Psychological Medicine (2009), **39**, 1967–1978. © Cambridge University Press 2009
doi:10.1017/S0033291709005807

ORIGINAL ARTICLE

A psychometric evaluation of the DSM-IV borderline personality disorder criteria: age and sex moderation of criterion functioning

S. H. Aggen[1]*, M. C. Neale[1,2], E. Røysamb[3,4], T. Reichborn-Kjennerud[3,5,6] and K. S. Kendler[1,2]

[1] *Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA*
[2] *Departments of Psychiatry and Human Genetics, Medical College of Virginia of Virginia Commonwealth University, Richmond; and the Virginia Institute for Psychiatry and Behavioral Genetics, Richmond, VA, USA*
[3] *Division of Mental Health, Norwegian Institute of Public Health, Norway*
[4] *Institute of Psychology, University of Oslo, Norway*
[5] *Institute of Psychiatry, University of Oslo, Norway*
[6] *Department of Epidemiology, Columbia University, New York, NY, USA*

Background. Despite its importance as a paradigmatic personality disorder, little is known about the measurement invariance of the DSM-IV borderline personality disorder (BPD) criteria; that is, whether the criteria assess the disorder equivalently across different groups.

Method. BPD criteria were evaluated at interview in 2794 young adult Norwegian twins. Analyses, based on item-response modeling, were conducted to test for differential age and sex moderation of the individual BPD criteria characteristics given factor-level covariate effects.

Results. Confirmatory factor analytic results supported a unidimensional structure for the nine BPD criteria. Compared to males, females had a higher BPD factor mean, larger factor variance and there was a significant age by sex interaction on the factor mean. Strong differential sex and age by sex interaction effects were found for the 'impulsivity' criterion factor loading and threshold. Impulsivity related to the BPD factor poorly in young females but improved significantly in older females. Males reported more impulsivity compared to females and this difference increased with age. The 'affective instability' threshold was also moderated, with males reporting less than expected.

Conclusions. The results suggest the DSM-IV BPD 'impulsivity' and 'affective instability' criteria function differentially with respect to age and sex, with impulsivity being especially problematic. If verified, these findings have important implications for the interpretation of prior research with these criteria. These non-invariant age and sex effects may be identifying criteria-level expression features relevant to BPD nosology and etiology. Criterion functioning assessed using modern psychometric methods should be considered in the development of DSM-V.

Received 25 March 2008; Revised 2 March 2009; Accepted 11 March 2009; First published online 29 April 2009

Key words: DSM-IV borderline criteria, item analyses, measurement invariance.

Introduction

Borderline personality disorder (BPD), one of the cluster B Axis II personality disorders in DSM-IV (APA, 1994), is a complex syndrome characterized by pervasive patterns of instability in emotion regulation, interpersonal relationships, self-image and self-control (Skodol *et al*. 2002). The nine DSM-IV BPD diagnostic criteria specify the core cognitive, behavioral and interpersonal features that identify and differentiate BPD from other personality and psychiatric disorders.

A central assumption made when diagnoses of BPD are compared across populations or between subgroups within populations is that measurement invariance (MI; Meredith, 1993) holds for the diagnostic criteria; that is, the criteria set construct individual differences on the disorder phenotype in the same way across groups. MI is a central concept in psychometrics. It states that individuals with the same factor score should have the same probability for a given observed response regardless of group membership. Tests for MI determine whether items of a test or criteria for a disorder function equivalently across

* Address for correspondence: S. H. Aggen, Ph.D., Virginia Commonwealth University, Department of Psychiatry, PO Box 980126, Richmond, VA 23298-0126, USA.
(Email: saggen@vcu.edu)

groups. Investigating differential item functioning (Holland & Wainer, 1993) typically involves testing whether item discrimination (slope) and difficulty (threshold location) parameters are invariant in the comparison groups. Discrimination is an index of how sharply each criterion probabilistically distinguishes differences on the underlying disorder construct. Thresholds are the locations on the underlying continuous factor where each criterion has a 0.5 probability of being positive.

Among the possible population characteristics that might impact MI for the BPD criteria, sex and age are obvious choices. Although gender bias for DSM personality disorder criteria has been much discussed (Widiger, 1998; Lindsay *et al*. 2000; Flanagan & Blashfield, 2003), we are aware of only one study that investigated this question rigorously using item response modeling (Jane *et al*. 2007). Another study examined differential functioning in young *versus* old individuals for some DSM-IV personality criteria but BPD was unfortunately not included (Balsis *et al*. 2007).

The aim of this study was to investigate the MI of the nine DSM-IV BPD diagnostic criteria in an epidemiologic sample of Norwegian twins. The BPD criteria were first tested for unidimensionality. Next, a common factor model that accounts for the correlated twin structure was used to test for and quantify any age, sex, and age by sex interaction moderation effects on the individual BPD criterion factor loadings and thresholds that depart from expectations derived from the covariate effects on the factor mean and variance. Such covariate effects found to moderate measurement features of the individual BPD criteria may represent: (1) confounds to establishing coherent individual differences on the disorder phenotype, (2) threats to valid group comparisons, and (3) possible sources of criteria-level functioning that are of etiologic and nosologic interest to research on BPD.

