Psychological Assessment
1997, Vol. 9, No. 1, 3–14

1040-3590/97/$3.00

An Item Response Theory Analysis of the Hare Psychopathy Checklist—Revised

David J. Cooke
Glasgow Caledonian University and Greater Glasgow Health Board Mental Health and Community Trust

Christine Michie
Glasgow Caledonian University

Hare's Psychopathy Checklist—Revised (PCL–R; R. D. Hare, 1991) is the measure of choice for measuring psychopathic personality disorder. An item response theory (IRT) approach was adopted to analyze both test and item functioning. Data from 2,067 North American participants were analyzed. The analysis confirmed that the test was appropriate for both the diagnosis of psychopathic personality disorder and for making measures of trait strength. Two correlated but distinct factors underpin scores on the PCL–R: Factor 1, Selfish, Callous, and Remorseless Use of Others, and Factor 2, Chronically Unstable and Antisocial Life style. Items related to Factor 1 are generally more discriminating and provide more information about the trait than items relating to Factor 2. Future uses of IRT procedures in the analysis of PCL–R data are discussed.

Hare's Psychopathy Checklist (PCL; Hare, 1980) and its recent revision (PCL–R; Hare, 1991) are rating scales designed to measure psychopathic personality disorder. The PCL–R uses information collected by interview and file review to rate the participant on 20 characteristics. These characteristics include the behavioral, affective, and interpersonal characteristics thought to define psychopathic personality disorder (Cleckley, 1976; Hare, 1970). The PCL–R is regarded as the instrument of choice for measuring psychopathic personality disorder (Conoley & Impara, 1995).

The whole test can be used to provide either a diagnosis of psychopathy or a trait measure of psychopathic personality disorder (Hare et al., 1990). The total test score provides an estimate of a higher order construct underpinned by two facets of psychopathic personality disorder (Hare et al., 1990). Detailed analysis has revealed that the factor structure underpinning the PCL and PCL–R can be best described in terms of two distinct yet correlated factors (Cooke, 1995a; Hare et al., 1990; Harpur, Hare, & Hakstian, 1989). The first factor can be characterized as representing the "selfish, callous, and remorseless use of others" and is specified by core personality traits including superficiality, habitual lying, manipulativeness, and callousness, together with a lack of affect, guilt, remorse, and empathy. The second factor, which can be characterized as "chronically unstable and antisocial lifestyle," is specified by characteristics including the need for stimulation, poor behavioral controls, lack of realistic long-term goals, impulsivity, and juvenile delinquency.

The use of the PCL–R in forensic and clinical settings is increasing (Forth, Hart, & Hare, 1990; Hare, 1991; Harris, Rice, & Quinsey, 1993; Hart, Kropp, & Hare, 1988; Rice, Harris, & Cormier, 1992; Serin, Peters, & Barbaree, 1990). The PCL–R has good internal consistency and interrater reliability; it also has a stable factor structure (Cooke, 1995a; Hare et al., 1990). The value of the PCL–R in forensic and clinical settings is supported by evidence of its predictive validity: PCL–R scores predict a variety of antisocial behaviors, including criminal violence, recidivism following release from prison or hospital, and response to correctional treatment programs (e.g., Forth et al., 1990; Hart et al., 1988; Ogloff, Wong, & Greenwood, 1990; Harris, Rice, & Cormier, 1991; Serin et al., 1990).

The validity of the instrument is further sustained—perhaps unusually for a clinical instrument—by an impressive array of laboratory evidence. This evidence illustrates differences, among other things, in the psychopaths' ability to interpret the emotional tone of language (Hare & Jutai, 1988; Williamson, Harpur, & Hare, 1991), their use of hand gestures (Gillstrom & Hare, 1988), their responses to aversive stimuli (Hare, 1978), and their ability to shift attentional focus when faced with competing signals for reward and punishment (Newman & Kosson, 1986; see Hare, 1991, and Cooke, Forth, & Hare, 1996, for a review of this evidence).

Evaluating Tests Through the Use of Item Response Theory Methods

Nunnally and Bernstein (1994) indicated that item response theory (IRT) methods can be regarded as the central component of modern psychometrics. Although these techniques have been widely used with tests of skill, they have rarely been used with personality tests (Nunnally & Bernstein, 1994). IRT methods can answer many of the questions that are important both for test developers and test users. IRT procedures allow detailed

David J. Cooke, Department of Psychology, Glasgow Caledonian University, Glasgow, Scotland, and Department of Forensic Clinical Psychology Services, Greater Glasgow Health Board Mental Health and Community Trust, Glasgow, Scotland; Christine Michie, Department of Psychology, Glasgow Caledonian University, Glasgow, Scotland.

We thank Robert Hare and Joseph Newman for providing us with their raw data. Joseph Newman's data were collected with support from National Institute of Mental Health Grant NH37711.

Correspondence concerning this article should be addressed to David J. Cooke, who is now at Forensic Clinical Psychology Services, Douglas Inch Centre, 2 Woodside Terrace, Glasgow, G3 7UY Scotland.

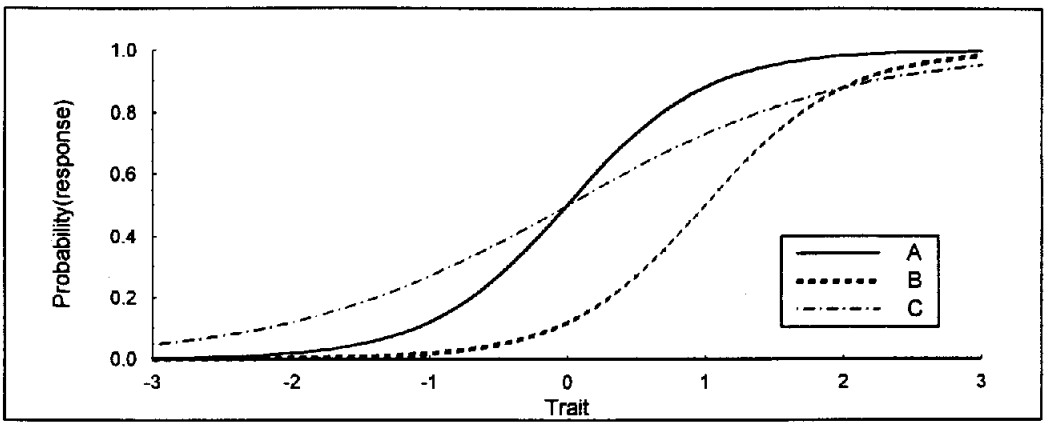

Figure 1. Hypothetical item characteristic curves.

examination of the properties of individual items. Item characteristic curves (ICC) are central features of IRT analyses.

