

A Multigroup Item Response Theory Analysis of the Psychopathy Checklist—Revised

Daniel M. Bolt
University of Wisconsin—Madison

Robert D. Hare
University of British Columbia

Jennifer E. Vitale
Hampden-Sydney College

Joseph P. Newman
University of Wisconsin—Madison

Item response theory was used to investigate the functioning of the Psychopathy Checklist—Revised (PCL–R; R. D. Hare, 1991, 2003) in several offender populations. With male criminal offenders ($N = 3,847$) as a reference group, differential item functioning analyses were performed for 3 comparison groups: female criminal offenders ($N = 1,219$), male forensic psychiatric patients ($N = 1,246$), and male criminal offenders scored from file reviews ($N = 2,626$). Results are discussed in the context of the 2-factor, 4-facet model for the PCL–R (R. D. Hare, 2003; J. Parker, G. Sitarenios, & R. D. Hare, 2003). Application of a multigroup graded response model to all 4 groups suggests scalar equivalence may hold at least approximately for each population, although the PCL–R provided slightly greater information about the latent trait of psychopathy for male criminal offenders scored from the standard procedure.

Psychopathy is characterized by a set of affective, interpersonal, lifestyle, and socially deviant features, including egocentricity, deceptiveness, callousness, impulsivity, irresponsibility, lack of empathy or remorse, shallow emotions, poor behavioral controls, early behavior problems, and a proneness to antisocial behaviors (Cleckley, 1976; Hare, 1991). The most widely used measure of psychopathy is the Psychopathy Checklist—Revised (PCL–R; Hare, 1991, 2003), a 20-item instrument scored on the basis of interview and file information. Each item is scored as 0 (*not present*), 1 (*possibly present*), or 2 (*definitely present*), resulting in total PCL–R scores that range from 0 to 40. A cut score of 30 is commonly used to distinguish individuals with psychopathy from those without psychopathy for research purposes (Hare, 1991, 2003). The PCL–R has demonstrated good internal consistency, test–retest, and interrater reliability across diverse populations (see, e.g., Alterman, Cacciola, & Rutherford, 1993; Hare et al., 1990; Vitale, Smith, Brinkley, & Newman, 2002). Moreover, the validity of the PCL–R is well established, in both basic and applied settings. In prison populations, PCL–R scores have been shown to

predict violent behavior and recidivism, revocation of parole, and poor participation in and response to therapeutic interventions (see, e.g., Hare & McPherson, 1984; Hart, 1998; Hemphill, Templeman, Wong, & Hare, 1998; Salekin, Rogers, & Sewell, 1996), among other outcomes. In laboratory studies, PCL–R scores have also correlated with various language (Brinkley, Bernstein, & Newman, 1999; Hare, Williamson, & Harpur, 1988), emotional (Day & Wong, 1996; Lorenz & Newman, 2002; Williamson, Harpur, & Hare, 1991), and physiological (e.g., Arnett, Smith, & Newman, 1997; Ogloff & Wong, 1990; Patrick, Bradley, & Lang, 1993) deficits associated with psychopathy.

In the mental health and criminal justice systems, an assessment of psychopathy may have serious implications for individuals and society. For example, a PCL–R score often is used as an indicator of treatment amenability as well as a risk factor for recidivism and violence and may contribute to parole board decisions about early release from custody. It is therefore not surprising that the PCL–R has become the focus of intense psychometric scrutiny, especially regarding its generalizability across diverse populations. Most generalizability studies have focused on the structural equivalence of the instrument, as can be assessed with multigroup exploratory or confirmatory factor analyses (e.g., Brandt, Kennedy, Patrick, & Curtin, 1997; Cooke, 1995; Cooke, Kosson, & Michie, 2001; Hare et al., 1990; Kosson, Smith, & Newman, 1990; Parker, Sitarenios, & Hare, 2003; Salekin, Rogers, & Sewell, 1997; Windle & Dumenci, 1999). There remains, however, a pressing need to verify the generalizability of the PCL–R score metric (Cooke et al., 2001; Cooke & Michie, 1999). Scalar equivalence is said to hold when test scores represent the same levels of the construct across diverse populations (Van de Vijver & Leung, 1997). Factor structure equivalence does not imply scalar equivalence, as item mean scores are typically ignored in factor analysis. However, the issue of scalar equivalence should always be of great concern with

Daniel M. Bolt, Department of Educational Psychology, University of Wisconsin—Madison; Robert D. Hare, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada; Jennifer E. Vitale, Department of Psychology, Hampden-Sydney College; Joseph P. Newman, Department of Psychology, University of Wisconsin—Madison.

Robert D. Hare is the author of the Psychopathy Checklist—Revised (PCL–R), published by Multi-Health Systems, and receives royalties from its sale.

Correspondence concerning this article should be addressed to Daniel M. Bolt, Department of Educational Psychology, University of Wisconsin—Madison, 1025 West Johnson, Room 859, Madison, WI 53706. E-mail: dmbolt@facstaff.wisc.edu

psychological instruments, including the PCL–R. Hare (1991, 1998) has noted that scoring of some PCL–R items is likely to be influenced by societal or cultural factors. When the effects of such factors are accumulated across items, the meaning and implications of PCL–R total scores can become distorted. Cooke and Michie (1999), for example, reported that the lack of scalar equivalence across Scottish and North American criminal offenders was substantial enough to suggest that a cut score of 25 with Scottish offenders was equivalent to a cut score of 30 with North American offenders.

Item response theory (IRT) provides an appealing framework for studying scalar equivalence (Embretson & Reise, 2000). The attractiveness of IRT follows from its invariance properties. By modeling the trait–item score relationship, IRT can characterize differences in item and test functioning in a way that is not affected by differences in the trait distributions across the groups being compared (Embretson & Reise, 2000). When the trait–item score relationship is found to be different across populations, an item is said to exhibit differential item functioning (DIF). Essential to any DIF analysis is the identification of items that can be assumed to perform equivalently across populations. These items contribute to defining the metric of a latent trait (e.g., psychopathy) against which the remaining items can be studied for DIF. Once these

items are identified, an IRT-based DIF study of the PCL–R makes it possible to investigate whether individual psychopathy characteristics become manifest at different levels of psychopathy across populations. Items that display DIF not only are of questionable validity but also have the potential to contribute to bias in the total scores. By studying the cumulative effects of DIF across items, we can ascertain whether the same total PCL–R score may represent different latent levels of psychopathy across groups, implying a lack of scalar equivalence.

It is important to note that differentially functioning items need not imply a lack of scalar equivalence. For example, Cooke et al. (2001) compared Caucasian and African American criminal offenders and identified five PCL–R items that performed differentially across race (Items 3, 9, 13, 14, and 15; see Table 1 for a list of the PCL–R items). However, because the DIF observed across items occurred in opposing directions (some items resulted in higher scores for each group), it cancelled at the test score level—that is, the same total PCL–R scores were still expected at each level of psychopathy for Caucasian and African American criminal offenders. For this reason, Cooke et al. suggested the overall effect of DIF on total PCL–R scores was negligible.

A frequent limitation when one is performing DIF analyses with the PCL–R is its large demands in terms of sample size. In this

Table 1
Psychopathy Checklist—Revised Items and Descriptive Statistics Across Four Studied Populations

PCL–R item ^a	Facet	Factor	Population							
			Male off		Fem off		Psych		File	
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1. Glibness/superficial charm	1	1	0.84	0.71	0.71	0.75	0.58	0.70	0.45	0.67
2. Grandiose sense of self-worth	1	1	0.92	0.74	0.67	0.74	0.65	0.72	0.60	0.75
3. Need for stimulation/proneness to boredom	3	2	1.27	0.72	1.14	0.79	1.25	0.71	0.78	0.81
4. Pathological lying	1	1	0.99	0.71	0.84	0.73	0.90	0.69	0.53	0.71
5. Conning/manipulative	1	1	1.06	0.77	1.05	0.80	0.90	0.74	0.92	0.82
6. Lack of remorse or guilt	2	1	1.46	0.67	1.24	0.76	1.39	0.67	1.30	0.76
7. Shallow affect	2	1	0.97	0.74	0.86	0.76	1.13	0.69	0.61	0.74
8. Callous/lack of empathy	2	1	1.19	0.72	0.92	0.79	1.12	0.68	1.00	0.81
9. Parasitic lifestyle	3	2	1.08	0.67	1.13	0.76	1.08	0.72	0.75	0.77
10. Poor behavioral controls	4	2	1.21	0.78	0.94	0.84	1.39	0.68	0.98	0.81
11. Promiscuous sexual behavior			1.16	0.81	0.94	0.89	1.10	0.86	1.01	0.88
12. Early behavior problems	4	2	0.94	0.83	0.54	0.76	1.06	0.84	0.51	0.76
13. Lack of realistic, long-term goals	3	2	1.20	0.74	0.95	0.80	1.03	0.72	0.83	0.81
14. Impulsivity	3	2	1.30	0.69	1.19	0.78	1.46	0.62	1.06	0.77
15. Irresponsibility	3	2	1.36	0.68	1.38	0.72	1.26	0.63	1.08	0.76
16. Failure to accept responsibility	2	1	1.32	0.72	1.05	0.78	1.28	0.72	1.27	0.78
17. Many short-term marital relationships			0.75	0.84	0.76	0.84	0.44	0.73	0.27	0.61
18. Juvenile delinquency	4	2	1.18	0.86	0.58	0.77	1.04	0.85	0.71	0.83
19. Revocation of conditional release	4	2	1.48	0.79	1.25	0.87	1.27	0.85	0.96	0.90
20. Criminal versatility	4	2	1.26	0.79	0.94	0.82	1.08	0.83	0.78	0.86

Note. PCL–R = Psychopathy Checklist—Revised; Male off = male criminal offenders ($N = 3,847$); Fem off = female criminal offenders ($N = 1,219$); Psych = male forensic psychiatric patients ($N = 1,246$); File = male criminal offenders scored only from file review ($N = 2,626$).