Method

Participants

The twin sample came from the Norwegian Institute of Public Health Twin Panel (NIPHTP; Harris *et al*. 2002). Twins were identified through the national Medical Birth Registry, established 1 January 1967. The current panel includes information on 15370 like- and unlike-sexed twins born 1967–1979. Two questionnaire studies were conducted: in 1992 (twins born 1967–1974) and in 1998 (twins born 1967–1979). Altogether, 12700 twins received the second questionnaire, and 8045 responded after one reminder (63%). The sample included 3334 pairs and 1377 single responders.

Data for the present study were taken from an extensive interview of Axis I and Axis II psychiatric disorders. Participants were recruited from 3153 complete twin pairs who agreed to participate. An additional 68 pairs were drawn directly from the NIPHTP. Of these 3221 eligible pairs, 0.8% were unwilling or unable to participate, in 16.2% of the pairs only one twin agreed to be interviewed, and 38.2% did not respond after two contacts. In total, 2794 twins (44% of those eligible) were interviewed. The final sample consisted of 1022 males and 1772 females from 669 monozygotic (MZ) and 717 dizygotic (DZ) pairs and 22 singleton responders. Zygosity was determined by a combination of genotyping and questionnaire data with a predicted misclassification rate of $<1\%$. The mean age of this sample was 28.2 (S.D. = 3.9) years. The age range was 19–36 years. Approval was received from the Norwegian Data Inspectorate and the Regional Ethical Committee. Written informed consent was obtained from all participants after receiving a full description of the study.

Measures and rating scaling

The Structured Interview for DSM-IV Personality (SIDP-IV; Pfohl *et al*. 1995) is a comprehensive semi-structured diagnostic interview designed to assess all DSM-IV Axis II forms of psychopathology. The instrument uses non-pejorative questions organized into topical sections. The specific DSM-IV diagnostic criteria associated with each set of questions were rated as follows: 0 = not present or limited to rare and isolated examples, 1 = subthreshold (some evidence of the behavioral characteristic, but not sufficiently pervasive to be considered present), 2 = present (the behavioral characteristic is expressed consistently for most of the past 5 years), 3 = strongly present (the characteristic is present and associated with subjective distress and functional impairment in social, occupational or intimate relationships). Interviewers used the '5-year rule' requiring that the particular behaviors, cognitions and feelings must have been present and persisted over the 5 years prior to the interview.

A brief summary of the nine DSM-IV BPD diagnostic criteria is given in Table 1, along with frequencies and sample proportions for the four rating categories by males and females. As seen in a prior Norwegian study (Torgersen *et al*. 2001), endorsement rates tended to be low for all criteria. Given the rarity of the 'strongly present' response, we examined the information content of all the rating options for each BPD criterion using a version of the partial credit

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

Table 1. *Summary of interviewer ratings used to categorize the degree of presence of each of the nine DSM-IV borderline personality behavioral criteria in male and female Norwegian twins*

DSM-IV BPD criteria		Sex	Rating categories							
			Frequency				Proportion			
			0	1	2	3	0	1	2	3
BPD-1	Avoid real or imagined abandonment	Male	953	58	8	3	93.2	5.7	0.8	0.3
		Female	1616	119	30	7	91.2	6.7	1.7	0.4
BPD-2	Unstable interpersonal relationships	Male	918	82	16	6	89.8	8.0	1.6	0.6
		Female	1542	169	43	18	87.0	9.5	2.4	1.0
BPD-3	Identity unstable self-image	Male	1003	17	1	1	98.1	1.7	0.1	0.1
		Female	1726	41	5	0	97.4	2.3	0.3	0
BPD-4	Self-damaging impulsivity	Male	750	198	53	21	73.4	19.4	5.2	2.1
		Female	1518	203	42	9	85.7	11.5	2.4	0.5
BPD-5	Recurrent suicidal; self-mutilating	Male	980	29	11	1	96.0	2.8	1.1	0.1
		Female	1651	85	19	17	93.2	4.8	1.1	1.0
BPD-6	Affective instability	Male	869	114	33	6	85.0	11.2	3.2	0.6
		Female	1353	288	103	27	76.4	16.3	5.8	1.5
BPD-7	Chronic feelings of emptiness	Male	942	67	10	3	92.2	6.6	1.0	0.3
		Female	1548	159	44	21	87.4	9.0	2.5	1.2
BPD-8	Inappropriate, intense anger	Male	885	112	18	7	86.6	11.0	1.8	0.7
		Female	1411	293	61	7	79.6	16.5	3.4	0.4
BPD-9	Stress-related paranoid ideation	Male	969	46	6	0	94.9	4.5	0.6	0
		Female	1670	85	15	2	94.2	4.8	0.8	0.1

model (Masters, 1982) in Multilog (Thissen, 1991). The results indicated that this rating option did provide useful information and was thus retained for the MI analyses. As all covariate moderation effects are estimated to be identical across all thresholds within a criterion, only effects for the threshold between rating categories 1 (subthreshold) and 2 (present) are reported.

