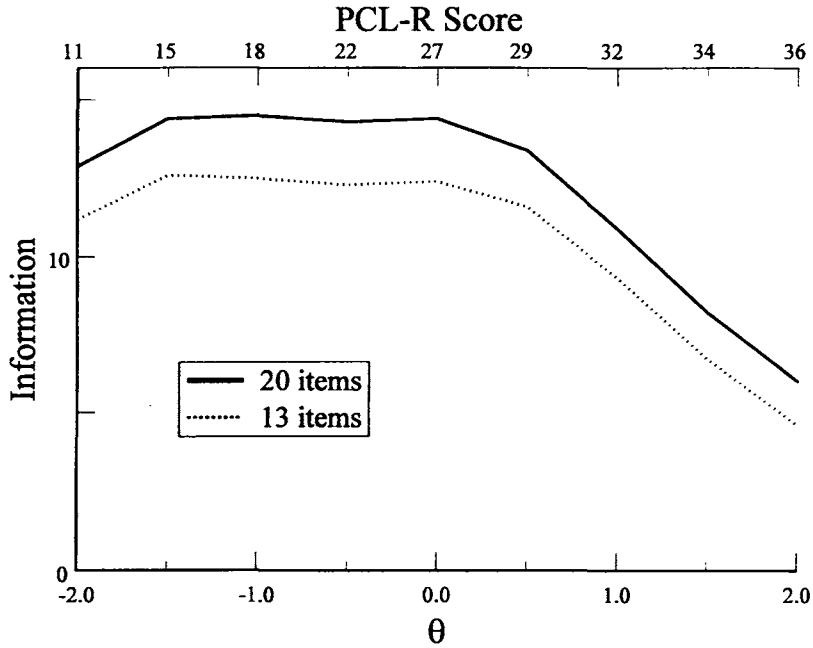

Figure 2. Test information function for 13- and 20-item versions of the Psychopathy Checklist—Revised (PCL–R).

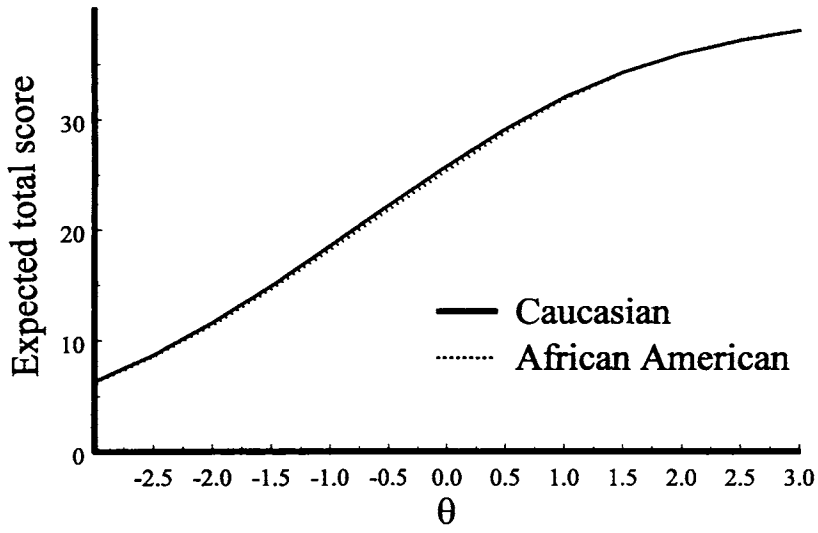

Figure 3. Test characteristics curves for 20-item Psychopathy Checklist—Revised for Caucasian and African American participants.

Discussion

Overall our analyses indicate little difference of any substance in the performance of the PCL–R in Caucasian and African American participants. In some sense, this result is reassuring as it indicates that the PCL–R is not inherently biased, although these findings do not preclude biased application. Both the similar underlying factor structure and the evidence for similar item functioning for most, but not all, items add to the empirical basis for using the PCL–R in African American male inmates. Test level analyses also point to similar functioning for the PCL–R in African American and Caucasian inmates. In short, the 13 (or 20) items of the PCL–R appear to be similarly useful in diagnosing psychopaths of both ethnic groups.