These trace lines plot the probability of a response given the level of the underlying skill, trait, or characteristic being measured. These trace lines can be calculated from empirical data (see later discussion for details). For the purpose of illustration, three hypothetical ICCs for a positive response are plotted in Figure 1. Curve A has a steep slope, thus as the level of the underlying trait increases past a critical level, the probability of a positive response on the item moves rapidly from a low value to a high value. This item discriminates well at this level of the trait. Curves A and B are parallel, thus they are equally discriminating; however, Item B discriminates at a higher level of the trait. If an IRT analysis of the Vocabulary subtest of the Wechsler Adult Intelligence Scale—Revised (WAIS–R; Weschler, 1981) were carried out, then Curve A might relate to an easy item such as "bed" or "winter," whereas Curve B might relate to a hard item such as "remorse" or "perimeter." Curve C relates to an item that has less discriminative power, as the probability of a positive response only changes slowly with increases in the characteristic of interest.

Careful analyses of these trace lines can assist in at least three ways. First, they can assist in the elimination of items that do not provide any significant information about the trait of interest. Second, they can be used to select items that give accurate assessment across the whole range of a test, or by way of contrast, they can be used to select items that cluster around the diagnostic cutoff and thus provide maximum discrimination in this critical range of the trait. Third, and perhaps of greatest significance, they can be used to identify item bias or, in the more neutral terminology of IRT, differential item function (DIF). DIF occurs when an item is more discriminating or is more difficult or more extreme in one group as compared with another; careful consideration of trace lines can assist in identifying racial, gender, or other biases in a test.

The Relationship Between Classical Test Theory and Item Response Theory

Traditionally, the psychometric properties of the PCL–R have been assessed in terms of classical test theory (CTT; Hare, 1980, 1991; Hare et al., 1990; Harpur, Hakstian, & Hare, 1988; Hart & Hare, 1989); in this article it is demonstrated that IRT methods can increase our understanding of the test's functioning.

CTT and IRT models are overlapping theoretical frameworks for understanding test performance rather than competing frameworks (Hulin, Drasgow, & Parsons, 1983b). IRT models make stronger assumptions than CTT models, in particular, the assumption of local independence and the assumption of logistic relationships between item responses and the underlying trait. IRT models have significant advantages over CTT models, particularly when it comes to the consideration of test bias (Hambleton, 1989; Lord, 1980). Key concepts in CTT including item difficulty (proportion of participants with a positive rating), item discrimination (corrected item-to-total correlation), alpha reliability, and optimal cutoffs are all dependent on the characteristics of the standardization sample. For example, the number of participants rated as having engaged in "juvenile delinquency" will be higher in a prison sample than in a general population sample, thus the item difficulty will vary by sample (Hambleton, 1989; King, King, Fairbank, Schlenger, & Surface, 1993; Nunnally & Bernstein, 1994). Gibbons, Clark, Cavanaugh, and Davis (1985), in an early application of IRT to a clinical assessment procedure, demonstrated the sample-dependent nature of item discrimination. They confirmed that symptoms that reliably discriminated between high and low depression in a sample of psychiatric patients were much less effective in discriminating depression among those who are physically ill.

Corrected item-to-total correlations and alpha reliability coefficients will tend to be higher in heterogeneous samples (Nunnally & Bernstein, 1994). CTT estimates of reliability are less informative than IRT estimates. Estimates of test reliability using CTT are estimated for the mean of the standardization sample, whereas IRT models allow the precision of measurement to be established at any point on the underlying latent trait (King et al., 1993).

IRT curves are not dependent on the sample used to generate the curves (Hambleton, 1989). Thus it is less likely that true group differences will be mistaken for bias. The ability to distin-

guish between true group differences and bias has both political and social importance, which is particularly important now that the PCL–R is being used for significant real-life decisions, including discharge from hospital and release on parole (Shepard, Camilli, & Williams, 1984; Thissen, Steinberg, & Gerrard, 1986).

The total score on the PCL–R, in common with most clinical rating scales, is obtained by the simple addition of item scores. This procedure assumes that equal ratings on each item of the scale represent an equivalent level of psychopathic personality disorder. Nunnally and Bernstein (1994) argued that the number of correct answers is not linearly related to a underlying latent trait; individuals with the same number of correct responses on a test can be shown to have significantly different levels of skill. It may be the case that certain characteristics only become apparent at very high levels of disorder: The occurrence of one of these extreme symptoms would, therefore, have greater diagnostic significance. Gibbons et al. (1985) showed that symptoms of depression such as "work inhibition" and "fatigue" occurred even in mild cases of depression, whereas symptoms such as "feeling like a failure," "guilt," "suicidal thoughts" and "loss of social interest" only occurred in very depressed patients.

A further consequence of CTT models is that a participant's score is dependent on the version of the test used; with IRT methods it is possible to estimate the individual's position on the latent trait independent of the version of the test used. This is important in relation to the suite of instruments measuring psychopathic personality disorder, including the original PCL, the PCL–R, and the screening version (Hart, Hare, & Forth, 1994). In addition, the PCL–R may be scored on the basis of interview and file review or on the basis of file review alone. IRT procedures allow the estimation of the same underlying trait irrespective of the instrument or data collection method used.

A IRT approach may confer certain practical advantages on the test developer and the test user. If the test developer's primary objective is to develop a diagnostic instrument, an instrument designed to allocate participants into a "case" category, then items should be designed with item difficulty parameters or thresholds (b_i, see later discussion for definition) that cluster around the diagnostic cutoff point. When the diagnostic decision is the principal concern, it may be possible to use a short version of the test, using only those items with thresholds near the diagnostic cutoff point. Given that the PCL–R is a time-consuming procedure, a reduction in the number of items could be useful in many practical settings.

Method

Sample

Data from 10 North American samples—8 Canadian and 2 American—were obtained from Robert Hare and Joseph Newman. The Canadian samples include four hospital samples, 80 consecutive remands to a forensic hospital in British Columbia (Hart & Hare, 1989), 163 patients in the forensic psychiatric unit of the Mental Health Centre at Penetanguishene, 132 patients in the Regional Psychiatric Center in Saskatoon, Saskatchewan, 65 patients of a forensic out-patient clinic in Vancouver and four prison samples, 106 prisoners assessed at the Institute Phillipe Pinel de Montréal, 121 inmates of Oakalla provincial prison in British Columbia, 322 inmates of Matsqui federal medium security institution in British Columbia, and 87 inmates of a medium security prison in Kingston, Ontario (see Hare, 1991, for further details of these samples). The data from the United States consisted of two samples of federal prisoners—838 White prisoners and 153 Black prisoners.

Results

Choice of Model

A range of different item response models has been developed (Holland & Wainer, 1993; Hulin, Drasgow, & Parsons, 1983b; Lord, 1980); an important step in model building is the selection of an appropriate mathematical function. The qualities of the data are the most important determinants of which mathematical function should be selected. Item scores on the PCL–R fall into one of three ordered categories: 0 = *does not apply;* 1 = *applies to a certain extent or there is uncertainty that it applies;* 2 = *definitely applies.* The trichotomous nature of the item scores means that the model must include three trace lines or item characteristic curves (ICCs) for each item. Given the ordering of responses, it is parsimonious to make the assumption that as the underlying trait increases the probability of being in Category 0 will decrease, the probability of being in Class 1 will increase then decrease, whereas at high levels of the latent trait the probability of being in Class 2 will increase.