Item level analysis

Unidimensional structure

To examine the dimensionality of the nine BPD diagnostic criteria, confirmatory factor analyses (CFA) were carried out for the total sample and separately for males and females in Mplus (Muthen & Muthen, 2004) using a robust weighted least squares means and variance adjusted (WLSMV) estimator that has been shown to perform well with ordinal data (Flora & Curran, 2004). Omnibus fit indices and parameter standard errors were adjusted to account for the twin non-independence.

The overall fits of the CFA models were assessed by the comparative fit index (CFI; Bentler, 1990) and the Tucker–Lewis index (TLI; Tucker & Lewis, 1973). Both indexes gauge the relative reduction in misfit for a restrictive single-factor model compared to a null model (Bentler & Bonett, 1980). Values above 0.90 and 0.95 are generally considered acceptable and very good, respectively.

Statistical model and fit comparisons

A path diagram of the single-group common factor model developed to test for differential covariate moderation of the BPD criteria characteristics is presented in Fig. 1. The term covariate is used to refer to 'fixed' variables that can have an impact on the item characteristics (i.e. factor loadings and thresholds); here specifically age, sex and their interaction. The twin non-independence is modeled by fitting separate factor models for each member of the twin pairs labeled Twin 1 and Twin 2. Observed variables are drawn as boxes (□), unobserved variables (factors) are solid-line circles (○), triangles (△) are unit constants for estimating means, and diamonds (◇) denote definition variables for incorporating the observed covariates in the model. Broken line circles are special nodes used to estimate the covariate moderation effects. Single-headed arrows (→) indicate linear regression effects and double-headed arrows (↔) represent variances and covariances.

As a conceptual aid, the model is partitioned into three sections. The top section labeled '1' shows how the covariate effects for the factor mean ($B\rightarrow$) and

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

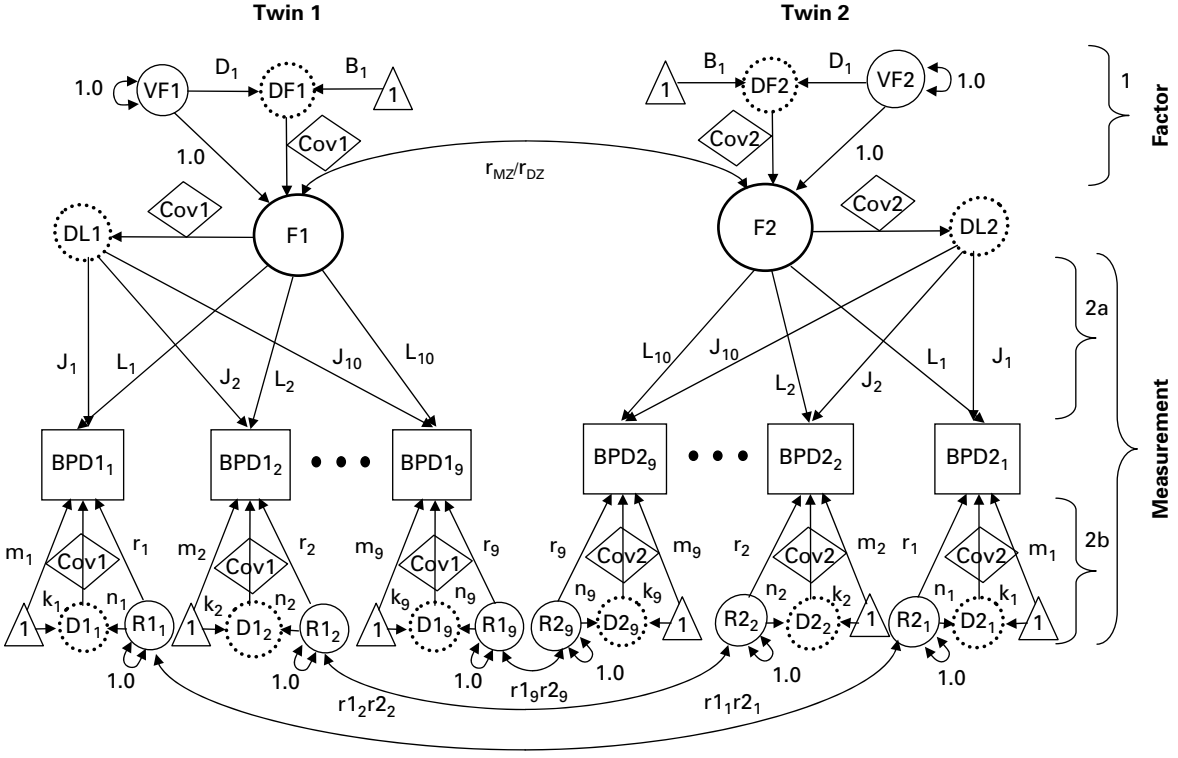

Fig. 1. Path diagram of the common factor model used to test and estimate moderation effects of age, sex, and age by sex interaction on the BPD symptom criteria. Notation: (□) observed variables; (○) unobserved variables (factors); (△) unit constants for estimating means and threshold covariate effects; (◇) definition variables for incorporating covariate effects (e.g. Cov); broken line circles, special nodes used to estimate the covariate moderation effects (e.g. DF and DL); (→) linear regression effects; (↔) variances and covariances, with 1.0 indicating fixed values; VF, factor variance; r_{MZ}/r_{DZ}, estimated monozygotic (MZ) and dizygotic (DZ) twin 1/twin 2 factor correlations; $r1_ir2_i$, twin 1/twin 2 correlations between same BPD criterion residuals.