^a Copyright © 1990, 1991, Robert D. Hare, Ph.D. and Multi-Health Systems Inc. All rights reserved. In the USA, P.O. Box 950, North Tonawanda, NY 14120-0950, 1-800-456-3003. In Canada, 3770 Victoria Park Avenue, Toronto, Ontario M2H 3M6, 1-800-268-6011. Internationally, +1-416-492-2627. Fax, +1-416-492-3343. Reproduced with permission.

article, we report on an IRT analysis with larger samples than any previously reported IRT analyses of the PCL-R. The large samples permitted assessment of differences across groups that have typically been too small to study with adequate power. Because DIF in the PCL-R based on race has been previously investigated by Cooke et al. (2001), we focus on other characteristics in this analysis.

Gender

Both gender differences and similarities have been observed in psychopathy research. On the basis of PCL-R scores, female offenders appear to display lower base rates of psychopathy than do male offenders. For example, Rutherford, Cacciola, Alterman, and McKay (1996) failed to find any women in their study with PCL-R scores above 30. A review of research studies of psychopathy in female populations conducted by Vitale et al. (2002) generally found base rates of psychopathy ranging from 9% to 23% for women and from 15% to 30% for men. In the second edition of the PCL-R manual (Hare, 2003), approximately 15% of male offenders and 7.5% of female offenders had a PCL-R score above 30. Gender differences have also been observed in several external correlates of the PCL-R scores, including anxiety, negative affectivity, and intelligence (Vitale et al., 2002). Conversely, the factor structure of the PCL-R appears to be much the same for female and male offenders (Hare, 2003; Parker et al., 2003; Warren et al., 2003). Furthermore, there are similarities between female and male offenders in some criminal correlates (Loucks & Zamble, 2000; Richards, Casey, & Lucente, 2003; Vitale et al., 2002), associations with other personality disorders (Rutherford et al., 1996; Salekin et al., 1997; Warren et al., 2003), and response to treatment (Richards et al., 2003). In this analysis, we attempted to clarify whether observed gender differences might be due to gender-related bias in some PCL-R items.

Forensic Psychiatric Versus Criminal Offenders

The PCL-R is currently used extensively in forensic psychiatric facilities as well as prisons. Approximately 10% of male forensic psychiatric offenders described by Hare (2003) have a PCL-R score of 30 or higher. Previous analyses of the instrument across settings have supported a generalizable factor structure (Hare, 2003; Hare et al., 1990; Parker et al., 2003). However, it has been argued that individual PCL-R characteristics will be observed in differing degrees across settings. For example, several of the hallmark features of psychopathy are also defining characteristics of other disorders. Hart & Hare (1989) noted that "psychopaths and psychotics exhibit some of the same symptoms and behaviors, including impulsivity, shallow affect, seemingly illogical antisocial behavior, and chronic disturbances in interpersonal functioning" (p. 211). At the same time, other PCL-R items (e.g., Item 19, "Revocation of conditional release"; Item 20, "Criminal versatility") are linked to the violation of court-imposed rules and a wide range of criminal activities and thus may be disproportionately prevalent among criminal offenders. Taken together, it seems likely that there will be some DIF when criminal offenders and forensic psychiatric patients are compared, although its implica-

tions for scalar equivalence and the potential need for different diagnostic cut scores are less clear.

File-Only Versus Standard Administration

It is not always possible to conduct the interview portion of a standard PCL-R administration. Consequently, the PCL-R is sometimes scored using only file reviews (Hare, 2003; Harris, Rice, & Cormier, 1991). Although PCL-R scores from file reviews generally correlate highly with scores based on both file reviews and interviews, the file-only scores are often lower than those obtained with the standard procedure (Grann, Langstrom, Tengstrom, & Stalenheim, 1998; Hare, 2003; Wong, 1988). For example, only approximately 6% of file reviews described by Hare (2003) have a PCL-R score of 30 or higher. These findings may be due to the lack of information needed to confidently score certain PCL-R characteristics. Wong (1988) and Hare (2003) have suggested that the interpersonal and affective items, which rely largely on the integration of interviewer impressions and collateral information, may be particularly difficult to evaluate without the interview. In this study, we sought to identify items that show DIF when scored without an interview and also to determine whether these differences support establishment of alternative cut scores when the PCL-R is scored from files.

Method

Description of Data

PCL-R item ratings for a total of 8,938 offenders were analyzed. The complete data set, reported in the second edition of the PCL-R manual (Hare, 2003), consists of 14 samples, including 7 from the original 1991 sample (Hare, 1991) that were previously used by Cooke and Michie (1997) in an IRT analysis of the PCL-R. For the current analysis, the pool of offenders was separated into four new samples: (a) male criminal offenders ($N = 3,847$), (b) female criminal offenders ($N = 1,219$), (c) male forensic psychiatric patients ($N = 1,246$), and (d) male criminal offenders for whom the PCL-R was scored only from file data ($N = 2,626$). For all respondents, the PCL-R assessments were performed by trained raters and, with the exception of the last sample, were based on both interviews and file reviews.

Table 1 lists the PCL-R items and item statistics for each sample. Also reported is a categorization of the items based on recent large-sample, multigroup analyses that yielded a two-factor, four-facet hierarchical model of the PCL-R (Hare, 2003; Parker et al., 2003). The model is based on 18 of the 20 PCL-R items categorized according to four psychopathy facets, labeled as follows: *Facet 1, Interpersonal* (Items 1, 2, 4, and 5); *Facet 2, Affective* (Items 6, 7, 8, and 16); *Facet 3, Lifestyle* (Items 3, 9, 13, 14, and 15); and *Facet 4, Antisocial* (Items 10, 12, 18, 19, and 20). The four facets load onto two higher order factors: Factor 1, Interpersonal/Affective and Factor 2, Lifestyle/Antisocial. Factors 1 and 2 are identical to the original PCL-R factors (Hare, 1991; Hare et al., 1990), except that Item 20 ("Criminal versatility") now loads on Factor 2. Facets 1, 2, and 3 are identical to, respectively, Factors 1, 2, and 3 identified by Cooke and Michie (2001) in their factor analysis of a selected set of 13 PCL-R items. Facet 4 includes 5 items that were excluded by Cooke and Michie (2001) but that clearly form a fourth factor, or facet.

When one is conducting traditional IRT analyses, unidimensionality is assumed, implying that only one latent dimension underlies the data. Despite some debate over the appropriate factor structure for the PCL-R, all extant models describe highly intercorrelated factors, implying the

strong influence of a single underlying general factor (i.e., psychopathy). Nevertheless, we expected that some multidimensionality would be present in the current data sets. We used several criteria to examine the amount of multidimensionality in the data from each group, including (a) the eigenvalues of the reduced polychoric correlation matrix, (b) the fit of a single-factor model to the polychoric correlation matrix, and (c) the general factor saturation (GFS) estimate (Zinbarg, Barlow, & Brown, 1997). The eigenvalues of the reduced polychoric correlation matrix were 4.807, 0.855, 0.312, and 0.193 for male criminal offenders; 4.508, 1.308, 0.397, and 0.225 for female criminal offenders; 4.429, 1.194, 0.926, and 0.336 for forensic psychiatric patients; and 6.123, 1.588, 0.546, and 0.505 for file reviews. The ratio of the first-to-second eigenvalues, a common criterion for assessment of unidimensionality, ranged from 5.6:1 for male criminal offenders to 3.4:1 for female criminal offenders. These ratios are similar to those observed in previous studies (e.g., Cooke & Michie, 1997) and although supportive of a strong first dimension, are clearly indicative of some multidimensionality in the data.

However, despite this multidimensionality, a single-factor model still fit the four polychoric correlation matrices quite well. In each group, a single-factor model was fit using the weighted least squares estimation procedure in LISREL 8 (Jöreskog & Sörbom, 1996), and goodness of fit was assessed with the minimum-fit-function chi-square test, comparative fit index (CFI; Bentler, 1990), Tucker–Lewis index (TLI; Tucker & Lewis, 1973), and root-mean-square error of approximation index (RMSEA; Steiger, 1990). Results suggest a decent fit for all four groups: male criminal offenders, $\chi^2(170, N = 3,847) = 2,382.9, p < .01, CFI = .94, TLI = .93, RMSEA = .065$; female criminal offenders, $\chi^2(170, N = 1,219) = 1,183.3, p < .01, CFI = .92, TLI = .91, RMSEA = .074$; forensic psychiatric patients, $\chi^2(170, N = 1,246) = 1,330.0, p < .01, CFI = .91, TLI = .90, RMSEA = .084$; and file reviews, $\chi^2(170, N = 2,626) = 1,948.2, p < .01, CFI = .95, TLI = .95, RMSEA = .077$. Thus, it appears that a single factor was relatively successful in accounting for the interitem polychoric correlations.

Finally, the GFS estimate was used to quantify the influence of the single underlying factor. The GFS represents the proportion of total score variance that can be attributed to the factor. To estimate the GFS, the two-factor, four-facet hierarchical model was fit, but now with an additional third-order factor added to account for the intercorrelation between the two second-order factors. Note that relative to the previous factor analysis, the current model was fit to the interitem covariance matrix, because the item variances are needed to quantify the total test variance, and the polychoric matrix removes information about the variances of the items. Results suggest that the model provided a reasonable fit within the four samples: male criminal offenders, $\chi^2(132, N = 3,847) = 2,225.0, p < .01, CFI = .87, TLI = .84, RMSEA = .066$; female criminal offenders, $\chi^2(132, N = 1,219) = 853.4, p < .01, CFI = .85, TLI = .83, RMSEA = .070$; forensic psychiatric patients, $\chi^2(132, N = 1,246) = 1,207.8, p < .01, CFI = .79, TLI = .76, RMSEA = .085$; and file reviews, $\chi^2(132, N = 2,626) = 2,577.0, p < .01, CFI = .84, TLI = .83, RMSEA = .089$. Cooke et al. (2001) have suggested that GFS estimates above .50 indicate sufficient unidimensionality for IRT. For the current data, the GFS estimates ranged from .56 (for female offenders, standard administration) to .71 (for male offenders, standard administration) across the four samples. When the two-factor, four-facet model was modified to allow the two excluded items (Item 11, “Promiscuous sexual behavior”; Item 17, “Many short-term marital relationships”) to load directly on the general factor, the GFS estimates for each sample stayed approximately the same. Consequently, all 20 PCL–R items were deemed to measure sufficiently a single underlying factor such that IRT modeling would be informative for each sample.