Factor Structure

To ensure generalizability across groups, comparability of factor structure is a necessary but not sufficient condition. The findings of this study are in sharp contrast to those of the only other published study that has examined the comparability of factor structures across racial or ethnic groups (Kosson et al., 1990). Kooson et al. (1990) concluded that the factor structures are not comparable for Caucasian and African American participants and suggested that findings "raise the possibility that the personality dynamics of Black and White psychopaths may be somewhat different" (p. 258). Kosson et al. based their conclusions in part on exploratory factor-analytic techniques and a comparison of congruence coefficients, techniques that are now viewed as less appropriate for this task than CFA (Cooke & Michie, 2001; Van de Vijver & Leung, 1997). In the current CFA, no cross-group differences are detected when the three-factor hierarchical model is fitted; that is, no cross-group differences are detected in any of the model parameters, including the loadings, variances, disturbances, and errors. Thus, the more powerful and more appropriate methods of CFA indicate that the results of Kosson et al. (1990) do not generalize to these samples.

The factor analysis also provided evidence that the three-factor model developed by Cooke and Michie (2001) on North American and European samples can be cross-validated on these two independent samples. Cross-validation of a structural model, using samples independent of those used to develop the model, enhances the plausibility of the model, and cross-validation across groups increases the generalizability of the model (Byrne, 1994; Van de Vijver & Leung 1997).

Effects at the Item Level

Overall, current analyses suggest relatively few significant differences in the item functioning of the PCL–R as a function of ethnicity; the effect sizes for the five significant differences are small. Regardless of which set of items is considered, there are no differences in the discriminating power of the items. There are also no differences in the extremity of any of the items loading on the Arrogant and Deceitful Interpersonal Style and Deficient Affective Experience factors, those dimensions most closely linked to the personality core thought to underlie psychopathy.

Only items that load on the Irresponsible Behavioral Style factor demonstrate statistically significant DIF, and the magnitude of DIF is small. Although it is important to replicate these differences before we can be certain the specific cases of DIF are meaningful, they suggest that it is with respect to this dimension of psychopathy that researchers and clinicians should be most careful in speculating about the traits of African American individuals with high PCL–R scores. Moreover, because Irresponsible Behavioral Style (often referred to as Factor 2 in the two-dimensional view of psychopathy) is the dimension of psychopathy most closely associated with Antisocial Personality Disorder, caution may also be warranted in making conclusions about the specific traits of African Americans with this diagnosis.

Effects at the Test Level

Because African American participants obtained positive ratings at lower levels of the latent trait than Caucasian participants for

some items and at higher levels for others, the DIF canceled, resulting in little effect on total test scores. Although African American participants earned slightly higher PCL–R scores overall, there was little discernible difference between groups in the TCCs. Moreover, the nonsignificant rDTF value corroborates the absence of a difference. On average, the difference in PCL–R scores is 0.5 when estimates derived using one set of model parameters are compared with estimates derived from the other set of parameters.

Examination of the TCCs and the test information functions also indicates that, at high levels of the construct, individuals with very different levels of the construct have very similar scores. We are not aware of this finding being reported before. This indicates that the PCL–R has poor discrimination, and therefore is relatively poor, at measuring the diversity of psychopathic traits above the diagnostic cutoff. This was true for both ethnic groups. An advantage of IRT methods, with an explicit measurement model, is that they permit examination of the impact of either modifying current items, or writing new items, on the assessment of the disorder (Steinberg & Thissen, 1996). These results indicate that to measure the diversity within the higher range of the latent trait, it may be necessary to either develop new items for this specific purpose or refine and extend the range of possible scores within existing items. Clinical experience suggests that those who score high on the PCL–R are heterogeneous in terms of presentation and clinical need; current findings suggest that the PCL–R is not good at characterizing this heterogeneity. An alternative strategy for making distinctions among subsets of high PCL–R scorers is to include additional measures of psychopathy-relevant traits (Vassileva, Kosson, & Conrod, 2001).