factor variance ($D\rightarrow$) are specified using the definition variables and special nodes (DF2). These factor level effects are of direct substantive interest and serve as a reference for identifying 'pure' forms of differential covariate moderation of criteria factor loadings and thresholds (Borsboom *et al*. 2002).

Sections 2a and 2b denote the measurement portion of the model. Section 2a identifies the BPD criteria factor loadings and their covariate moderation. Factor loadings are similar to linear regressions of the individual criteria onto the factor. They index the strength of relationship of each criterion with the factor. Factor loadings can be transformed and are equivalent to discrimination parameters in the two-parameter normal theory-based item response model. Larger values indicate steeper slopes. Factor loadings are labeled $L_{\#}$ with their corresponding covariate moderation effects denoted $J_{\#}$. The J's are estimates of the direction and magnitude of how the covariate effects on each BPD criterion loading depart from expectations of the covariate effects at the factor level.

Section 2b shows how the threshold locations ($m_{\#}$) and their differential moderation effects ($k_{\#}$) are obtained. With ordinal data, the covariate moderation parameters $k_{\#}$ estimate changes in threshold locations that deviate from covariate effect expectations on the factor mean. Separate MZ and DZ correlations (r_{MZ}/r_{DZ}) are allowed for the twin1–twin2 common BPD factors (F1 and F2). Specific variances ($r_{\#}$) for each BPD criteria are obtained by formulae calculation. Residual correlations across twins for the same BPD criteria are also estimated. Parameter labels with subscripts (e.g. B_1, D_1, L_i, J_i and K_i) indicate parameters constrained to be equal across twin 1 and twin 2 whereas model element labels without subscripts can take different values. For example, Cov1 and Cov2 indicate that the covariate age, sex, and age by sex interaction definition variables can take on different values for members of a twin pair.

The Mx software (Neale *et al*. 2004) was used to implement a full-information (Bock *et al*. 1988) marginal maximum likelihood ((MML; Bock & Aitkin, 1981) estimation procedure that can accommodate both ordinal and quasi-continuous moderation variables when obtaining model fits and parameter estimates. Optimization is carried out on the raw data

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

by integrating over the factor distribution using a 10-point Gaussian quadrature (Neale *et al*. 2006). The definition variables make it possible to estimate age, sex, and age by sex interaction covariate effects for the factor mean, factor variance and all moderation effects for each BPD criterion factor loading and threshold using the entire sample. This single-group approach has several advantages. First, it limits losses in statistical power because there is no need to partition the sample into groups (e.g. males and females). Second, when including age as a moderator, the continuous linear effect over the full age range of the sample can be estimated without imposing some arbitrary cut-point to define groups (e.g. young *versus* old).

Model comparison tests

To identify significant covariate effects at the factor level and test for differential moderation on the criteria loadings and thresholds, a hierarchical sequence of model comparisons was carried out. First, a baseline model was specified. This model allowed no covariate moderation for any of the factor or criteria parameters shown in Fig. 1 [i.e. factor mean (B), variance (D) or factor loadings ($J_{\#}$) and thresholds ($k_{\#}$)]. This model represents complete MI with respect to the covariates. If the fit of this model cannot be improved upon, there is no evidence of any age, sex or age by sex interaction influences at any level of the model.

Next, a model with all age, sex, and age by sex interaction effects on both the factor mean and variance was compared to the baseline model. If this multivariate test produces a significant reduction in model-data misfit, some factor-level covariate effects are significant. Further comparisons were performed to identify which factor mean and variance covariate effects were statistically reliable.

In the second phase, models allowing covariate moderation for all nine BPD criteria factor loadings and thresholds were compared to the model including all factor mean and variance covariate effects. By first accounting for covariate effects on the factor mean and variance, estimated factor loading and threshold covariate effects represent 'pure' forms of differential item functioning (Borsboom *et al*. 2002). If this multivariate comparison produced a significant likelihood ratio test, additional comparisons are performed to identify which covariate moderation effects on factor loadings and thresholds are responsible for the significant multivariate result. Model fits were assessed by likelihood ratio χ^2 tests and the Akaike Information Criterion (AIC; Akaike, 1981, 1987), where smaller values reflect a better balance of explanatory power and parsimony.

Finally, bootstrapping was performed to obtain 95% confidence intervals (CIs) for all significant factor loading and threshold covariate moderation effects. To illustrate graphically the form of the differential functioning for each BPD criterion, effects were expressed using four points: (1) no covariate effects (operationalized as males with no age effects), (2) male age effect, (3) sex effect (male–female sex difference), and (4) age by sex interaction effect (sex effect plus the age by sex interaction effect).