It should be noted that despite the statistical confirmation of some multidimensionality in each group, there are several compelling reasons in the current study for use of a unidimensional IRT model to study DIF. First, unidimensionality in its strictest sense is usually not necessary to

realize the benefits of IRT (Smith & Reise, 1998), provided a dominant first dimension is present, as appears to be the case for the current data sets. Second, relative to other IRT applications, the implications of multidimensionality on DIF analyses are better understood (e.g., Drasgow, 1987). Specifically, multidimensionality has the potential to produce DIF when a unidimensional IRT model is applied; that is, an item that actually performs the same across groups may appear to perform differentially if the groups have different distributions on the secondary factor. In this respect, being aware of the secondary dimensions need not hinder the DIF analysis but can contribute to providing an interpretation of DIF when it occurs. Finally and most important, because most practical use of the PCL–R involves a single total score, it was considered more informative to study item performance against a single latent trait rather than against multiple traits.

IRT and the Graded Response Model

IRT is a statistical framework for modeling items scored using discrete categories. In IRT, item scores are modeled as a function of a unidimensional latent trait, denoted θ (e.g., level of psychopathy), that is assumed to underlie scores on the items. Cooke and Michie (1997, 1999) found the graded response model (GRM; Samejima, 1969) to be suitable for the PCL–R. For items scored in three categories, the GRM characterizes each item according to three parameters: b_1 and b_2 are item threshold parameters, and a is an item discrimination parameter. The b_1 and b_2 parameters relate to the θ levels needed before scores of 1 and 2 tend to be observed on the item. The a parameter indicates the degree to which the item score is influenced by the latent trait. Items with large positive a s (e.g., > 1.0) discriminate well with respect to θ .

Figure 1A illustrates cumulative score probability curves for three PCL–R items on the basis of their GRM estimates for male criminal offenders: Item 1 “Glibness/superficial charm” ($\hat{a} = 1.19, \hat{b}_1 = -0.65, \hat{b}_2 = 1.46$); Item 3, “Need for stimulation/proneness to boredom” ($\hat{a} = 1.30, \hat{b}_1 = -1.43, \hat{b}_2 = 0.24$); and Item 17, “Many short-term marital relationships” ($\hat{a} = 0.65, \hat{b}_1 = 0.05, \hat{b}_2 = 1.76$). The corresponding option characteristic curves, indicating the probability of scoring in each category, are shown in Figure 1B. The meaning of GRM estimates is perhaps best conveyed through a comparison of these curves across items. For example, Item 3, having lower b_1 and b_2 estimates, achieves higher scores at lower levels of the trait. Item 3 thus represents a PCL–R characteristic that is more commonly observed among individuals with low psychopathy levels than does Items 1 or 17. Item 17 has a lower a parameter than Items 1 and 3, implying its probability curves will be less concentrated along the θ scale. Practically speaking, Item 17 would be a less useful indicator of psychopathy than would Item 1 or 3.

GRM items can also be illustrated according to their *expected response functions* (ERFs). An ERF represents the expected item score as a function of θ and is computed as the sum of the item score categories weighted by their probabilities. The ERFs for Items 1, 3, and 17 are shown in Figure 1C. As in Figures 1A and 1B, the ERFs also make apparent that Item 3 represents a psychopathy characteristic more commonly observed at low psychopathy levels, and that Item 17 is a less discriminating PCL–R characteristic. Generally, ERFs are easier to inspect than the option characteristic curves, as there exists only one curve per item.

Comparison of ERFs across populations also provides a useful way of interpreting DIF. When one is fitting an IRT model, DIF is technically said to exist when an item has different parameter values (e.g., different a s, b_1 s, or b_2 s) across groups. However, when DIF exists, the ERFs will also necessarily differ (Chang & Mazzeo, 1994). ERFs provide an attractive way of evaluating the implications of DIF as it is often only the expected scores that are of interest. Moreover, as noted below, it is typically the ERFs that provide the basis for quantifying the amount of DIF in an item.

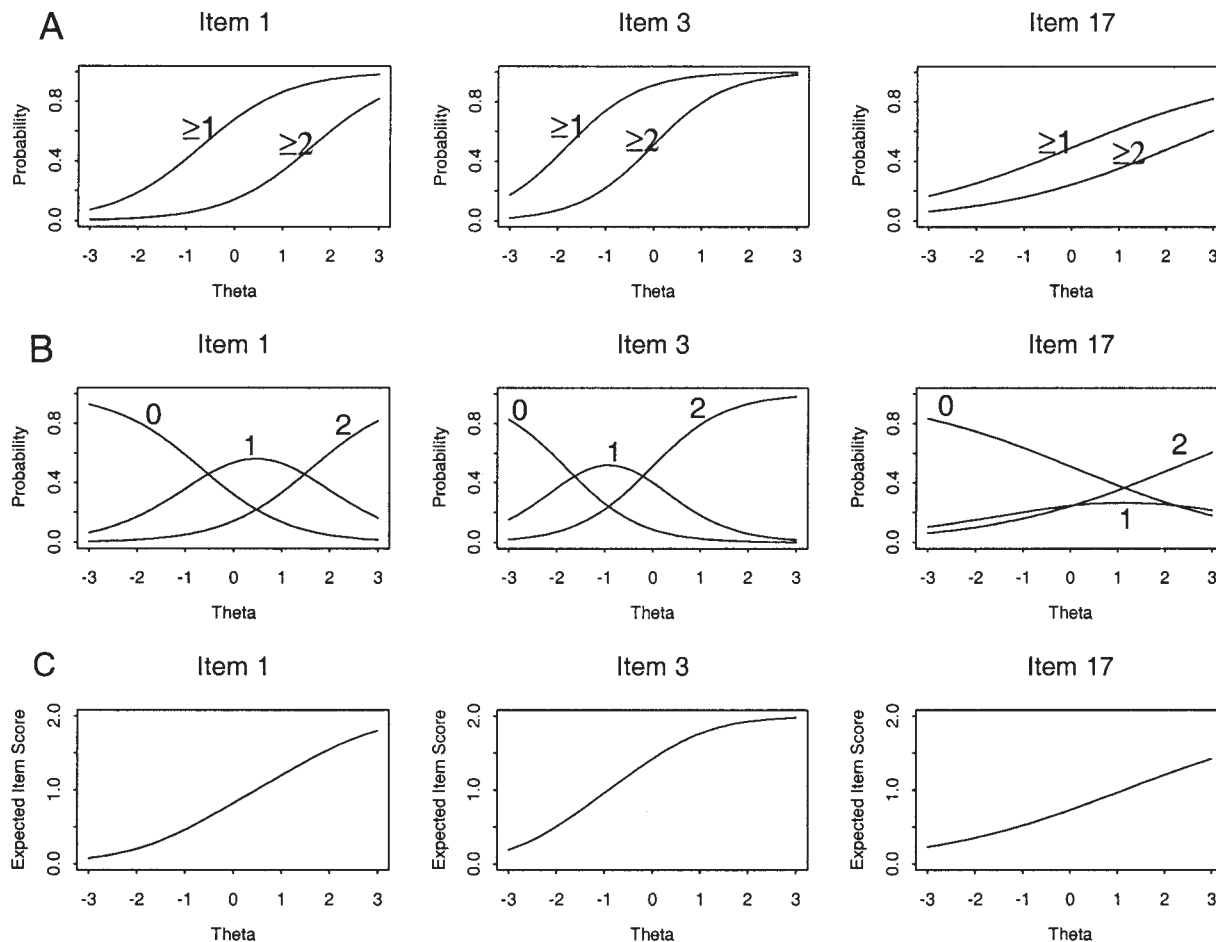


Figure 1. Illustration of estimated item category characteristic curves and expected response functions for three Psychopathy Checklist—Revised items. A: Estimated cumulative score probability curves. B: Estimated option characteristic curves. C: Estimated expected response functions.

Assessing GRM Fit

Due in part to its statistical properties (Samejima, 1997), the GRM is one of the most suitable of polytomous item response models for modeling item scores based on ordered score ratings. To evaluate the appropriateness of the GRM for the data considered in this study, we randomly selected two nonoverlapping samples of 3,000 respondents from the complete data set of 8,938 respondents, and we inspected several goodness-of-fit indices. The first sample of 3,000 was used to obtain GRM item parameter estimates; the second sample was used as a cross-validation sample to evaluate the fit of the GRM. Three types of chi-square indices (Drasgow, Levine, Tsien, Williams, & Mead, 1995) were inspected with the program MODFIT (Stark, 2001). These indices evaluate the fit of the GRM with respect to the first-order, second-order, and third-order joint frequencies of the item scores, respectively (for details, see Drasgow et al., 1995). Essentially the first-order indices evaluate whether the model-implied score probabilities given the trait are consistent with the empirical probabilities observed for individual items; the second- and third-order tests are also sensitive to local dependence among scores across item pairs and item triples. Drasgow et al. (1995) suggested that model fit is good provided the adjusted chi-square divided by degrees of freedom is 3 or less. In the current analysis, the average adjusted chi-square index for the individual items was 1.731, with 16 of the 20 items being in the range of good fit (<3). All items except

Item 18 “Juvenile delinquency” had chi-square values below 5. (The larger misfit of Item 18, adjusted $\chi^2/df = 9.10$, can likely be attributed to its large DIF for female offenders, as demonstrated later.) Thus, these results appear to support use of the GRM in the current application. For the item doubles and triples, the average adjusted chi-square indices were 3.721 and 3.639, respectively, with only approximately half of the pairs–triples being in the range of good fit. Consistent with the previous factor analyses, these results possibly point to the presence of some local dependence among item pairs and triples. The likely explanation for local dependence is the known multidimensionality in the data. However, as noted in the factor analysis of these data, the local dependence does not appear to be so large as to contradict the presence of a single pervasive underlying factor. Thus, for the reasons considered earlier, we still consider a unidimensional GRM useful in studying group differences.