Two-parameter (Parameters a and b_i) logistic functions provide the most appropriate mathematical expressions for describing the trace lines for items such as those in the PCL–R. The interrelation between the probability of each possible response to an item and the latent trait can be summarized by three values of the parameters a, b_1, and b_2. The trace lines that describe how the probability of the Response 0 (i.e., $(P(0))$ and the Response 2 (i.e., $(P(2))$ vary with the level of the trait and are S-shaped curves that are mirror images of each other. These curves are illustrated in Figure 2.

The slopes at the points of inflection are of the same magnitude but opposite in direction and are determined by Parameter a. The a parameter is, therefore, a measure of the discriminating power of the item (Hulin et al., 1983b). The position of the points of inflexion are given by the threshold parameters b_1 for $(P(0))$ and b_2 for $(P(2))$. At these levels of the trait, the probability crosses the 0.5 probability level. The Parameter b_i provides measures of item difficulty or extremity or frequency of a behavior or attitude. Increases in the value of b_i move the curve to the right, increasing the item's level of extremity, unpopularity, or difficulty (Hulin et al., 1983b). Given that the items are assessed by a trained rater, the probability of scoring 1 or 2 at very low levels of the trait should be zero, thus there is no necessity to include a guessing parameter in the model.

A model that fits these assumptions is Samejima's graded model, which is underpinned by two parameter logistic functions (Thissen, 1991). For a unidimensional trait for Item i, the probability of each response, given that the underlying level of the trait is Θ, is given by the following equations:

$$P(Response = 0|\Theta) = 1 - \frac{1}{1 + \exp[-a_i(\Theta - b_{i1})]}, \quad (1)$$

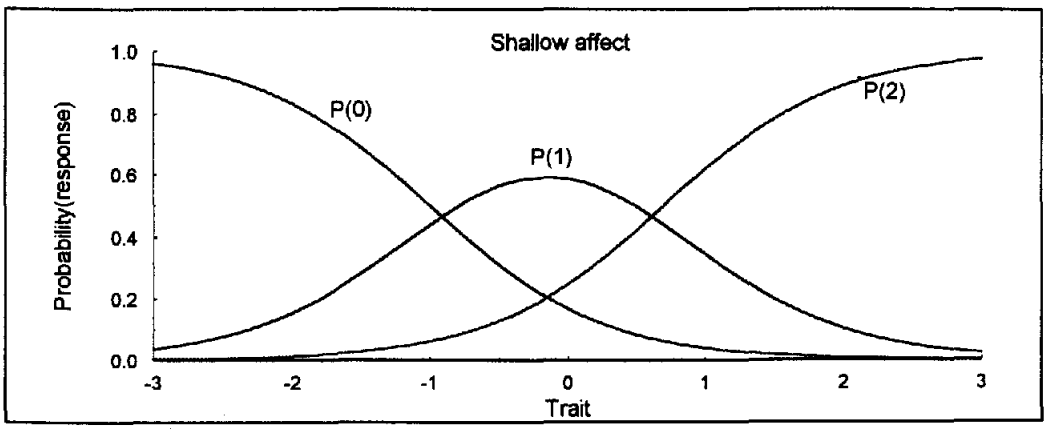

Figure 2. Item characteristic curve for item "shallow affect."

$$P(Response = 1|\Theta) = \frac{1}{1 + \exp[-a_i(\Theta - b_{i1})]} - \frac{1}{1 + \exp[-a_i(\Theta - b_{i2})]}, \quad (2)$$

and

$$P(Response = 2|\Theta) = \frac{1}{1 + \exp[-a_i(\Theta - b_{i2})]}, \quad (3)$$

where Θ is the underlying trait, a is the slope of the trace lines at the inflexion points, b_1 is the value of Θ below which the probability of the item being rated 0 is below 0.5, and b_2 is the value of Θ above which the probability of the item being rated 2 is above 0.5.

Assumption of Homogeneity

Although IRT models are known to be robust to departures from homogeneity (Hulin, Drasgow, & Parsons, 1983a), homogeneity is an assumption underpinning the model. There is no agreed benchmark for the homogeneity of a scale; however, an evaluation of a scale's homogeneity may be made by considering Cronbach's alpha, corrected item-to-total correlations, and the comparative percentages of variance absorbed by the first and second unrotated components in a component analysis.

Hare (1991) argued that within North America the PCL–R shows considerable homogeneity: Cronbach's alpha ranges from .83 to .91 in the standardization samples, the mean corrected item-to-total correlation is .49 in the pooled standardization sample, and the ratio between the first and second unrotated components is approximately 3:1. For the current sample (N = 2,067) Cronbach's alpha is .80, the mean corrected item-to-total correlation is .40, and the ratio between the first and second unrotated components is 3.1:1.

Hare (1991) suggested that this finding of homogeneity is not inconsistent with the finding that the PCL and PCL–R are underpinned by two distinct yet correlated factors, factors that have an average correlation of .5 (Cooke, 1995a; Hare et al., 1990; Harpur et al., 1989). Hare (1991) argued that it makes heuristic sense to view the overall test score as representing a higher order construct entailing the two correlated factors. On theoretical grounds, therefore, it makes sense to consider item characteristic curves in relation to both the total test and the individual factors. Traditionally, users of the test have considered only the total score; however, increasingly users are becoming aware of the value of using factor scores as well (Hare et al., 1990; Harpur & Hare, 1991; Harpur et al., 1989). Thus it also makes sense, from a practical point of view, to consider item performance in relation to both the total score and individual factor scores.

Item Response Curves for the Full Test

There are 10 subsamples in total; however, some of the eight subsamples from Canada are very small in relation to the number of parameters in the IRT model. It was decided, therefore, to consider the data from Canada as consisting of two subsamples: the hospital participants (n = 440) and the prison participants (n = 636).

Samejima's graded model was fitted separately using *Multilog* (Thissen, 1991) to the data for each of the four resulting subsamples (i.e., two Canadian and two American samples). This full model has four sets of IRT parameters estimated by maximum likelihood, and the procedure yields a value of the statistic G^2 = −2log likelihood. The method of generalized likelihood ratio testing (GLRT) was used to test whether the parameters could be considered equal in the four subsamples. The IRT model was refitted, constraining the parameters to be equal for all four subsamples but allowing the mean level of trait to vary between subsamples. According to the theory of GLRT, this results in an increase in G^2 over the baseline value from the full model. Under certain conditions this increase is distributed as a chi-square statistic with degrees of freedom equal to the number of extra parameters in the less constrained model. If the statistic is large enough to be significant, this indicates that there is differential item functioning (DIF; i.e., some of the items are behaving differently in the subsamples). In this case the increase in G^2 was not significant. This analysis demonstrates that the data from the four North American subsamples can be adequately explained by a graded model with

Table 1
Distribution of Individual PCL–R Items by Category and Parameters Fitted for Complete Test