Results

Unidimensional structure

Table 2 presents the results for the CFA models. For the female only and total samples, the CFI and TLI were good whereas for males they were acceptable. These findings suggest the BPD criteria set has a unidimensional structure in this population.

Factor loadings ranged from 0.48 to 0.79 in the full sample. The impulsivity criterion (BPD-4) had the lowest loading (i.e. least discriminating) in the combined and female samples. The order of loadings differed somewhat in males and females. For example, affective instability (BPD-6) was the most salient indicator of the BPD factor in females whereas the avoid abandonment (BPD-1) criterion was for males.

Factor-level covariate effects

Model-fitting comparisons for the covariate effects on the factor mean and variance are shown in Table 3. Model 1 is the baseline model. Model 2 includes all age, sex, and age by sex interaction effects on the factor mean and variance. This model significantly reduced the overall misfit [$\Delta\chi^2(6) = 36.9$, $p = 0.000$], resulting in a better (smaller) AIC compared to the baseline. Separate comparisons for the factor mean (model 2a) and variance (model 2b) covariate effects both showed significant improvements.

Controlling for factor variance effects, sex (model 3b) and sex by age interaction (3c) effects on the factor mean were significant but an age-only model (3a) was not. Controlling for effects on the factor mean, the BPD factor variance was significantly impacted by sex (model 4b) whereas age (4a) and the sex by age interaction (4c) were not significant. As shown in the last column of Table 3, the interaction of sex and age on the factor mean had a relatively large effect size. The pattern of effects indicates that, compared to males, females have a higher BPD factor mean (effect size 0.20) and this mean difference becomes more pronounce with age (0.49). Compared to males, the BPD factor variance for females was also larger (0.13). It is emphasized, however, that the interpretation of all

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

Table 2. *Confirmatory factor analysis results testing for the unidimensionality of the nine Axis II borderline personality disorder diagnostic criteria*

	Total sample ($n=2794$)			Females ($n=1772$)			Males ($n=1022$)		
	Loadings	S.E.	Res	Loadings	S.E.	Res	Loadings	S.E.	Res
BPD-1	0.65	0.03	0.58	0.62	004	0.62	0.72	005	0.48
BPD-2	0.76	0.02	0.42	0.79	002	0.38	0.68	004	0.54
BPD-3	0.71	0.04	0.50	0.70	005	0.51	0.71	009	0.49
BPD-4	0.48	0.03	0.77	0.53	004	0.71	0.56	004	0.69
BPD-5	0.70	0.03	0.51	0.73	004	0.47	0.66	006	0.58
BPD-6	0.79	0.02	0.38	0.83	002	0.32	0.66	004	0.55
BPD-7	0.67	0.03	0.55	0.70	003	0.51	0.55	006	0.70
BPD-8	0.64	0.03	0.59	0.64	003	0.59	0.65	004	0.58
BPD-9	0.67	0.03	0.55	0.69	004	0.53	0.66	006	0.56
Null	1898.9 (13)			1578.9 (13)			489.5 (13)		
χ^2	118.3 (23)			97.5 (21)			64.7 (21)		
CFI	0.95			0.95			0.91		
TLI	0.97			0.97			0.94		
RMSEA	0.04			0.05			0.05		

BPD, Borderline personality disorder; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; Loadings, factor loading estimates; S.E., factor loading standard error; Res, criterion residual variance; Null, chi-square fit for uncorrelated baseline model; χ^2, chi-square fit for single-factor model.

Degrees of freedom are given in parentheses. For the weighted least squares means and variance adjusted (WLSMV) estimator, degrees of freedom are estimated and data dependent.

Table 3. *Model comparisons testing for age, sex and age by sex interaction effects on the BPD factor mean and variance*

Model	−2lnL	df	Comparison model	$\Delta\chi^2$	Δdf	$p\Delta\chi^2$	AIC	Effect size
(1) No latent effects or criteria moderation	18328.8	25 097		–	–	–	−31865.8	
(2) Age, sex, and interaction factor mean and variance	18291.8	25 091	(1)	36.9	6	**0.000**	−31890.2	–
(2a) Age, sex, and interaction factor mean	18308.4	25 094	(1)	20.4	3	**0.000**	−31879.6	–
(2b) Age, sex, and interaction factor variance	18318.7	25 094	(1)	10.0	3	**0.018**	−31869.3	–
(3a) Age only factor mean	18318.1	25 093	(2b)	0.7	1	0.419	−31867.9	–
(3b) Sex only factor mean	18303.7	25 093	(2b)	15.0	1	**0.000**	−31882.3	**0.20**
(3c) Interaction factor mean	18291.7	25 091	(2b, 3a, 3b)	11.3	1	**0.000**	−31890.3	**0.49**
(4a) Age only factor variance	18306.9	25 093	(2a)	1.4	1	0.239	−31879.0	–
(4b) Sex only factor variance	18293.7	25 093	(2a)	14.7	1	**0.000**	−31892.3	**0.13**
(4c) Interaction factor variance	18291.7	25 091	(2a, 4a, 4b)	0.3	1	0.591	−31890.3	–

BPD, Borderline personality disorder; −2lnL, negative twice the log likelihood; df, degrees of freedom; $\Delta\chi^2$, difference in chi-square between models; Δdf, difference in degrees of freedom between models; $p\Delta\chi^2$, probability associated with chi-square difference; AIC, Akaike Information Criterion.