DIF Using the Likelihood Ratio Test

Assuming data have been collected from more than one group, the likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986) statistically tests for DIF by comparing the statistical fit of two models: one in which the parameters of the item being studied for DIF are set equal across groups (a *compact model*) and one in which the parameters are free to vary (an

augmented model). A separate log-likelihood statistic can be computed for each model. When the models are compared, the difference between -2 times the log-likelihood for each model follows a chi-square distribution under a null hypothesis of no DIF in the studied item.

In the current study, a separate DIF analysis was conducted for each of the three comparison samples, always with male criminal offenders as a reference group. For each DIF analysis, an iterative purification procedure (Lord, 1980, p. 220) was first used to identify anchor items (see also Cooke & Michie, 1999). Anchor items are items that fail to display DIF and thus have their item parameters constrained to equality across groups so as to define a common latent metric against which the remaining items can be evaluated for DIF. The iterative purification process starts by testing each item for DIF while using all of the remaining items as anchors. On the basis of the results of this analysis, a core collection of items that display no DIF is chosen as the initial anchor. The remaining items are then tested for DIF against this anchor; items are iteratively added to or subtracted from the anchor on the basis of whether they display DIF. The process terminates when all items on the anchor fail to display DIF and no additional items can be added without introducing DIF. In the current study, DIF testing was performed with respect to all three GRM item parameters (a , b_1 , b_2) resulting in chi-square tests having three degrees of freedom. We used the computer program, IRTLRFID (Thissen, 2001), both to apply the iterative purification procedure to identify anchor items and to conduct subsequent LR tests for the nonanchor items. In this program, the metric indeterminacy of the latent trait is resolved by setting the mean and variance of the latent trait to 0 and 1, respectively, for the reference group.

Quantification of DIF

Cohen, Kim and Baker (1993) and Wainer (1993) discussed indices based on either the signed or the unsigned vertical distances between ERFs as a means of quantifying DIF. Because the θ metric is continuous, in the current analysis these indices were estimated through a discrete approximation in which the vertical difference between ERFs was computed at θ nodes separated at 0.1 intervals from -3.0 to 3.0 . The difference between ERFs across nodes is then averaged using the distribution of the latent trait

at each node in the comparison sample as a weight. For the average unsigned ERF difference (AUD), the absolute value of the difference is used, whereas for the average signed ERF difference (ASD), the signed difference is used. The ASD differs from the AUD in that it is affected by changes in the direction of DIF across levels of the latent trait. In other words, the magnitude of the ASD can be less than the AUD provided the differences between ERFs cancel across trait levels. When this occurs, the DIF is referred to as *nonuniform DIF* (see, e.g., Van de Vijver & Leung, 1997).

Figure 2 provides an illustration of the ERFs for male offenders and female offenders with respect to Item 9 (“Parasitic lifestyle”) and Item 10 (“Poor behavioral controls”). Item 9 results in higher expected scores for female offenders than male offenders at low levels of the latent trait but produces lower expected scores for female offenders at high levels of the latent trait. This item displays nonuniform DIF and will produce an AUD index that is of greater magnitude than the ASD index. By contrast, Item 10 results in higher expected scores for male offenders across nearly all levels of the latent trait, implying primarily uniform DIF and an ASD index that will be of approximately the same magnitude as the AUD index.

In interpreting the magnitude of the AUD and ASD indices, we adopted criteria based on previous recommendations (Roussos & Stout, 1995). For three category items, an AUD or ASD with magnitude greater than .20 corresponds to a large amount of DIF, .10–.20 a moderate amount of DIF, and less than .10 a small amount of DIF. Both the statistical significance of DIF and its magnitude contributed to determining which items should function as anchors in the previously described purification process. Items were assigned to the anchor if they either (a) contained a statistically insignificant amount of DIF or (b) had an AUD less than .05, implying a very small amount of DIF.

A Multigroup GRM

A limitation of performing separate DIF analyses for the three comparison samples is that the resulting GRM parameter estimates are not on a common metric. Therefore, to better evaluate the implications of DIF, we fit a multigroup GRM in which the item parameters of all four samples

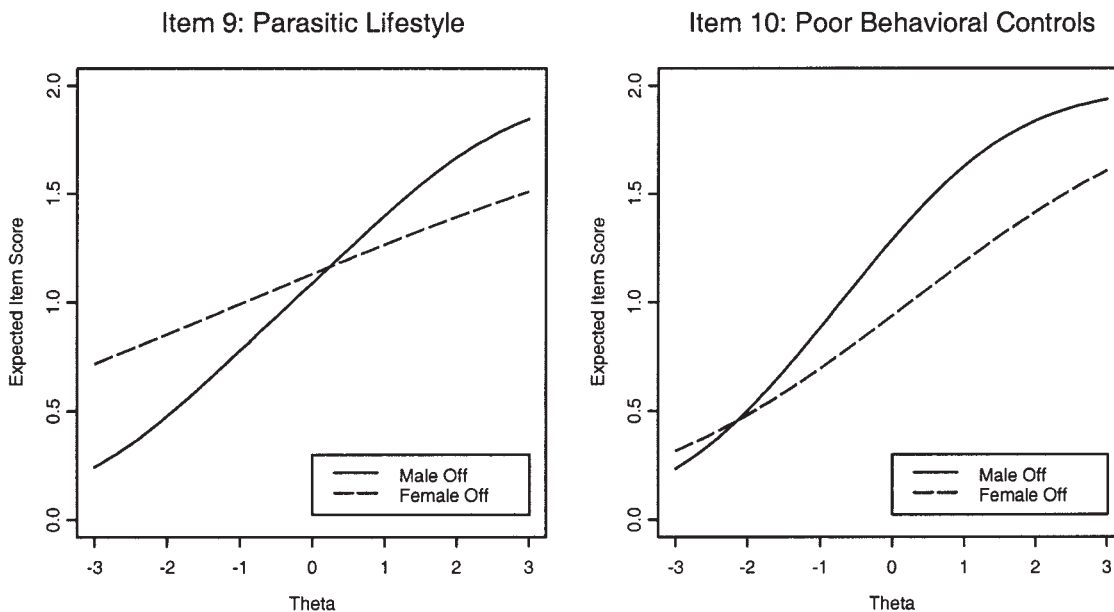


Figure 2. Illustration of items displaying nonuniform and uniform differential item functioning. Male Off = male criminal offenders; Female Off = female criminal offenders.

were estimated simultaneously. A multigroup GRM allows for separate GRM estimates for each sample but through anchor items connects the estimates of all four samples to a common latent metric. In this analysis, we used the items found to function as anchors in the DIF analyses as anchors in the multigroup IRT model, and thus constrained their parameters to be equal to the corresponding parameters in the reference sample (male criminal offenders). Parameters for all DIF items were left unconstrained. The program MULTILOG (Thissen, 1991) was used to fit the resulting model. In MULTILOG, a mean of the latent trait (i.e., a mean latent psychopathy level) is also estimated for each sample. The standard deviation of the latent trait for each group is set to 1.0. All parameter estimates are reported with respect to the logistic metric of the GRM.

There are several ways in which the multigroup model estimates can be used to compare groups. First, and most important, is that the expected total PCL-R scores at each trait level can be compared. In IRT, the relationship between the latent trait and expected total score is illustrated by a test characteristic curve and is computed as the simple sum of the ERFs at each trait level. Differential test functioning exists when the test characteristic curves differ across groups. Finding differential test functioning implies a lack of scalar equivalence.

Second, the distributions of the latent trait across groups can be compared (Meredith, 1993). When substantial differential test functioning exists, the groups cannot be compared with respect to test score distributions because test scores fail to indicate equivalent levels of the trait. However, the groups can be compared with respect to the latent metric because this is defined by items that function equivalently. To test for differences in the latent mean estimates, an LR test can be applied in the same way as when used to test an item for DIF.

Finally, the measurement precision provided by PCL-R items can be compared across groups. DIF may contribute not only to differential test functioning, but also to variability in the degree of accuracy with which the

latent trait is measured for each group. In IRT, measurement precision is quantified through information functions (see, e.g., Cooke & Michie, 1997). An item information function indicates the accuracy with which the latent trait can be estimated from the item score and typically varies across trait levels (Embretson, 1996). Computationally, it is the square of the reciprocal of the standard error of the trait estimate based on the item score (see Baker, 1992, p. 246, for the information function formula for the GRM). The sum of item information functions produces a test information function, which indicates the overall amount of precision collectively provided by all item scores. By comparing information functions across groups, we thus gain insight into another potential consequence of DIF.