	Category						Item parameters		
	0		1		2				
Item	*n*	%	*n*	%	*n*	%	*a*	b_1	b_2
Glibness/superficial charm	790	38	838	41	438	21	1.3	−0.2	1.6
Grandiose sense of self-worth	666	32	841	41	559	27	1.4	−0.5	1.2
Need for stimulation	287	14	648	31	1,128	55	1.5	−1.5	0.1
Pathological lying	559	27	871	42	633	31	1.4	−0.7	1.1
Conning/manipulative	610	30	803	39	648	31	1.4	−0.6	1.1
Lack of remorse or guilt	207	10	621	30	1,230	60	1.6	−1.7	−0.1
Shallow affect	416	20	826	40	819	40	1.6	−1.0	0.7
Callous/lack of empathy	312	15	847	41	905	44	1.9	−1.2	0.5
Parasitic lifestyle	401	20	1,009	49	630	31	0.9	−1.6	1.4
Poor behavioral controls	442	22	699	34	914	44	0.9	−1.4	0.6
Promiscuous sexual behavior	631	31	528	26	861	43	0.7	−0.9	0.8
Early behavior problems	726	38	471	25	697	37	0.9	−0.4	0.9
Lack of long-term goals	334	16	756	37	971	47	1.2	−1.5	0.4
Impulsivity	178	9	613	30	1,271	62	1.4	−2.0	−0.2
Irresponsibility	179	9	699	34	1,175	57	1.3	−2.0	0.0
Failure to accept responsibility	389	19	736	36	933	45	1.0	−1.5	0.5
Short-term marital relationships	1,075	58	354	19	412	22	0.6	0.8	2.4
Juvenile delinquency	677	34	390	19	943	47	0.7	−0.8	0.5
Revocation of conditional release	460	23	373	19	1,130	58	0.7	−1.5	−0.2
Criminal versatility	765	37	706	34	587	29	0.8	−0.5	1.6

Note. PCL–R = Psychopathy Checklist—Revised.

the same parameters. Although the mean level of psychopathic personality disorder varies among these subsamples, the items behave similarly in all four populations sampled. The fitted parameters are displayed in Table 1. Examination of the threshold parameters (b_1 and b_2) for the aggregate North American sample indicates that there is a sizable variation in the degree of the trait at which items are more likely than not to receive a 1 or 2 rating, respectively. This variation in parameter values confirms that the PCL–R is a good measure of trait strength because there are items that discriminate at all levels of the trait. The thresholds are roughly in line with the prevalence rates for each category (i.e., high thresholds are associated with low levels in Category 2).

Table 1 reveals that items such as "callous/lack of empathy," "shallow affect," and "lack of remorse or guilt" are particularly discriminating (i.e., large *a* parameters), whereas "juvenile delinquency," "many short-term marital relationships," and "revocation of conditional release" are the least discriminating (i.e., small *a* parameters).

Three item characteristic curves are plotted in Figure 3 for illustration. Figure 3, top panel, contains the curves for the item "glibness/superficial charm." The curve for the probability of a 2 response is moderately steep, showing that the item has good discrimination and that a 2 response tends to occur only at high levels of the trait. By way of contrast, the item "irresponsibility" has the same slope for a 2 response but has a higher probability of occurring at lower levels of the trait than "glibness/superficial charm." Figure 3, bottom panel, contains the curves for "callous/lack of empathy"; this item has a very steep slope for a 2 response, indicating that it is a very discriminating item.

Examination of the threshold parameters indicates that items that load on Factor 1, Selfish, Callous, and Remorseless Use of Others, have larger threshold parameters than items that load on Factor 2, Chronically Unstable and Antisocial Lifestyle (Mann Whitney *U* test for b_1 and b_2, $p < .05$). This result is consistent with the findings on a Scottish data set (Cooke & Michie, 1995) and with findings based on the Screening Version of the PCL as it is applied in North American samples (Cooke, Michie, Hart, & Hare, 1995).

The total score of the PCL–R is used for both diagnostic purposes (making categorical decisions) and for estimating trait strength. The total score is obtained by the simple addition of item scores. The foregoing IRT analysis reveals that the relative importance of the items varies with the level of the trait. To determine whether the total PCL–R score is an adequate estimate of the latent trait for diagnostic purposes, cross-classification tables were generated using the standard cutoff points of 30 and 20 on the PCL–R and equivalent cutoff points for the estimated underlying trait. The relationship between Θ (Θ being the estimate of the latent trait from the complete test) and the total PCL–R score was examined. The relationship was virtual unity ($r = .98$) and linear. The standard cutoff of 30 on the PCL–R total score was equivalent to 1.06 (effectively 1) on the latent trait, and a cutoff of 20 on the PCL–R total score was equivalent to −0.05 (effectively 0) on the latent trait. The cutoff of 30 and 1 resulted in 233 participants being misclassified (misclassification rate = 11%), whereas the cutoff of 20 and 0 resulted in 64 participants being misclassified (misclassification rate = 3.2%).

As noted earlier, previous analysis indicated that these data were underpinned by two distinct but correlated factors (Cooke,