Values in bold indicate significant reductions in model-data misfit and effect sizes significantly different from zero.

factor-level covariate effects can be compromised or altered if severe forms of differential functioning are present at the level of the individual criterion.

BPD criteria moderation effects

Table 4 presents the results testing for BPD criteria that varied differentially as a function of age, sex, and age by sex interactions given the covariate effects on the BPD factor. The first two lines give fits for the baseline model (BL) and a model (FC) with all mean and variance effects at the factor level (same as models 1 and 2 in Table 3). Although only three of the six factor mean and variance covariate effects were statistically significant (see Table 3), all six were retained to test for differential moderation at the criterion level.

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

Table 4. *Model comparisons testing for age, sex, and age by sex interaction differential moderation effects of the nine DSM-IV BPD criteria factor loadings and threshold locations*

								Age		Sex		Interaction	
	−2lnL	df	$\Delta\chi^2$	Comparison model	Δdf	$p\Delta\chi^2$	AIC	σ^2	μ	σ^2	μ	σ^2	μ
Baseline model, BL (no moderation)	18328.8	25 097	–	–	–	–	−31865.8	–	–	–	–	–	–
Factor level effects													
Factor mean and variance effects (FC)	18291.8	25 091	36.9	BL	6	0.00	−31890.2	–	–	**0.14**	**0.24**	–	0.02
BPD criterion level effects													
BPD criteria								λ	τ	λ	τ	λ	τ
1. Avoid real or imagined abandonment	18282.8	25 085	9.0	FC	6	0.18	−31887.2	–	–	–	–	–	–
2. Unstable interpersonal relationships	18283.4	25 085	8.4	FC	6	0.21	−31886.6	–	–	–	–	–	–
3. Identity unstable self-image	18285.8	25 085	6.0	FC	6	0.43	−31884.2	–	–	–	–	–	–
4. Self-damaging impulsivity	18149.7	25 085	142.1	FC	6	0.00	**−32020.4**	–	–	**−0.16**	**0.27**	**0.29**	**0.52**
5. Recurrent suicidal; self-mutilating	18288.5	25 085	3.2	FC	6	0.78	−31881.5	–	–	–	–	–	–
6. Affective instability	18274.5	25 085	17.2	FC	6	0.01	**−31895.5**	–	–	–	**−0.17**	–	–
7. Chronic feelings of emptiness	18284.6	25 085	7.2	FC	6	0.31	−31885.4	–	–	–	–	–	–
8. Inappropriate, intense anger	18280.7	25 085	11.1	FC	6	0.08	−31889.3	–	–	–	–	–	–
9. Stress-related paranoid ideation	18287.9	25 085	3.8	FC	6	0.71	−31882.0	–	–	–	–	–	–

BPD, Borderline personality disorder; −2lnL, negative twice the log likelihood; df, degrees of freedom; $\Delta\chi^2$, difference in chi-square between models; Δdf, difference in degrees of freedom between models; $p\Delta\chi^2$, probability associated with chi-square difference; AIC, Akaike Information Criterion; μ, factor mean; σ^2, factor variance; λ, factor loading; τ, criterion threshold location.

Values in bold indicate significant reductions in model-data misfit and effect sizes significantly different from zero.

However, the effect sizes reported in the six far-right columns for the FC model are factor mean and variance covariate effects obtained when correcting for all significant differential criteria factor loading and threshold effects. Note that the factor mean interaction effect size is now much smaller (0.02 compared to 0.49 in Table 3) and is no longer significantly different from zero.

As shown in the criterion-level model fitting results of Table 4 (first eight columns), having controlled for factor-level effects, the performance of both BPD-4 (self-damaging impulsivity) and BPD-6 (affective instability) displayed significant forms of differential functioning. Differential moderation was particularly strong [$\Delta\chi^2(6) = 142.1$, $p = 0.000$] for impulsivity. In the six lower right columns of Table 4, the estimated differential effects sizes for factor loadings (λ) and thresholds (τ) are reported. Effects for impulsivity were complex as both λ and τ displayed differential sex ($\lambda = -0.16$, $\tau = 0.27$) and age by sex ($\lambda = 0.29$, $\tau = 0.52$) moderation. The performance of the instability of affect criterion was differentially moderated by sex but only for the threshold ($\tau = -0.17$).

Figure 2 illustrates the nature of these differential covariate moderation effects. For each BPD criterion, four points (medians) with 95% bootstrapped CIs are used to display the differential patterns of age and sex effects on (*a*) the factor loadings and (*b*) the thresholds. Significant effects are denoted by points labeled 1–4. Criteria with no significant effects have identical median bootstrap values and 95% CIs for all four points. The left-most point labeled 1 is a reference point indicating no differential moderation. Differential age moderation for males is given by the adjacent point labeled 2. The point labeled 3 displays the differential sex effect. Finally, the point labeled 4 adds in any age by sex interaction effect.