Results

DIF

Table 2 reports results from the DIF analysis conducted for each comparison sample. For each item found to display significant DIF, the LR chi-square statistic, the ASD index, and the AUD index are reported. Items for which no indices are reported served as anchor items. For the ASD, a positive value indicates a higher expected item score (conditional on trait level) for the comparison group, whereas a negative value indicates a higher expected score for the reference group (male criminal offenders, standard administration). For example, Item 18 ("Juvenile delinquency") has an ASD of $-.510$ for female offenders, implying that for male and female offenders having the same latent levels of psychopathy, female offenders receive, on average, ratings $.510$ lower than male offenders. It is interesting to note that the absolute magnitudes of

Table 2
Differential Item Functioning Statistics for Female Offender, Forensic Psychiatric, and File Review Samples Compared With Male Criminal Offenders, File and Interview

Item	LR χ^2			Average signed ERF difference			Average unsigned ERF difference		
	Female	Psych	File	Female	Psych	File	Female	Psych	File
1		91.2	179.3		-.167	-.174		.167	.174
2		74.9	94.7		-.170	-.066		.170	.082
3									
4			241.5			-.193			.193
5	122.8		145.1	.202		.215	.222		.239
6			84.2			.188			.193
7		140.7	106.7		.248	-.052		248	.112
8			219.3			.223			.241
9	130.3		65.8	.151		.017	.160		.072
10	71.8	145.7		-.108	.268		.143	.268	
11	90.0	43.4	165.8	-.078	.011	.168	.102	.080	.245
12	121.5	66.6		-.269	.215		.269	.215	
13	38.5	35.5		-.113	-.107		.113	.107	
14	40.3	129.0	145.6	.011	.226	.219	.052	.226	.219
15	86.0			.146			.157		
16			82.1			.148			.148
17	41.4	120.2	287.8	.104	-.246	-.252	.168	.247	.254
18	322.4			-.510			.510		
19	50.5	45.6	80.3	-.147	-.173	-.211	.164	.173	.211
20	82.9		102.2	-.202		-.127	.210		.127

Note. All chi-square tests had 3 degrees of freedom. LR = likelihood ratio; ERF = expected response function; Female = female criminal offenders ($N = 1,219$); Psych = male forensic psychiatric patients ($N = 1,246$); File = male criminal offenders scored only from file review ($N = 2,626$).

the ASD and AUD indices are very similar across all DIF items. This suggests that when DIF occurs, it generally occurs uniformly (i.e., in the same direction across all levels of the unidimensional trait). Thus, the current interpretation of DIF results will focus primarily on the ASD index.

Not surprisingly, there is considerable variability in the direction and magnitude of DIF for individual items across comparison samples. For female offenders, large differences from the reference group are observed for Items 5 (“Conning/manipulative”), 12 (“Early behavior problems”), 18 (“Juvenile delinquency”), and 20 (“Criminal versatility”), with female offenders achieving higher scores on Item 5 and lower scores on Items 12, 18, and 20.

For forensic psychiatric patients, large amounts of DIF were observed for Items 7 (“Shallow affect”), 10 (“Poor behavioral controls”), 12 (“Early behavior problems”), 14 (“Impulsivity”), and 17 (“Many short-term marital relationships”), with the patients achieving higher scores than the reference group for Items 7, 10, 12, and 14 but lower scores for Item 17. Items 1, 2, and 19 also produced substantial amounts of DIF, all resulting in lower scores for forensic psychiatric patients. The findings for Items 7, 10, and 14, in particular, appear to be consistent with Hart and Hare’s (1989) prediction that such characteristics are likely to be present in forensic psychiatric patients because of similar symptomology. As anticipated, some cancellation of DIF appears to occur through items such as Item 19, which is more closely tied to criminal offenses.

Of all three comparison samples, the largest number of DIF items was observed for the file reviews. Items 5 (“Conning/manipulative”), 8 (“Callous/lack of empathy”) and 14 (“Impulsivity”) produced large DIF with higher scores for file reviews, whereas Items 17 (“Many short-term marital relationships”) and 19 (“Revocation of conditional release”) displayed the largest amounts of DIF in the reverse direction. Relative to the findings for female offenders and forensic psychiatric patients, the causes of DIF for file reviews were generally more difficult to interpret. In part this may be due to the possibility that the file review sample differed from the reference group in ways other than the method of scoring (note that there was not random assignment to the PCL–R scoring conditions).

Some general observations can be made with respect to the types of items that tended to display DIF, based on the two-factor, four-facet model of the PCL–R. For example, among female offenders, Factor 1 items (according to the two-factor model) generally displayed smaller amounts of DIF than the non-Factor 1 items. Facet 4 (Antisocial) items, in particular, consistently produced lower scores for female offenders. These general findings are consistent with those of Cooke et al., (2001), who failed to find DIF in any of the Factor 1 items in a DIF study examining race differences. Recalling that DIF is itself a validity criterion, finding less DIF in Factor 1 items appears to support the frequent claim that Factor 1 items may lie closer to the core of the psychopathy disorder (Hare, 2003), as they appear to be less affected by other characteristics associated with gender or race.

For file reviews, Factor 1 items were more prone to display DIF. All eight of the Facet 1 (Interpersonal) and Facet 2 (Affective) items were found to display statistically detectable DIF. Such findings are consistent with Wong’s (1988) and Hare’s (2003) suggestion that scoring these items would be more difficult to rate

without an interview. It is interesting, however, that these items did not display DIF in a consistent direction. Items 1, 2, 4, and 7 resulted in lower scores for file reviews, whereas Items 5, 6, 8, and 16 resulted in higher scores for file reviews. Thus it appears that the absence of information needed to score these items resulted in some that were overestimated, whereas others were underestimated. It is interesting that the direction of DIF is also associated with the mean score of the items in the reference population observed in Table 1. Specifically, Items 5, 6, 8, and 16 are the Factor 1 items with the highest means among male criminal offenders, standard administration, while Items 1, 2, 4, and 7 have the lower means. Thus, one explanation for this finding may be that in the absence of information needed to score Factor 1 items, raters rely more heavily on the known marginal base rates of the characteristics in rating Factor 1 items for the file reviews.

Multigroup GRM Analysis

Table 3 reports estimates from the multigroup GRM. Estimates in bold represent items for which parameters were left unconstrained in the multigroup model (due to their display of DIF), and items not in bold represent anchor items. For standard administration of the PCL–R with male criminal offenders, the GRM estimates were largely consistent with those observed in previous IRT analyses of the PCL–R (Cooke & Michie, 1997, 1999). Factor 1 items (Facets 1 and 2) tended to have slightly higher discrimination estimates and often higher threshold estimates than Factor 2 (Facets 3 and 4) items. Both item features contribute to making Factor 1 items more useful than Factor 2 items in identifying the prototypical individual with psychopathy, a finding that also applies to the three comparison groups despite the presence of DIF in some items.

The item parameter estimates for the multigroup model corresponded quite closely to the results observed from the DIF analyses. For example, Item 18, observed in the DIF analysis to produce lower scores for female compared with male offenders, also had higher threshold estimates for female compared with male offenders. It is more interesting that the majority of items found to display DIF produced lower discrimination estimates in the comparison sample. For example, among female offenders, all of the DIF items resulted in lower discrimination estimates. The discrimination estimates for file reviews were not consistently lower; however, two affective items, Item 7 (“Shallow affect”) and Item 8 (“Callous/lack of empathy”), showed a substantial decline in discrimination, perhaps reflecting greater difficulty in evaluating these items relative to other Factor 1 items when only file review was used.

Differential Test Functioning

Figure 3 illustrates the test characteristic curves for each of the three comparison groups against the test characteristic curve for the reference group (male criminal offenders, standard administration). To evaluate whether the difference between curves was significant, we used a chi-square test based on the differential functioning of items and tests (DFIT) procedure (Flowers, Oshima, & Raju, 1999). The DFIT test statistic is computed as a weighted difference between curves in which the estimated trait distribution

Table 3
Graded Response Model Item Parameter Estimates, Multigroup Analysis

Item	Male off			Fem off			Psych			File		
	<i>a</i>	<i>b</i> ₁	<i>b</i> ₂	<i>a</i>	<i>b</i> ₁	<i>b</i> ₂	<i>a</i>	<i>b</i> ₁	<i>b</i> ₂	<i>a</i>	<i>b</i> ₁	<i>b</i> ₂
1	1.19	-0.65	1.46	1.19	-0.65	1.46	1.03	-0.01	2.01	1.18	-0.27	1.33
2	1.20	-0.71	1.24	1.20	-0.71	1.24	1.06	-0.21	1.76	1.27	-0.70	0.76
3	1.30	-1.43	0.24	1.30	-1.43	0.24	1.30	-1.43	0.24	1.30	-1.43	0.24
4	1.26	-1.07	1.13	1.26	-1.07	1.13	1.26	-1.07	1.13	1.21	-0.54	1.02
5	1.41	-0.90	0.77	0.89	-1.63	0.39	1.41	-0.90	0.77	1.20	-1.50	-0.05
6	1.58	-1.83	-0.18	1.58	-1.83	-0.18	1.58	-1.83	-0.18	1.58	-2.32	-0.89
7	1.23	-0.95	1.04	1.23	-0.95	1.04	1.04	-1.90	0.72	0.79	-0.72	1.43
8	1.84	-1.18	0.47	1.84	-1.18	0.47	1.84	-1.18	0.47	1.48	-1.64	-0.26
9	0.96	-1.71	1.14	0.64	-2.43	0.55	0.96	-1.71	1.14	1.05	-1.21	0.65
10	1.02	-1.63	0.20	0.87	-1.10	0.55	1.08	-2.43	-0.20	1.02	-1.63	0.20
11	0.92	-1.29	0.41	0.84	-0.85	0.29	0.54	-1.59	0.42	0.68	-1.74	-0.28
12	1.16	-0.42	0.82	0.94	0.16	1.59	0.97	-1.07	0.35	1.16	-0.42	0.82
13	1.03	-1.47	0.44	0.94	-1.26	0.61	1.00	-1.50	0.99	1.03	-1.47	0.44
14	1.11	-2.02	0.32	1.07	-1.87	-0.03	1.24	-2.72	-0.29	1.22	-2.06	-0.20
15	1.15	-2.19	0.10	0.82	-2.97	-0.55	1.15	-2.19	0.10	1.15	-2.19	0.10
16	0.98	-1.95	0.21	0.98	-1.95	0.21	0.98	-1.95	0.21	1.16	-2.44	-0.85
17	0.65	0.05	1.76	0.30	-0.51	3.01	0.22	3.71	7.77	0.33	3.67	6.24
18	0.91	-1.00	0.28	0.89	0.03	1.58	0.91	-1.00	0.28	0.91	-1.00	0.28
19	0.67	-2.37	-1.12	0.37	-3.04	-0.88	0.51	-2.20	-0.43	0.73	-1.39	-0.24
20	0.85	-1.56	0.24	0.69	-1.32	0.84	0.85	-1.56	0.24	0.63	-0.93	0.64

Note. Boldface implies different item parameter values in the corresponding group according to the differential item functioning analysis. Male off = male criminal offenders; Fem off = female criminal offenders; Psych = male forensic psychiatric patients; File = male criminal offenders scored only from file review.

of the comparison group determines the weights. The DFIT tests suggested a significant difference between the test characteristic curves for female criminal offenders, $\chi^2(1219, N = 1,219) = 1,557.3, p < .01$, and file reviews, $\chi^2(2625, N = 2,626) = 2,4289.5, p < .01$, versus the reference group, but not for forensic psychiatric patients, $\chi^2(1245, N = 1,246) = 1,246.1, p = .45$.