It is noteworthy that there is relatively little difference in the information provided by the 13-item and 20-item versions, suggesting that little measurement precision is lost by reducing test length by a third. This finding has both practical and theoretical implications. From a practical perspective, the PCL–R is a time-consuming procedure to carry out; little improvement in precision is gained by assessing the additional 7 items. Similar content in some items means that test information is overestimated, and thus the difference in information provided by the two versions is actually less than it appears (Cooke & Michie, 2001).

Although our findings are somewhat reassuring in relation to the generalizability of the PCL–R to African American participants, it is important to emphasize the limitations of the present findings. First, the IRT and factor analyses do not unequivocally demonstrate that African American and Caucasian individuals earning high scores on the PCL–R are characterized by similar underlying mechanisms. The validation of the psychopathy construct across ethnic groups remains a stepwise process. In particular, it remains possible that PCL–R items could function similarly to allow the identification of African American and Caucasian offenders with a high versus low proportion of psychopathic traits, yet these psychopathic and nonpsychopathic participants might still differ from each other in some underlying emotional and cognitive mechanisms. In particular, whereas similar deficits have been reported for African American and Caucasian psychopaths with respect to affect recognition and divided attention during left-hemisphere activation (Kosson, 1998; Kosson et al., 2001), differences in passive avoidance, attention to peripheral contingencies, and appraisal appear less robust in African American than in Caucasian psychopaths (Kosson, 1998; Newman & Schmitt, 1998; Vitale & Newman, 1998).

Having established a common metric for the measurement of the construct of psychopathy, it is also important to examine forms of validity other than construct validity. In the context of the criminal-justice application of the instrument, predictive validity is particularly important. For example, current evidence suggests that PCL–R scores in African American inmates are equally predictive of lifetime criminal activity and criminal versatility (Kosson et al., 1989; Kosson et al., 1990). Whether PCL–R scores are equally predictive of nonviolent and violent recidivism in different ethnic groups is an important question for future research: Bias in predictive validity can occur even in the absence of measurement bias (Millsap, 1997).

To the extent that these findings of measurement invariance across ethnicity generalize to other studies, it rules out one possible explanation for observed differences in performance on experimental tasks (e.g., Kosson, 1998; Kosson et al., 1990; Newman & Schmitt, 1998). However, other explanations require exploration, including ethnic "differences in motivational factors, perceptions of the task requirements, or the fact that the experimenters in these studies all have been White" (Bodholdt, Richards, & Gacono, 2000, p. 67).

Bodholdt et al. (2000) argued that the scoring of some items may be affected by ethnically based rater–offender interactions; they also argued that even observed differences on experimental tasks (e.g., Kosson, 1998; Kosson et al., 1990; Newman & Schmitt, 1998) may be a consequence of the experimenter being Caucasian. Such Rater × Participant Ethnicity effects can be assessed. One explanation for the cross-cultural effects reported by Cooke, Hart, Michie, and Hare (2001) is that Scottish raters systematically underrate PCL–R items. However, using a split-plot factorial design in which Scottish and Canadian raters scored videotaped interviews of Scottish and Canadian prisoners, Cooke et al. found no main effect of rater nationality or Rater Nationality × Prisoner Nationality interaction. The impact of Rater × Participant Ethnicity could be assessed in the same manner.

In conclusion, the overall finding that the PCL–R was not biased in this study does not mean that the test cannot be used in a biased fashion. Awareness of that possibility is important, particularly given the serious implications that a high PCL–R score can have for an individual's liberty.

References

American Educational Research Association & American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barrett, P., & Eysenck, S. B. G. (1984). The assessment of personality factors across 25 countries. *Personality and Individual Differences, 5*, 615–632.

Bijnen, E. J., & Poortinga, Y. H. (1988). The questionable value of cross-cultural comparisons with the Eysenck Personality Questionnaire. *Journal of Cross-Cultural Psychology, 19*, 193–202.

Bodholdt, R. H., Richards, H. R., & Gacono, C. B. (2000). Assessing psychopathy in adults: The Psychopathy Checklist—Revised and Screening Version. In C. B. Gacono (Ed.), *The clinical and forensic assessment of psychopathy: A practitioner's guide* (pp. 55–86). Mahwah, NJ: Erlbaum.