Examining the pattern of all BPD factor loadings (Fig. 2*a*), no differential moderation was detected for eight of the nine criteria. For these eight criteria, median bootstrapped factor loadings ranged between about 0.50 and 0.75. The impulsivity criterion was differentially moderated by sex and an age by sex interaction. For males, the factor loading is not differentially moderated by age (line connecting points 1 and 2 is flat). However, for females, the factor loading unexpectedly changes with age. In young females, the impulsivity criterion discriminates among individuals differenced on the BPD factor poorly, with a factor loading of about 0.30. Discrimination increases with age so that for the oldest females in the sample, this criterion distinguishes more sharply in females, compared to males, than would be expected based on the estimated factor mean and variance covariate effects.

Differential moderation effects for thresholds are shown in Fig. 2*b*. Two general features are noted. First, the individual BPD criteria differ in their location on the BPD factor. The 'identity disturbance' criterion is most informative at higher scores on the factor whereas, for example, the 'affective instability' functions optimally at lower levels of the factor. Second, in this population-based sample, all criteria primarily distinguish among BPD factor scores above the mean (i.e. the zero point on the factor scale) and tend to operate within a fairly restricted region between about 1.6 and 3.0.

Seven of the nine BPD criterion showed no differential moderation. The criterion of 'affective instability' had a significant but modest differential sex effect, with men reporting less affective instability compared to women, given the same BPD factor level conditional on the factor-level covariate effects. For the impulsivity criterion, differential moderation was pronounced and more complex. There was no differential age moderation for males (i.e. points 1 and 2 are identical). Indeed, for males, it was the most commonly reported BPD criterion. However, for young females with a given factor level conditional on factor covariate effects, this behavioral feature was differentially reported less often (i.e. higher threshold) compared to young males. This male–female discrepancy for reporting impulsivity increases disproportionately even more with age. Combined with the differential moderation effects on the factor loading, the impulsivity criterion displayed a particularly egregious form of differential functioning that impacted and altered the age by sex interaction effect on the factor mean.

Discussion

Model testing of measurement invariance for the nine DSM-IV BPD criteria yielded three key results. First, despite being proposed by a committee and developed with little psychometric guidance, these nine criteria identify a relatively coherent single factor in a general population sample. That a unidimensional structure adequately accounted for the pattern of associations in the BPD criteria set was essential to the subsequent MI model testing and interpretation of results (McDonald, 1981).

Research on the structural organization of the BPD criteria has produced mixed and sometimes inconsistent findings. Although often treated as a categorically singular disorder, it has been characterized as multidimensional in the sense of not being seated in a single diathesis (Paris, 2007). Taxometric studies have shown that the DSM-IV BPD criteria are not consistent with a classification representation but rather fall