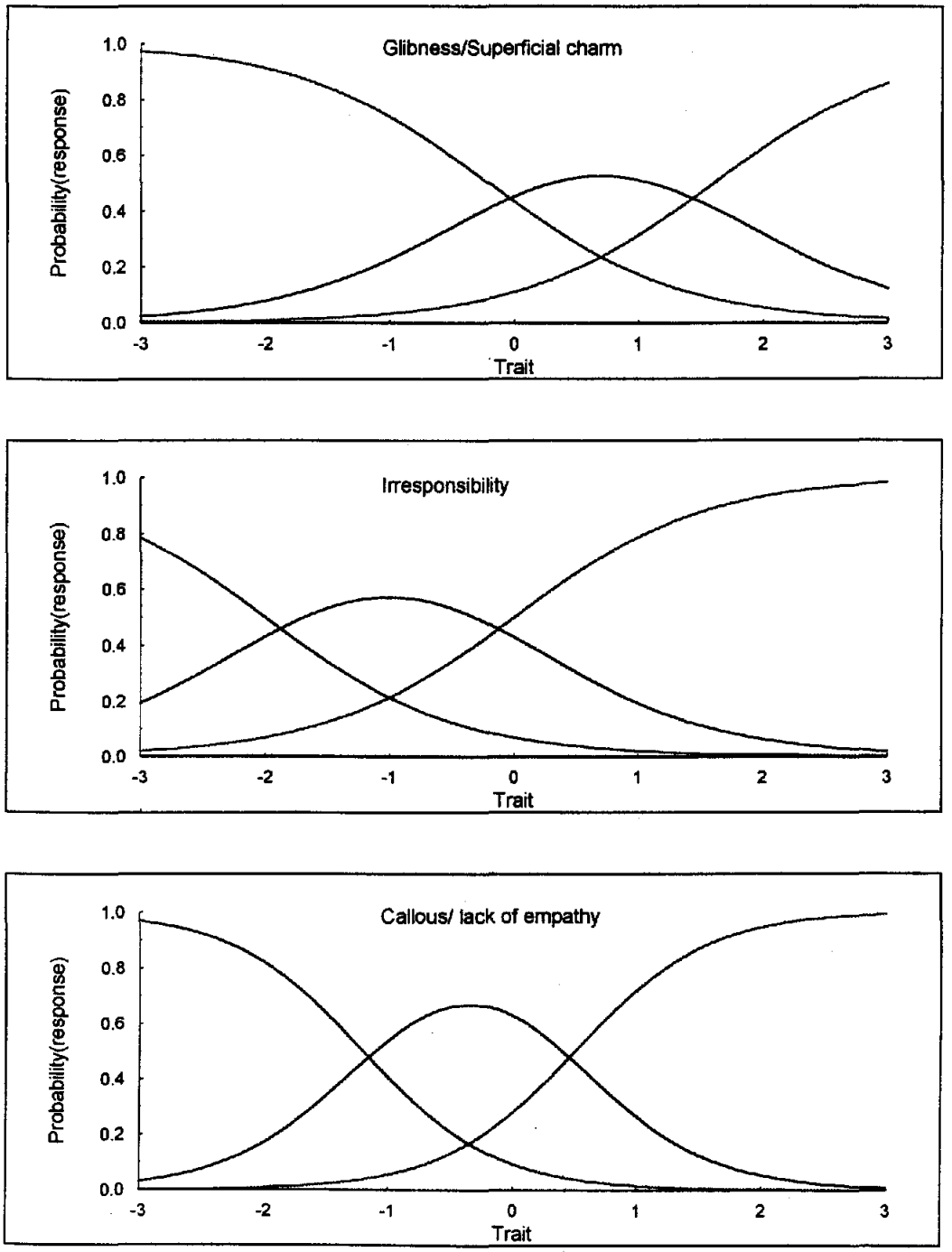

Figure 3. Item characteristic curves for items "glibness/superficial charm," "irresponsibility," and "callous/lack of empathy."

1995a; Harpur et al., 1989). Because Samejima's graded model is designed to apply to unidimensional underlying traits, the two factors were analyzed separately.

Factor 1: Selfish, Callous, and Remorseless Use of Others

Once again, Samejima's graded model was applied to the estimation of the item characteristic curves; on this occasion the model was used to assess the eight items that load on Factor 1. The parameters of the model fitted for all participants are shown in Table 2. The larger the slope, the more discriminating the item. The slope or *a* parameters of the ICC vary to some degree, from 1.2 to 2.0. Two items—"lack of remorse or guilt" and "grandiose sense of self-worth"—are particularly discriminating. The values for the thresholds show that an individual is likely to be assessed as Category 2 for some items, such as

Table 2
Parameters For Each PCL–R Item Fitted for Separate Factors

Item	Factor	Item parameters		
		a	b_1	b_2
Glibness/superficial charm	1	1.7	−0.3	1.3
Grandiose sense of self-worth	1	1.9	−0.5	0.9
Need for stimulation	2	2.0	−1.5	−0.2
Pathological lying	1	1.5	−0.8	0.9
Conning/manipulative	1	1.4	−0.8	0.9
Lack of remorse or guilt	1	2.0	−0.6	−0.2
Shallow effect	1	1.6	−1.1	0.5
Callous/lack of empathy	1	1.9	−1.3	0.3
Parasitic lifestyle	2	1.2	−1.5	0.9
Poor behavioral controls	2	1.0	−1.5	0.3
Promiscuous sexual behavior	None			
Early behavior problems	2	1.3	−0.5	0.5
Lack of long-term goals	2	1.4	−1.6	0.1
Impulsivity	2	2.0	−1.8	−0.4
Irresponsibility	2	1.4	−2.2	−0.3
Failure to accept responsibility	1	1.2	−1.4	0.3
Short-term marital relationships	None			
Juvenile delinquency	2	1.0	−0.8	0.1
Revocation of conditional release	2	0.7	−1.7	−0.4
Criminal versatility	None			

Note. PCL–R = Psychopathy Checklist—Revised.

"glibness and superficial charm," only at very high levels of the underlying trait. Other items, such as "lack of guilt or remorse," discriminate at lower levels of the trait.

Factor 2: Chronically Unstable and Antisocial Lifestyle

Using *Multilog,* the logistic curves of Samejima's graded model were fitted to the data for nine items that load on Factor 2. The results for all participants are shown in Table 2. The items with the highest slopes, the items that discriminate best, are "need for stimulation" and "impulsivity." "Revocation of conditional release" is the least discriminating item. In general, participants are less likely to be assessed as Category 0 and more likely to be assessed as Category 2 for the items that compose Factor 2. This is reflected in the thresholds that are all centrally placed in the range. "Impulsivity" has a threshold at a comparatively low level of the trait, whereas "parasitic lifestyle" has a threshold that is high on the trait.

Information Functions

The item parameter estimates provided by Samejima's graded model can be used to produce item information functions, not only for the individual items in relation to the specific latent traits but also for the whole test. Item information is asymptotically 1 over the square of the standard error; the more information provided by an item or by a test, the greater the precision with which the item or test score is estimated. Item information provides an estimate of the item's reliability at different points on the latent trait. Examination of test information allows determination of the precision of estimates at different points on the test. If the primary purpose of a test is to provide diagnostic information, then information should be clustered near the diagnostic cutpoint. If a test is designed to measure trait strength, then information should be maximized across as broad a range of the trait as is possible. Examination of individual item information functions allows the estimation of the overall precision of these items, and in particular this reveals the point on the trait where precision is maximal.

The information for the test, for the two factors, and the 20 items of the complete tests at different levels of the trait are shown in Table 3. Total PCL–R scores and Factor 1 and Factor 2 scores are provided in order that the level of information at particular scores can be assessed. The information functions have two maxima at b_1 and b_2. This pattern is most obvious for "callous/lack of empathy," with other items showing a plateau between the two thresholds.

Examination of the information functions for the trait estimated from all the items and the traits specified by the individual factors indicates that for the complete test and for Factor 1 the maximum amount of information is near the center of the distribution. This is not unexpected given that the quality of the estimate is higher at the point where there is the greatest number of cases to estimate it. By way of contrast, the maximum amount of information on Factor 2 is at or around the trait level of −0.5. There is a distinct fall-off in information above a trait level of 1.0; this level is equivalent to a the diagnostic cutoff of 30 on the PCL–R, and thus this loss of precision is not important because the diagnostic decision has been made.

Nunnally and Bernstein (1994) indicated that it possible to estimate the relative efficiency of tests at a particular point on a trait by calculating the ratio of their respective item information functions. For example, at the diagnostic cutoff of 1.0, the Trait Factor 1, although only containing eight items, has a relative efficiency of 74% compared with that of the full test. The relative efficiency of Factor 2 at this point, although it contains one more item than Factor 1, is only 46%.

The functioning of items can be further explored by examining the distribution of item information across the trait. Figure 4, top panel, contains four Factor 1 items (i.e., "glibness/superficial charm," "grandiose sense of self-worth," "pathological lying," and "conning/manipulative") that all have high levels of information, with the maximum information being at high levels of the trait around the diagnostic cutoff points of 0 and 1 (equivalent to PCL–R scores of 20 and 30, respectively). Figure 4, middle panel, illustrates four Factor 1 items (i.e., "callous/lack of empathy," "lack of remorse," "shallow affect," and "failure to accept responsibility") that plateau at lower levels of the trait, and all, other than "failure to accept responsibility," display high levels of information. The bottom panel of Figure 4 illustrates the information function of four Factor 2 items: "need for stimulation," "impulsivity," "irresponsibility," and "lack of goals." These items display high levels of information at low levels of the trait and provide comparatively little information at higher levels of the trait. The other eight items had information functions that were low and flat, suggesting that they contribute little to the estimation of the trait.