Bontempo, R. (1993). Translation fidelity of psychological scales: An item

response theory analysis of an Individualism–Collectivism Scale. *Journal of Cross-Cultural Psychology, 24*, 149–166.

Brandt, J. R., Kennedy, W. A., Patrick, C. J., & Curtin, J. J. (1997). Assessment of psychopathy in a population of incarcerated adolescent offenders. *Psychological Assessment, 9*, 429–435.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. London: Sage.

Cleckley, H. (1976). *The mask of sanity* (5th ed.). St Louis, MO: Mosby.

Conoley, J. C., & Impara, J. C. (1995). *Twelfth mental measurement yearbook*. Lincoln, NE: Buros Institute.

Cooke, D. J. (1998). Psychopathy across cultures. In D. J. Cooke, A. E. Forth, & R. D. Hare (Eds.), *Psychopathy: Theory, research and implications for society* (pp. 13–45). Dordrect, The Netherlands: Kluwer Academic.

Cooke, D. J., Forth, A., & Hare, R. D. (1998). *Psychopathy: Theory, research and implications for society*. Dordrect, The Netherlands: Kluwer Academic.

Cooke, D. J., Hart, S. D., Michie, C., & Hare, R. D. (2001). *The impact of rater nationality on Psychopathy Checklist Revised scores: Canada and Scotland compared*. Unpublished manuscript.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist—Revised. *Psychological Assessment, 9*, 3–14.

Cooke, D. J., & Michie, C. (1999). Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology, 108*, 55–68.

Cooke, D. J., & Michie, C. (2001). Refining the construct of psychopathy: Towards a hierarchical model. *Psychological Assessment, 13*, 171–188.

Cooke, D. J., Michie, C., & Clark, D. (2000). *Psychopathy across cultures: A replication and extension*. Unpublished manuscript.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1998). The functioning of the Screening Version of the Psychopathy Checklist—Revised: An item response theory analysis. *Psychological Assessment, 11*, 3–13.

Doninger, N. A., & Kosson, D. S. (2001). Interpersonal construct systems among psychopaths. *Personality and Individual Differences, 30*, 1263–1281.

Draguns, J. G. (1986). Culture and psychopathology: What is known about the relationship. *Australian Journal of Psychology, 38*, 329–338.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341–349.

Forth, A. E., & Mailloux, D. L. (2000). Psychopathy in youth: What do we know? In C. B. Gacono (Ed.), *The clinical and forensic assessment of psychopathy: A practitioner's guide* (pp. 25–54). Mahwah, NJ: Erlbaum.

Gottfredson, L. S. (1994). The science and politics of race norming. *American Psychologist, 49*, 955–963.

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

Hare, R. D. (1998). The Hare PCL–R: Some issues concerning its use and misuse. *Legal and Criminological Psychology, 3*, 101–119.

Hare, R. D., Cooke, D. J., & Hart, S. D. (1999). Psychopathy and sadistic personality disorder. In T. Millon, P. H. Blaney, & R. D. Davies (Eds.), *Oxford textbook of psychopathology* (pp. 555–584). London: Oxford University Press.

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version* (1st ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Hart, S. D., & Hare, R. D. (1997). Psychopathy: Assessment and association with criminal conduct. In D. M. Stoff, J. Maser, & J. Breiling (Eds.), *Handbook of antisocial behaviour* (pp. 22–35). New York: Wiley.

Hemphill, J. F., Hare, R. D., & Wong, S. (1998). Psychopathy and recidivism: A review. *Legal and Criminological Psychology, 3*, 139–170.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning* (1st ed.). Hillsdale, NJ: Erlbaum.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28*, 192–218.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kosson, D. S. (1998). Divided visual attention in psychopathic and nonpsychopathic offenders. *Personality and Individual Differences, 24*, 373–391.

Kosson, D. S., Smith, S. S., & Newman, J. P. (1989, June). *Applicability of the psychopathy construct to Black male inmates: A preliminary investigation*. Paper presented at the fourth annual meeting of the International Society for the Study of Individual Differences, Heidelberg, Germany.

Kosson, D. S., Smith, S. S., & Newman, J. P. (1990). Evaluating the construct validity of psychopathy in Black and White male inmates: Three preliminary studies. *Journal of Abnormal Psychology, 99*, 250–259.