When one is comparing test characteristic curves, it can be particularly important to focus on differences at $\theta = 1.5$, the approximate trait level at which a PCL-R score of 30 is observed. Although there were differences across groups here, the differences were generally small (<2 score points), and the amount of differential test functioning also remained small across other levels of the latent trait. Male criminal offenders receiving the standard administration had slightly higher expected scores at high levels of the trait and slightly lower expected scores at low levels of the latent trait, suggesting the PCL-R may be slightly more discriminating for this group than for the comparison groups.

Test characteristic curves can also be computed for the total scores on each facet. Figure 3 displays the facet characteristic curves, computed as the sum of ERFs across items within each facet type. For female offenders, the most substantial difference was observed for Facet 4 (Antisocial), in which female offenders received lower scores than male offenders. Recall that all five items from this facet displayed DIF such that female offenders received lower scores. The only other substantial difference occurred for Facet 2 (Affective), in which the file reviews achieved higher scores than the reference group, largely due to Items 6, 8, and 16, all of which produced DIF such that file reviews achieved higher scores.

Latent Mean Comparisons

Table 4 reports the latent mean estimates and PCL-R score descriptive statistics for the four groups. The latent mean for male criminal offenders, standard administration, was arbitrarily fixed to 0. All three comparison groups were found to have statistically different latent means compared with this group. Female offenders had an estimate nearly a half standard deviation below that observed for the reference group. The estimate for the file reviews was nearly a full standard deviation lower. Because scalar equivalence appeared to be approximately satisfied for all groups, differences in the latent mean estimates led to basically the same conclusions as when the distributions of PCL-R scores were compared.

The latent mean estimate for the file reviews is perhaps most interesting. The much lower trait estimate suggests that the lower PCL-R scores observed for file reviews are not due to differential functioning of the PCL-R. Of course it is conceivable that the file reviews may have differed from the standard administration respondents in respects other than the method of scoring and did actually have a lower distribution of the trait. Another possibility is that a majority of the PCL-R items exhibit a consistent bias in which the file reviews consistently received lower scores. For instance, it is possible that the anchor items from the file review analysis in reality act in concert to produce lower scores for file reviews. This would imply that the anchor was not functioning effectively in defining a common latent metric for the standard administration and file review offenders. More is said on this issue in the Discussion section.

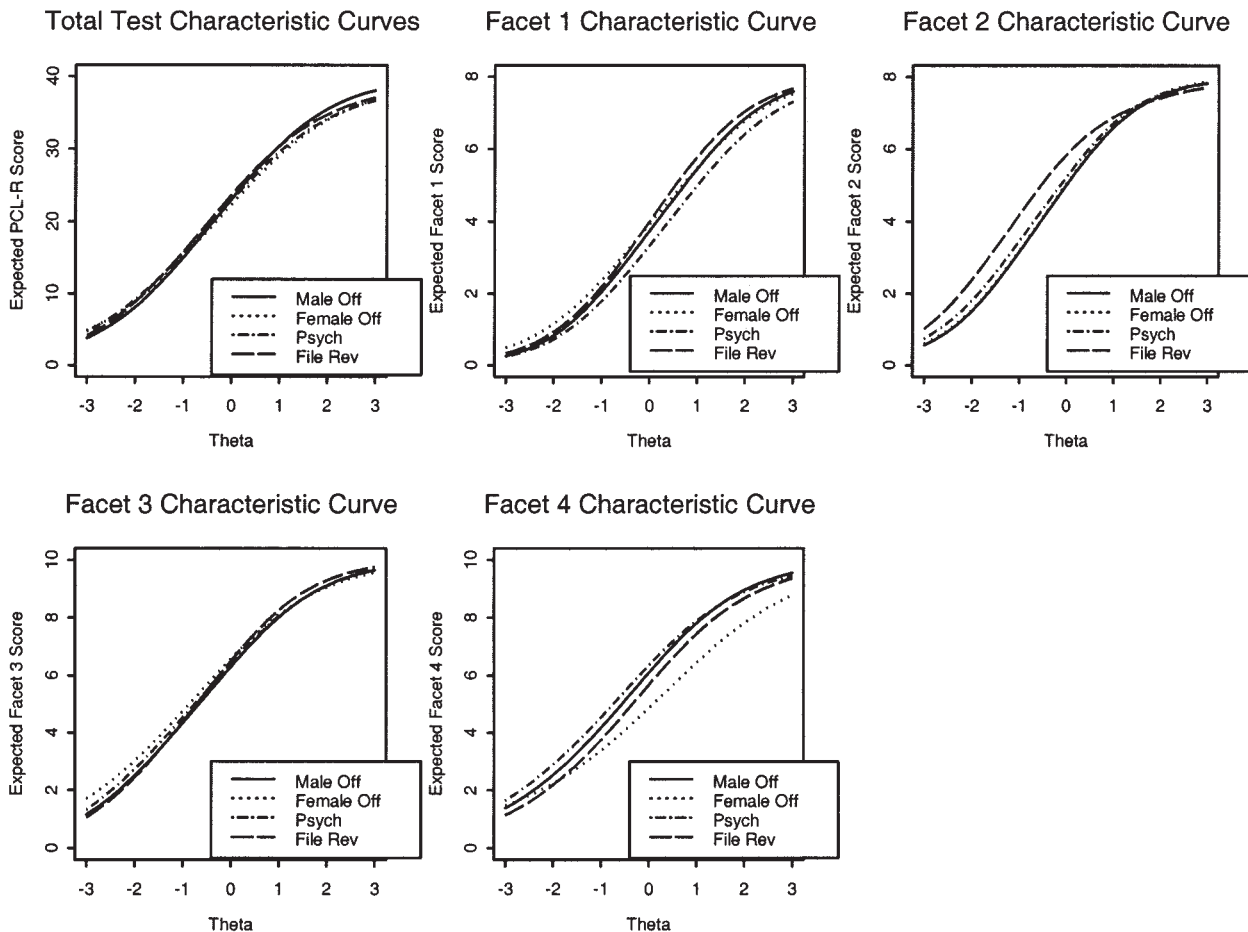


Figure 3. Estimated test and facet characteristic curves for reference (male criminal offenders) and comparison populations based on multigroup graded response model parameter estimates. Facet 1 = Interpersonal; Facet 2 = Affective; Facet 3 = Lifestyle; Facet 4 = Antisocial; PCL-R = Psychopathy Checklist—Revised; Male Off = male criminal offenders; Female Off = female criminal offenders; Psych = male forensic psychiatric patients; File Rev = male criminal offenders scored only from file review.

Comparing Information Functions

On the basis of the GRM estimates in Table 3, Table 5 displays the item information functions for male criminal offenders, stan-

dard administration, across 13 levels of the latent trait. In Table 5, both the within-item and the between-item variability in information across trait levels is apparent. Items such as 6 (“Lack of remorse or guilt”) and 8 (“Callous/lack of empathy”) provide

Table 4
Psychopathy Checklist—Revised (PCL-R) Score Distributions and Estimates of Latent Trait Means Across Studied Groups

Group	Total PCL-R scores				Latent trait means				
	N	M	SD	α	Estimate	SE	χ^2	df	p
Male off	3,847	22.9	7.6	.84	0.00				
Female off	1,219	19.0	7.5	.82	-0.45	0.04	144.2	1	< .001
Psych	1,246	21.2	6.8	.81	-0.19	0.04	28.5	1	< .001
File	2,626	16.1	8.2	.86	-0.96	0.03	864.0	1	< .001

Note. Male off = male criminal offenders; Female off = female criminal offenders; Psych = male forensic psychiatric patients; File = male offenders scored only from file review.