Factor 1 items tended to have more information than Factor 2 items; comparing the average maxima of Factor 1 and Factor 2 items revealed significant differences, Mann-Whitney U test,

Table 3
Information × 100 For Individual PCL–R Items, Total PCL–R Score, and Factor 1 and Factor 2 Scores, at Various Levels of the Trait

	Estimated trait								
Item	−2.0	−1.5	−1.0	−0.5	0.0	0.5	1.0	1.5	2.0
Glibness/superficial charm	14	23	34	43	47	47	47	46	41
Grandiose sense of self-worth	18	31	45	54	56	55	56	51	39
Need for stimulation	47	58	60	60	59	51	36	22	12
Pathological lying	24	37	48	53	53	53	53	46	34
Conning/manipulative	21	34	48	56	57	57	56	48	34
Lack of remorse or guilt	61	69	68	69	67	52	33	18	9
Shallow affect	34	52	64	65	64	65	59	42	25
Callous/lack of empathy	50	80	92	84	86	91	72	41	19
Parasitic lifestyle	19	21	21	21	21	21	21	21	19
Poor behavioral controls	21	24	25	26	26	25	22	19	15
Promiscuous sexual behavior	12	13	15	15	15	15	14	13	11
Early behavior problems	13	17	21	24	25	25	24	21	17
Lack of long-term goals	32	36	37	37	37	36	31	24	16
Impulsivity	49	50	50	50	47	37	25	15	9
Irresponsibility	45	45	45	46	45	39	28	18	11
Failure to accept responsibility	23	26	27	27	27	26	23	19	15
Short-term marital relationships	5	6	7	9	10	11	11	12	12
Juvenile delinquency	11	12	13	14	14	14	13	11	9
Revocation of conditional release	13	14	15	15	14	13	11	9	7
Criminal versatility	12	15	17	19	20	20	20	19	18
PCL–R score	0	5	10	15	20	25	30	35	40
Information for complete test	620	760	850	890	890	850	760	620	470
Factor 1 score	0	2	4	6	8	10	12	14	16
Information for Factor 1	410	550	640	680	690	640	560	430	300
Factor 2 score	0.0	2.2	4.5	6.8	9.0	11.2	13.5	15.8	18.0
Information for Factor 2	490	550	560	580	550	450	350	260	200

Note. PCL–R = Psychopathy Checklist—Revised.

$U(8,9) = 9$, $p = 0.01$. All Factor 2 items have low levels of information at the cutoff of 1.

Discussion

To our knowledge, this is first attempt to analyze PCL–R data using IRT methods. The IRT analysis of the four North American subsamples reveals remarkable consistency in the performance of this test in different settings and with different cultural groups. There is no evidence detectable in these comparatively large samples that suggests that the test is biased due to race or the presence of mental disorder. The analysis confirms that the PCL–R is a good measure of psychopathic personality disorder because all the items contribute to the estimate of the trait and there are different items that function efficiently, at different points, along the whole length of the trait.

It is reassuring to discover that the simple summation method for generating a total score on the PCL–R does not lead to any significant misclassification of cases. This finding reflects the fact that the b_i parameters are well distributed along the whole length of the trait.

Factor 1 is More Important Than Factor 2

Examination of the b_i parameters derived from this data set confirms (Cooke & Michie, 1995) results from a large representative sample of the Scottish prison population and results obtained with the Clinical Version of the PCL (Cooke & Michie, 1995; Cooke et al., 1995); these results demonstrated that Factor 1 items, those that load on the Selfish, Callous, and Remorseless Use of Others factor, have a statistically significant tendency to occur at higher levels of the trait than do items that load on Factor 2, the Chronically Unstable and Antisocial Lifestyle factor. Factor two items are positive even at comparatively low levels of the trait. Measures of information are related to but distinct from measures of extremity or difficulty (i.e., b_i parameters), and examination of the information functions indicates that Factor 1 items in general have greater precision in defining the trait than do Factor 2 items.

In the clinical literature, Factor 1 items have generally been regarded as the most central features of the disorder (Cleckley, 1976); this may reflect the fact that these items only become apparent in the extreme and perhaps most obvious cases of the disorder. Hare (1991) indicated that Factor 1 consistently correlates more highly with a prototypicality rating of psychopathy than does Factor 2. Despite Hare's (1991) argument that psychopathic personality disorder is best considered to be a higher order construct that overarches two distinct yet related subordinate constructs, it is clear that these constructs are not of equal importance. These findings confirm the clinical view that Factor 1 items are more protypical than Factor 2 items.

These results have implications for the classification of antisocial personality disorder under the fourth edition of the *Diag-*