Kosson, D. S., Suchy, U., Mayer, A. R., & Libby, J. (2001). *Facial affect recognition in criminal psychopaths*. Manuscript submitted for publication.

Loesel, F. (1998). Treatment and management of psychopaths. In D. J. Cooke, A. Forth, & R. D. Hare (Eds.), *Psychopathy: Theory, research and implications for society* (pp. 303–354). Dordrecht, The Netherlands: Kluwer Academic.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Hillsdale, NJ: Erlbaum.

Lyon, D., & Ogloff, J. R. P. (2000). Legal and ethical issues in psychopathy assessment. In C. B. Gacono (Ed.), *The clinical and forensic assessment of psychopathy* (pp. 139–173). Mahwah, NJ: Erlbaum.

MacCallum, R. C., & Austin, J. T. (2000). Application of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293–299.

Meredith, W., & Millsap, E. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289–311.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248–260.

Myers, W. C., Burket, R. C., & Harris, H. E. (1995). Adolescent psychopathy in relation to delinquent behaviors, conduct disorder, and personality disorders. *Journal of Forensic Sciences, 40*, 436–440.

Newman, J. P., & Schmitt, W. A. (1998). Passive avoidance in psychopathic offenders: A replication and extension. *Journal of Abnormal Psychology, 107*, 527–532.

Newman, J. P., Schmitt, W. A., & Voss, W. D. (1997). The impact of motivationally neutral cues on psychopaths: Assessing the generality of the response modulation hypothesis. *Journal of Abnormal Psychology, 106*, 563–575.

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

Okazaki, S., & Sue, S. (1995) Methodological issues in assessment research with ethnic minorities. *Psychological Assessment, 7*, 367–375.

Paris, J. (1998). Personality disorders in sociocultural perspective. *Journal of Personality Disorders, 12*, 289–301.

Quinsey, V. L. E., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk* (1st ed.). Washington, DC: American Psychological Association.

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based

internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Rice, M. E., Harris, G. T., & Cormier, C. A. (1992). An evaluation of a maximum security therapeutic community for psychopaths and other mentally disordered offenders. *Law and Human Behavior, 16*, 399–412.

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

Santor, D. A., & Ramsay, J. O. (1999). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10*, 345–359.

Seto, M. C., & Barbaree, H. E. (1999). Psychopathy, treatment behavior, and sex offender recidivism. *Journal of Interpersonal Violence, 14*, 1235–1248.

Steadman, H., Silver, E., Monahan, J., Appelbaum, P., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H., & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*, 83–100.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Measurement, 1*, 81–97.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: Perfect matching factor structures and still biased? *European Journal of Psychological Assessment, 11*, 194–201.

Thissen, D. (1991). *Multilog user's guide* (Version 6). (1st ed.). Mooresville, IN: Scientific Software.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J., & Poortinga, Y. H. (1994). Methodological issues in cross-cultural studies on parental rearing behavior and psychopathology. In C. Perris, W. A. Arrindell, & E. Eisemann (Eds.), *Parental rearing and psychopathology* (pp. 173–197). Chichester, UK: Wiley.

Vassileva, J., Kosson, D. S., & Conrod, P. (2001). *Classification of criminal offenders based on psychopathy and theoretically related constructs*. Manuscript submitted for publication.

Vitale, J. E., & Newman, J. P. (1998, May). *Social information processing in incarcerated adult male offenders*. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group difference on homogenous *and* heterogenous scales: An illustration with the MMPI. *Psychological Methods, 5*, 125–146.

Webster, C. D., Douglas, K., Eaves, D., & Hart, S. D. (1997). *HCR–20 assessing risk for violence* (2nd ed.). Vancouver, British Columbia, Canada: Simon Fraser University.

Zinbarg, R. E., Barlow, D. H., & Brown, T. A. (1997). Hierarchical structure and general factor saturation of the anxiety sensitivity index: Evidence and implications. *Psychological Assessment, 9*, 277–284.

Zumbo, B. D. (1999). *A handbook of the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Received September 19, 2000
Revision received May 23, 2001
Accepted May 28, 2001 ■