Table 5
Item Information for Male Criminal Offenders

Item	θ level												
	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
1	0.08	0.13	0.20	0.28	0.35	0.38	0.38	0.38	0.38	0.37	0.32	0.24	0.18
2	0.09	0.15	0.21	0.29	0.36	0.40	0.40	0.40	0.40	0.37	0.30	0.21	0.15
3	0.20	0.27	0.38	0.45	0.48	0.48	0.47	0.43	0.34	0.23	0.14	0.10	0.06
4	0.13	0.19	0.29	0.38	0.42	0.42	0.41	0.42	0.43	0.39	0.30	0.20	0.14
5	0.11	0.19	0.29	0.42	0.52	0.55	0.55	0.55	0.50	0.39	0.25	0.14	0.08
6	0.33	0.53	0.63	0.67	0.66	0.67	0.63	0.48	0.29	0.15	0.07	0.03	0.02
7	0.11	0.17	0.26	0.34	0.40	0.41	0.41	0.41	0.40	0.36	0.27	0.21	0.14
8	0.14	0.30	0.50	0.79	0.89	0.82	0.85	0.88	0.68	0.39	0.18	0.09	0.04
9	0.15	0.18	0.24	0.25	0.25	0.24	0.24	0.25	0.25	0.23	0.20	0.15	0.12
10	0.15	0.19	0.27	0.29	0.30	0.30	0.30	0.27	0.23	0.17	0.12	0.10	0.07
11	0.12	0.15	0.20	0.23	0.24	0.25	0.25	0.23	0.20	0.17	0.13	0.09	0.07
12	0.07	0.10	0.16	0.24	0.31	0.37	0.40	0.40	0.36	0.30	0.22	0.13	0.10
13	0.15	0.18	0.26	0.29	0.30	0.31	0.30	0.29	0.25	0.20	0.15	0.11	0.08
14	0.23	0.27	0.33	0.33	0.33	0.33	0.33	0.32	0.27	0.21	0.14	0.10	0.06
15	0.22	0.25	0.36	0.35	0.35	0.36	0.36	0.32	0.26	0.18	0.12	0.08	0.05
16	0.18	0.21	0.26	0.27	0.27	0.27	0.26	0.25	0.21	0.17	0.12	0.08	0.06
17	0.04	0.05	0.07	0.08	0.10	0.11	0.12	0.12	0.12	0.12	0.12	0.08	0.07
18	0.08	0.09	0.17	0.21	0.23	0.24	0.24	0.22	0.19	0.16	0.12	0.06	0.05
19	0.11	0.11	0.13	0.13	0.12	0.11	0.10	0.09	0.07	0.06	0.04	0.03	0.03
20	0.11	0.12	0.18	0.20	0.21	0.21	0.21	0.19	0.17	0.14	0.11	0.07	0.05
Total	2.80	3.83	5.39	6.49	7.09	7.23	7.21	6.90	6.00	4.76	3.42	2.30	1.62

several times greater information than Items such as 17 (“Many short-term marital relationships”) and 19 (“Revocation of conditional release”) and it is generally Factor 1 items that are most informative at the higher trait levels.

Each group’s total information function is illustrated in graphical form in the upper left panel of Figure 4. The contributions of the separate facets are shown in the remaining four panels. Across all groups, the shape of the total information function appears quite consistent, with maximum information provided at or just below $\theta = 0$ (corresponding to psychopathy levels expected to produce PCL-R scores of approximately 20). Slightly more information (approximately 15% to 25% more) appears to be provided for the reference group compared with the comparison groups at this trait level. At higher levels of the latent trait, it is the file reviews that receive the least amount of information of all four groups. This difference appears to be largely due to Facet 2 (and to a lesser extent Facet 1) and is expected because the Affective/Interpersonal items, which function most effectively at high levels of the trait, are more difficult to evaluate based only on file review. For female offenders, it is the Facet 3 (Lifestyle) and Facet 4 (Antisocial) items that provide noticeably less information. Consistent with the observations from the DIF analysis, these items reflect behaviors that not only are likely to be less prevalent among female offenders but also are more poorly related to psychopathy, based on their lower discrimination estimates. For forensic psychiatric patients, slightly less information is observed on all facets except Facet 3, in which the amount of information is nearly coincident with that of the reference group.

Discussion

DIF studies have the potential to provide insight into the validity of individual items (Van de Vijver & Leung, 1997) and have begun

to play an important role in assessing the generalizability of the PCL-R across diverse respondent populations. A common psychometric perspective on DIF attributes its occurrence to an item’s measurement of nuisance factors unrelated to the trait intended to be measured. In this respect, an item’s tendency to display DIF can (along with other psychometric criteria) serve as a criterion for item quality. Consistent with previous comparative DIF analyses of Caucasian and African American male offenders (Cooke et al., 2001), the current DIF analysis for female offenders found that PCL-R items reflecting social deviance (i.e., Factor 2) were more prone to display DIF than the affective/interpersonal (i.e., Factor 1) items. Facet 4 (Antisocial) items appeared to be the primary contributors to differential test functioning, although both Facet 3 (Lifestyle) and Facet 4 items provided less information for female offenders.

Despite the presence of a substantial number of DIF items for all comparison groups, the effects in terms of differential test functioning were found to be quite low. Generally, the cancellation of DIF observed for these comparison groups had an intuitive explanation. For example, among forensic psychiatric patients it was found that some items resulted in higher scores due to a similar symptomology with psychopathy, whereas others appeared to produce lower scores due to their association with criminal offenses. On the basis of the test characteristic curves, male criminal offenders given the standard PCL-R administration did achieve slightly higher scores than each comparison group at the high end of the trait scale, especially near $\theta = 1.5$, the trait level typically associated with psychopathy diagnosis and also had higher information functions. Such findings are perhaps not surprising, given the greater attention devoted to male criminal offender populations in the original development and validation of the instrument. Naturally, those characteristics judged most salient among male

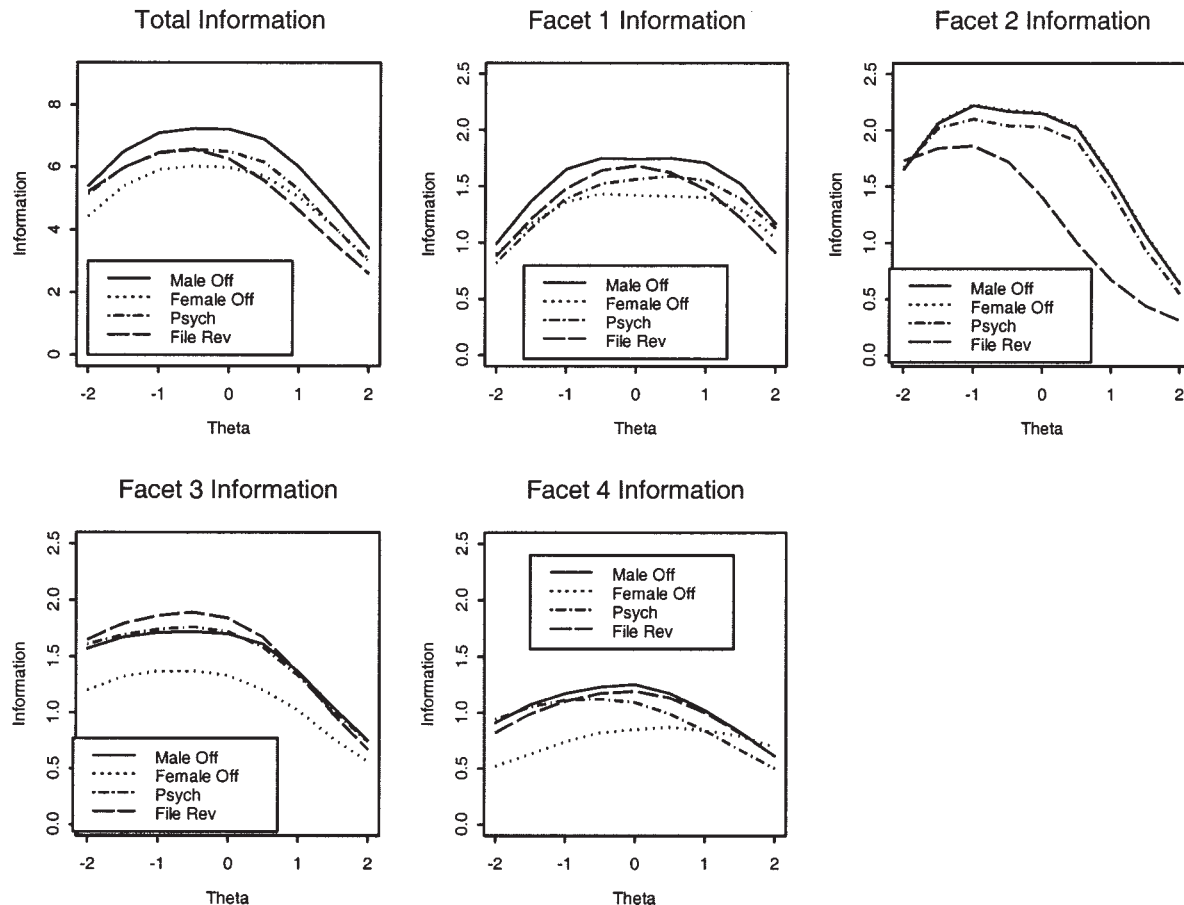


Figure 4. Estimated total and facet information curves for reference (male criminal offenders) and comparison populations based on multigroup graded response model parameter estimates. Facet 1 = Interpersonal; Facet 2 = Affective; Facet 3 = Lifestyle; Facet 4 = Antisocial; Male Off = male criminal offenders; Female Off = female criminal offenders; Psych = male forensic psychiatric patients; File Rev = male criminal offenders scored only from file review.

criminal offender psychopaths, if anything, should function slightly less well in identifying psychopaths from alternative populations.

Nevertheless, the fact that the PCL-R generally performs similarly in terms of both expected scores and information for the comparison groups is encouraging. The differences observed with respect to the test characteristic curves do not appear to require the use of different cut scores in identifying individuals with psychopathy. Likewise, the reduction in information for each comparison group is even lower at $\theta = 1.5$, with approximately a 10% decrease for female offenders and forensic psychiatric patients and approximately a 20% decrease for file reviews, relative to the reference group. Thus the PCL-R appears to remain an effective instrument for distinguishing individuals with psychopathy from those without psychopathy within each comparison group.

There are two reasons to be cautious in the use of IRT analyses as evidence of scalar equivalence, however. One reason is the exploratory approach used to find anchor items. The iterative purification approach used in the current study, similar to that used in previous studies, is not guaranteed to arrive at a final subset of

items that truly perform the same across groups. Indeed, this entire process is guided by the assumption that, in fact, a substantial proportion of the items do function equivalently. Correct identification of these items is critical in that it provides the entire basis for the scaling of the latent trait metric, and thus final conclusions regarding differential test functioning. Alternative approaches to selecting anchor items have been considered. One possibility is to specify them a priori, assuming sufficient information exists to identify which items should be unbiased (see, e.g., Orlando & Marshall, 2002). As several PCL-R studies have now found that Factor 1 items do generally function equivalently across race (and because Factor 1 items are sometimes said to lie closer to the core of psychopathy), it may be worthwhile to consider Factor 1 items as anchors for future analyses, especially comparisons across race and gender.