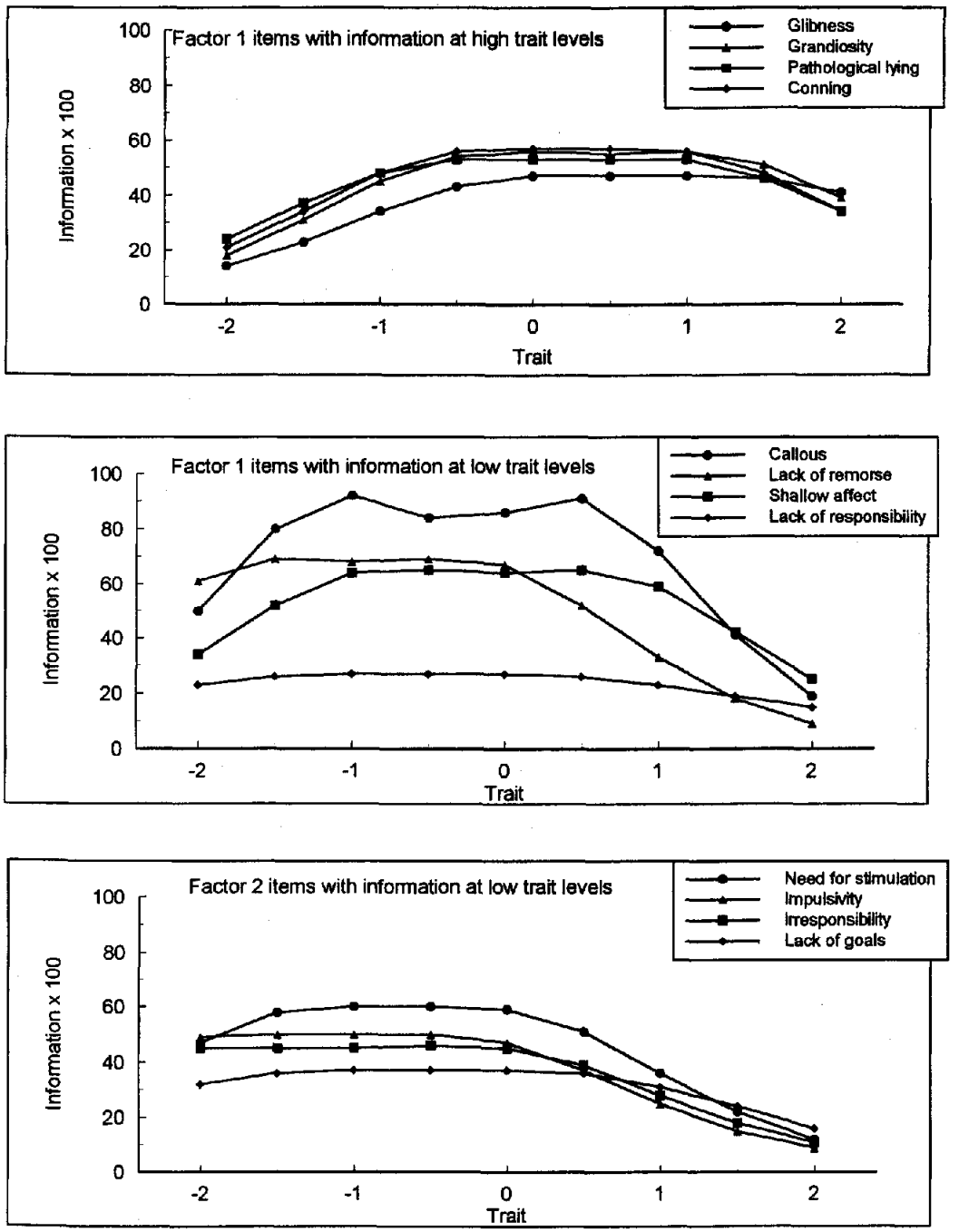

Figure 4. Selected item information functions.

nostic and Statistical Manual of Mental Disorders (*DSM–IV;* (American Psychiatric Association, 1994). Although early versions of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1952, 1968) focused on personality traits, this focus was shifted to antisocial behavior in the third edition (American Psychiatric Association, 1980) and its revision (American Psychiatric Association, 1987); the emphasis on behavioral characteristics continues in *DSM–IV.* There is a suggestion within *DSM–IV* that the Factor 1 items may, in certain contexts, have greater diagnostic significance.

> Lack of empathy, inflated self-appraisal, and superficial charm are features that have been commonly included in traditional conceptions of psychopathy and may be *particularly distinguishing* of Antisocial Personality Disorder in prison or forensic settings where criminal, delinquent, or aggressive acts are likely to be nonspecific. [italics added] (American Psychiatric Association, 1994, p. 647)

The results of this study confirm other findings (Cooke & Michie, 1995; Cooke et al., 1995) that the personality features are the core of the disorder and that they should therefore be given appropriate diagnostic prominence.

The finding that Factor 1 items are most likely to occur in the most extreme cases may explain the established finding that although Factor 2 items are related to socioeconomic status, educational attainment, and family of origin, Factor 1 items are independent of these variables (Hare, 1991; Hare et al., 1990; Harpur et al., 1989). It is likely that psychopathic personality disorder occurs as a consequence of the concatenation of biological vulnerability, critical early experiences, and social pressures (Paris, 1993). At extreme levels the trait emerges unmodified by the social context; the absence of any moderating effect of social context on Factor 1 items tends to implicate biological processes.

Other Applications of Item Response Theory

This study is an initial attempt to use IRT methods with data collected with the PCL–R. IRT methods could have broader application in this field if they are focused on three distinct problems: the problem of differential item functioning, the problem of cross-cultural differences, and the problem of equating different versions of the instrument, or indeed, the problem of equating different data collection methods.

Differential item functioning. There is a growing awareness that testing is not a neutral or value-free activity (McAllister, 1993; Zieky, 1993). Items and tests may perform differentially with different gender, cultural, and subcultural groups. The presence of test bias has important implications for tests such as the PCL–R that inform decisions regarding risk, dangerousness, and parole; fairness is paramount in these decisions. Item response theory models are extremely useful for detecting differential item functioning and differential test functioning (McAllister, 1993; Shepard et al., 1984; Thissen et al., 1986; Zieky, 1993). The PCL–R and its precursor has been standardized essentially on men in North America (Hare, 1991); although there is some information relating to women, juveniles, and Europeans, this information is extremely limited (e.g., Auf Klinteberg, Humble, & Schalling, 1992; Cooke, 1989; Forth et al., 1990; Haapasalo & Pulkkinen, 1992; Raine, 1985). Although no racial differences or differences due to mental disorder were found in item functioning in this study, this may merely reflect a lack of statistical power. It is unlikely, however, that lack of power can explain the absence of significant differences between the racial groups. There was considerable variation in the parameters across the constituent subsamples in the analyses: Differences between Black prisoners and White prisoners and differences between the combined Canadian and the combined U.S. samples were smaller than differences within the Canadian subsamples. Following Nunnally and Bernstein (1994), it is assumed that no racial differences exist in the performance of this test. "Unless one has evidence to the contrary, ethical and scientific considerations such as the law of parsimony dictate, assuming that there are no group differences in the attribute being measured" (Nunnally & Bernstein, 1994, p. 417).

Cross-cultural comparisons. A second application of IRT is in cross-cultural comparisons of constructs. Cooke (1995a), using traditional methods based on CTT, argued that the construct of psychopathy could be generalized from North America to Scotland. It was argued that the substantial differences in the prevalence of the disorder between Scotland and North America could not be attributed to differences in test functioning. Unfortunately, for the reasons outlined at the beginning of this article, CTT may not be an adequate tool for detecting differential item functioning (Bijnen & Poortinga, 1988; Compton et al., 1991; McAllister, 1993; Poortinga & Van de Vijver, 1987; Hulin, 1987; Shepard et al., 1984; Thissen et al., 1986; Zieky, 1993).

A primary problem that bedevils cross-cultural comparisons is that of ensuring that the construct of interest is measured using the same units in the different settings; IRT methods may be used to tackle the problem of metric equivalence (Reise, Widaman, & Pugh, 1993). To make meaningful comparisons of either prevalence estimates or difference in mean trait strength between two populations, it is essential to demonstrate not only that the same constructs or latent traits are being measured but also that they are being measured using the same units and have the same zero point. A concrete example from the physical sciences may clarify the nature of this problem: Although the Farenheit and Centigrade scales measure the same construct, the units and zero points differ. Within IRT models, because each item can be used to estimate the latent trait, it is possible to use those items that behave similarly in each setting as anchors and develop measures of the latent trait that have metric equivalence in the different settings. IRT methods are being applied currently to determine whether differential test functioning in North America and Scotland may account for the differences in estimated prevalence (Cooke & Michie, 1995).

Analysis of this type does not merely have significance for ensuring equivalence of measurement across settings, it may inform theorizing about the etiology of the disorder. It is likely that the expression of personality disorders, although not necessarily their etiology, will be influenced by, among other things, cultural pressures (Paris, 1993). For example, it might be hypothesized that "grandiosity" and "glibness/superficial charm" will be more frequently expressed in individualistic cultures where competitiveness is emphasized and independence and self-confidence is engendered (Cooke, 1995b). The relative extremity of items in different settings may have psychological significance: It will be interesting to determine whether "grandiosity" and "glibness/superficial charm" occur at more extreme levels of the trait in Scottish samples as compared with these North American samples.