A second reason for caution is due to the method used to score the PCL-R. Although rater effects are undoubtedly present to some degree, they are ignored in traditional IRT analyses, which model only the respondent-item score interaction. Of course it might be possible to observe consistency in how the scale is used

across groups (and under different scoring conditions) without the scale itself being a valid instrument for assessing psychopathy within each group. For example, it is conceivable that raters may develop a sophisticated understanding of the contingencies among items and apply those in largely the same way across groups, with less regard for whether those same contingencies should be applied in the same way for all comparison groups. In the current study, this possibility would be more likely in the file review analysis than in the analyses of the other groups. A naive interpretation of the results for file reviews would suggest that it is almost as good to score the PCL-R without the interview as with it. Unfortunately, the IRT analyses tell us only that the scale is being used in a largely consistent way when based on only file reviews, not that the ratings are valid. Clearly it is important to supplement the findings here with studies that also incorporate external validation criteria. In this regard, the empirical evidence consistently indicates that file reviews have much the same explanatory and predictive properties in the criminal justice system (recidivism, violence, response to treatment, etc.) as do standard PCL-R assessments (see Hare, 2003). At the same time, however, the observation of a much lower latent mean estimate for file reviews suggests that the offenders scored with file review may have differed in other respects from those scored with the standard administration, thus making the comparison of these groups not one solely based on mode of scoring. It also raises the possibility of a more systematic bias in file review scores that is not effectively studied with the use of IRT. Thus, until we have additional data on the external correlates of PCL-R assessments in a variety of other contexts (e.g., laboratory settings), we are cautious in lending too strong an interpretation to the file review analysis and the near equivalence of the test characteristic curves under both forms of PCL-R scoring.

The sample sizes considered in the current analysis were much larger than those considered in previous IRT DIF analyses with the PCL-R. Ankenmann, Witt, and Dunbar (1999) demonstrated that with the GRM, the LR test is able to detect items with moderate amounts of DIF 65% of the time when the sample sizes for the reference and focal groups are 500. Previous DIF studies using the PCL-R have often relied on samples much smaller than this and thus may not have been able to detect differences with sufficient power. We plan to revisit the possibility of race differences in the PCL-R by using these larger samples.

In conclusion, some PCL-R items performed differently in female offenders, male forensic psychiatric patients, and male offenders assessed from file reviews than they did in male offenders. Future revisions of the PCL-R will attempt to reduce these differences, although the fact that the items with DIF were not the same in each comparison group will make it difficult to develop a set of items that will show little or no DIF across a variety of settings. Meanwhile, the results of this study provide encouraging support for the scalar equivalence of PCL-R total scores in several different offender populations.

References

- Alterman, A. I., Cacciola, J. S., & Rutherford, M. J. (1993). Reliability of the Revised Psychopathy Checklist in substance abuse patients. *Psychological Assessment, 5*, 442-448.

- Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.
- Arnett, P. A., Smith, S. S., & Newman, J. P. (1997). Approach and avoidance motivation in incarcerated psychopaths during passive avoidance. *Journal of Personality and Social Psychology, 72*, 1413-1428.
- Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.
- Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246.
- 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.
- Brinkley, C. A., Bernstein, A., & Newman, J. P. (1999). Coherence in the narratives of psychopathic and nonpsychopathic criminal offenders. *Personality and Individual Differences, 27*, 519-530.
- Chang, H. H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391-404.
- Cleckley, H. (1976). *The mask of sanity*. (5th ed.). St. Louis, MO: Mosby.
- Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350.
- Cooke, D. J. (1995). Psychopathic disturbance in the Scottish prison population: The cross-cultural generalisability of the Hare Psychopathy Checklist. *Psychology, Crime, and Law, 2*, 101-118.
- Cooke, D. J., Kosson, D. S., & Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment, 13*, 531-542.
- 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.
- Day, R., & Wong, S. (1996). Anomalous perceptual asymmetries for negative emotional stimuli in the psychopath. *Journal of Abnormal Psychology, 105*, 648-652.
- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.
- Grann, M., Langstrom, N., Tengstrom, A., & Stalenheim, E. G. (1998). Reliability of the file-based retrospective ratings of psychopathy with the PCL-R. *Journal of Personality Assessment, 70*, 416-426.
- Hare, R. D. (1991). *Manual for the Revised Psychopathy Checklist* (1st ed.). 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*, 99-119.
- Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist* (2nd ed.). 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: Reliability and factor structure. *Psychological Assessment*, 2, 338–341.
- Hare, R. D., & McPherson, L. M. (1984). Violent and aggressive behavior by criminal psychopaths. *International Journal of Law and Psychiatry*, 7, 35–50.
- Hare, R. D., Williamson, S. E., & Harpur, T. J. (1988). Psychopathy and language. In T. E. Moffitt & S. A. Mednick (Eds.), *Biological contributions to crime causation* (pp. 68–92). Dordrecht, the Netherlands: Nijhoff Martinus.
- Harris, G. T., Rice, M. E., & Cormier, C. A. (1991). Psychopathy and violent recidivism. *Law and Human Behavior*, 15, 625–637.
- Hart, S. D. (1998). Psychopathy and risk for violence. In D. J. Cooke, A. E. Forth, & R. D. Hare, (Eds.), *Psychopathy: Theory, research, and implications for society* (pp. 355–373). Dordrecht, the Netherlands: Kluwer Academic.
- Hart, S. D., & Hare, R. D. (1989). Discriminant validity of the Psychopathy Checklist in a forensic psychiatric population. *Psychological Assessment*, 1, 211–218.
- Hemphill, J. F., Templeman, R., Wong, S., & Hare, R. D. (1998). Psychopathy and crime: Recidivism and criminal careers. In D. J. Cooke, A. E. Forth, & R. D. Hare (Eds.), *Psychopathy: Theory, research, and implications for society* (pp. 375–399). Dordrecht, the Netherlands: Kluwer Academic.
- Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Mahwah, NJ: Erlbaum.
- Kosson, D. S., Smith, S. S., & Newman, J. P. (1990). Evaluating the construct of psychopathy in black and white male inmates: Three preliminary studies. *Journal of Abnormal Psychology*, 99, 250–259.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lorenz, A. R., & Newman, J. P. (2002). Deficient response modulation and emotion processing in low-anxious Caucasian psychopathic offenders: Results from a lexical decision task. *Emotion*, 2, 91–104.
- Loucks, A. D., & Zamble, E. (2000). Predictors of criminal behavior and prison misconduct in serious female offenders. *Empirical and Applied Criminal Justice Review*, 1, 1–47.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Ogloff, J. P. R., & Wong, S. (1990). Electrodermal and cardiovascular evidence of a coping response in psychopaths. *Criminal Justice and Behavior*, 17, 231–245.
- Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychologica-Assessment*, 14, 50–59.
- Parker, J., Sitarenios, G., & Hare, R. D. (2003). *Large sample multigroup factor analyses of the Hare Psychopathy Checklist—Revised (PCL-R)*. Manuscript in preparation.
- Patrick, C. J., Bradley, M. M., & Lang, P. J. (1993). Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology*, 102, 82–92.
- Richards, H. J., Casey, J. O., & Lucente, S. W. (2003). Psychopathy and treatment response in incarcerated female substance abusers. *Criminal Justice and Behavior*, 30, 251–276.
- Roussos, L. A., & Stout, W. F. (1995). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, 33, 215–230.
- Rutherford, M. J., Cacciola, J. S., Alterman, A. I., & McKay, J. R. (1996). Reliability and validity of the Psychopathy Checklist in women methadone patients. *Assessment*, 3, 145–156.
- Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and the Psychopathy Checklist—Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice*, 3, 203–215.
- Salekin, R. T., Rogers, R., & Sewell, K. W. (1997). Construct validity of psychopathy in a female offender sample: A multitrait-multimethod evaluation. *Journal of Abnormal Psychology*, 106, 576–585.
- Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika Monograph*, 34(4, Pt. 2).
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.
- Smith, L. L. & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multi-dimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology*, 75, 1350–1362.
- Stark, S. (2001). MODFIT [Computer software]. Retrieved October 17, 2003, from <http://io.psych.uiuc.edu/irt>
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180.
- Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.
- Thissen, D. (2001). *IRTLRDIF v.2. 0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: University of North Carolina.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118–128.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry & Y. H. Poortinga (Eds.), *Handbook of cross-cultural psychology, Vol. 1: Theory and method* (2nd ed., pp. 257–300). Needham Heights, MA: Allyn & Bacon.
- Vitale, J. E., Smith, S. S., Brinkley, C. A., & Newman, J. P. (2002). The reliability and validity of the Psychopathy Checklist—Revised in a sample of female offenders. *Criminal Justice and Behavior*, 29, 202–231.
- Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Erlbaum.
- Warren, J. I., Burnette, M., South, C. S., Chauhan, P., Bale, R., Friend, R., & Van Patten, I. (2003). Psychopathy in women: Structural modeling and co-morbidity. *International Journal of Law and Psychiatry*, 26, 223–242.
- Williamson, S., Harpur, T. J., & Hare, R. D. (1991). Abnormal processing of affective words by psychopaths. *Psychophysiology*, 28, 260–273.
- Windle, M., & Dumenci, L. (1999). The factorial structure and construct validity of the Psychopathy Checklist—Revised among alcoholic inpatients. *Structural Equation Modeling*, 6, 372–393.
- Wong, S. (1988). Is Hare's Psychopathy Checklist reliable without the interview? *Psychological Reports*, 62, 931–934.
- 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.

Received June 11, 2003

Revision received December 1, 2003

Accepted January 5, 2004 ■