Comparing different versions of the test. Hulin et al. (1983b) indicated that an individual's position on the latent trait is independent of the version of the test used. This may have practical advantages. Research constraints may mean that it is not possible to interview participants and that PCL–R ratings must be made from case records (e.g., Harris, Rice, & Cormier, 1991; Harris, Rice, & Quinsey, 1993; Rice, Harris, & Cormier, 1992). IRT procedures can be used to compare information collected using file review alone with information collected by both interview and file review. Equally, IRT methods could be used to determine whether the screening version of the PCL (Hart et al., 1994) is measuring the same latent trait as the full PCL–R (Cooke et al., 1995).

In conclusion, IRT methods are likely to be useful procedures

for enhancing not only our knowledge of the functioning of the PCL–R and its constituent items but also our understanding of the etiology of this important disorder (Cooke, 1995b).

References

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders.* Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Auf Klinteberg, B., Humble, K., & Schalling, D. (1992). Personality and psychopathy of males with a history of early criminal behaviour. *European Journal of Personality, 6,* 245–266.

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.

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

Compton, W. M., Helzer, J. E., Hwu, H. G., Yeh, E. K., McEvoy, L., Tipp, J. E., & Spitznagel, E. L. (1991). New methods in cross-cultural psychiatry: psychiatric illness in Taiwan and the United States. *American Journal of Psychiatry, 148,* 1697–1704.

Conoley, J. C., & Impara, J. C. (1995). 12th mental measurement yearbook. Lincoln, NE: Buros Institute.

Cooke, D. J. (1989). Containing violent prisoners: An analysis of the Barlinnie Special Unit. *British Journal of Criminology, 29,* 129–143.

Cooke, D. J. (1995a). Psychopathic disturbance in the Scottish prison population: The cross-cultural generalisability of the Hare psychopathy checklist. *Psychology, Crime and Law, 2,* 101–108.

Cooke, D. J. (1995b). Psychopathic personality in different cultures. What do we know? What do we need to find out? *Journal of Personality Disorders, 10,* 23–40.

Cooke, D. J., Forth, A. E., & Hare, R. D. (1996). *Psychopathy: Theory, research and implications for society* (1st ed.). Dordrecht, The Netherlands.

Cooke, D. J., & Michie, C. (1995). *Psychopathy across cultures: An item response theory comparison of Hare's Psychopathy Checklist—Revised.* Manuscript submitted for publication.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1995). *The functioning of the clinical version of the Psychopathy Checklist: An item response theory analysis.* Manuscript submitted for publication.

Forth, A. E., Hart, S. D., & Hare, R. D. (1990). Assessment of psychopathy in male young offenders. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 342–344.

Gibbons, R. D., Clark, D. C., Cavanaugh, S. V., & Davis, J. M. (1985). Application of modern psychometric theory in psychometric research. *Journal of Psychiatric Research, 19,* 43–55.

Gillstrom, B. J., & Hare, R. D. (1988). Language-related hand gestures in psychopaths. *Journal of Personality Disorders, 2,* 21–27.

Haapasalo, J., & Pulkkinen, L. (1992). The Psychopathy Checklist and non-violent offender groups. *Criminal Behaviour and Mental Health, 2,* 315–328.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed; pp. 147–200). London: Collier MacMillan.

Hare, R. D. (1970). *Psychopathy: Theory and research.* New York: Wiley.

Hare, R. D. (1978). Electrodermal and cardiovascular correlates of psychopathy. In R. D. Hare & D. Schalling (Eds.), *Psychopathic disorder: Approaches to research* (pp. 107–143). Chichester, England: Wiley.

Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences, 1,* 111–119.

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

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The Revised Psychopathy Checklist: Descriptive statistics, reliability, and factor structure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 338–341.

Hare, R. D., & Jutai, J. W. (1988). Psychopathy and cerebral asymmetry in semantic processing. *Personality and Individual Differences, 9,* 329–337.

Harpur, T. J., Hakstian, R., & Hare, R. D. (1988). Factor structure of the psychopathy checklist. *Journal of Consulting and Clinical Psychology, 56,* 741–747.

Harpur, T. J., & Hare, R. D. (1991). *Psychopathy and violent behavior: Two factors are better than one.* Unpublished manuscript.

Harpur, T. J., Hare, R. D., & Hakstian, A. R. (1989). Two-factor conceptualization of psychopathy: Construct validity and assessment implications. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 6–17.

Harris, G. T., Rice, M. E., & Cormier, C. A. (1991). Psychopathy and violent recidivism. *Law and Human Behavior, 15,* 625–637.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315–335.

Hart, S. D., & Hare, R. D. (1989). The discriminant validity of the Psychopathy Checklist in a forensic psychiatric population. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 211–218.

Hart, S. D., Hare, R. D., & Forth, A. E. (1994). Psychopathy as a risk marker for violence: Development and validation of a Screening Version of the Revised Psychopathy Checklist. In J. Monahan & H. J. Steadman (Eds.), *Violence and mental disorder: Advances in risk assessment* (pp. 81–98). Chicago: University of Chicago.

Hart, S. D., Kropp, P. R., & Hare, R. D. (1988). The performance of male psychopaths following conditional release from prison. *Journal of Consulting and Clinical Psychology, 57,* 227–232.

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

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translation: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18,* 115–142.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983a). Applications of IRT to language translations. In *Item response theory: Application to psychological measurement* (pp. 185–209). Homewood, IL: Dow Jones-Irwin.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983b). Introduction to item response theory. In *Item response theory: Application to psychological measurement* (pp. 13–74). Homewood, IL: Dow Jones-Irwin.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E., & Surface, C. R. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 5,* 457–471.

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

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Newman, J. P., & Kosson, D. S. (1986). Passive avoidance learning in psychopathic and non-psychopathic offenders. *Journal of Abnormal Psychology, 95,* 252–256.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* McGraw-Hill.

Ogloff, J. R., Wong, S., & Greenwood, A. (1990). Treating adult psychopaths in a therapeutic community program within a correctional setting. *Behavioural Sciences and the Law, 8,* 81–90.

Paris, J. (1993). Personality disorders: A biopsychosocial model. *Journal of Personality Disorders, 7,* 255–264.

Poortinga, Y. H., & Van de Vijver, F. J. R. (1987). Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology, 18,* 259–282.

Raine, A. (1985). A psychometric assessment of Hare's checklist for psychopathy on an English prison population. *British Journal of Clinical Psychology, 24,* 247–258.

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.

Serin, R. C., Peters, R. D., & Barbaree, H. E. (1990). Predictors of psychopathy and release outcome in a criminal population. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 419–422.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artefacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Thissen, D. (1991). *Multilog (version 6)* [computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Wechsler, D. (1981). WAIS–R manual: Wechsler Adult Intelligence Scale—Revised. New York: Psychological Corporation.

Williamson, S., Harpur, T. J., & Hare, R. D. (1991). Abnormal processing of emotional words by psychopaths. *Psychophysiology, 28,* 260–273.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Received January 8, 1996
Revision received July 30, 1996
Accepted July 30, 1996 ■