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

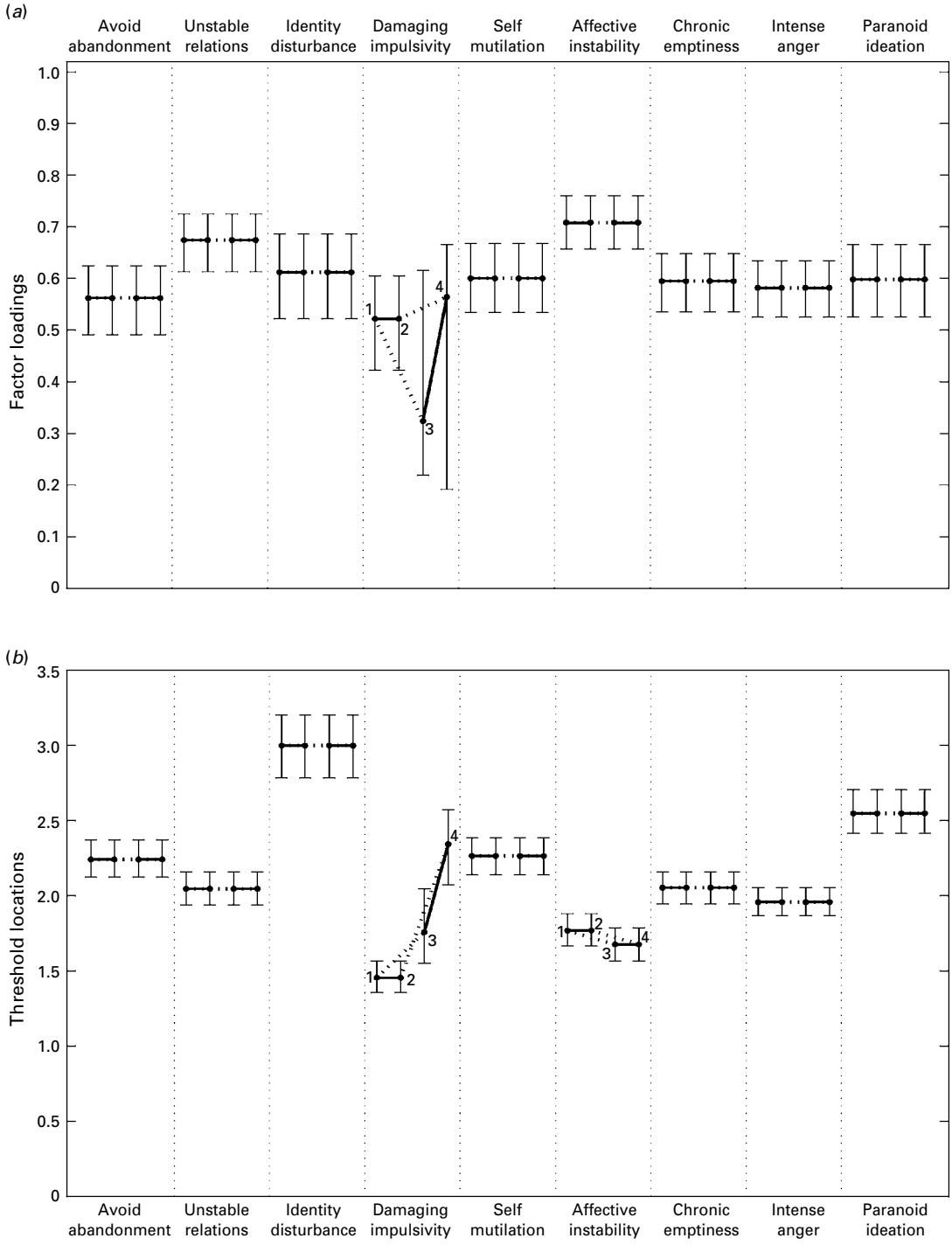

Fig. 2. Bootstrapping results illustrating significant differential age, sex, and age by sex interaction covariate moderation effects on the BPD criteria (*a*) factor loadings and (*b*) thresholds. Four points are used to show the form of the differential age, sex, and age by sex interactions effects for each DSM-IV BPD criteria. Each set of four points is separated by a broken vertical line. Criteria with significant differential moderation effects are labeled by points numbered 1–4. The left-most point is (*a*) the factor loading and (*b*) the threshold estimate ignoring any differential moderation effects. The next point to the right (2) shows the differential age moderation for males. The third point denotes the differential sex moderation effect (female). Finally, the fourth point adds in the differential moderation effect due to an age by sex interaction. Broken connecting lines between points 1 and 3 and between points 2 and 4 highlight differential sex and age by sex interaction moderation effects respectively.

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

along a continuum (Trull *et al*. 1990; Haslam, 2003; Rothschild *et al*. 2003). Exploratory principal component analyses of the BPD criteria have found three highly correlated factors (Sanislow *et al*. 2000; Blais *et al*. 1997; Taylor & Reeves, 2007). CFAs, however, have generally supported a unidimensional structure (Grilo *et al*. 2001; Sanislow *et al*. 2002; Johansen *et al*. 2004; Fossati *et al*. 2006). Most of these structural studies of the BPD diagnostic criteria have the limitation that they were carried out on clinically ascertained samples.

The second major set of findings of this study was that (i) the BPD factor mean significantly differed as a linear function of sex and an age by sex interaction and (ii) the factor variance for females was larger than that of males. These differences can be interpreted substantively. Females, on average, have a higher 'true' level of BPD compared to males. This is consistent with prior studies suggesting that the prevalence of BPD in clinical and most community samples, including one from Norway, is greater in females than males (for a review see Torgersen *et al*. 2001). The significant age by sex interaction effect on the factor mean would have been a novel finding. However, this factor mean effect was importantly found to be linked to the strong differential age by sex interaction effect found for the impulsivity criterion. Accounting for these differential moderating effects nullified the BPD factor mean age by sex interaction effect. Thus, in this case, a significant factor-level effect was found to be due to differential functioning of a single criterion.

The third and most important finding in this sample was that, having taken into account the effects of age, sex and their interaction at the factor level, the hypothesis of MI for the nine DSM-IV BPD criteria was rejected. That is, the measuring properties of the set of BPD criteria were not invariant with respect to age, sex and their interaction. However, the lack of MI was found to be due to only two of the nine criteria: impulsivity (BPD-4) and affective instability (BPD-6). For affective instability, the failure of MI was easily described. This criterion's threshold was differentially moderated by sex. That is, controlling for both the factor-level and covariate effects at the factor level, males report affective instability less often than did females. For impulsivity, MI failures were more pervasive and complex. Both the factor loading and threshold location were differentially moderated by sex and by the interaction of age by sex. As depicted in Fig. 2, the impulsivity factor loading and threshold displayed pronounced changes as a function of age in women that could not be predicted by the covariate effects at the level of the factor. As women age, the level of BPD liability required for a 0.5 probability of endorsing the impulsivity criterion, and also its discriminating power (i.e. the degree to which this criterion reflects the underlying liability to BPD), increased disproportionately. In other words, in younger women, impulsivity was relatively more prevalent (lower threshold) but rather uninformative (low factor loading) in discriminating among levels of the BPD factor. This pattern might arise if there are, in young women, many 'non-specific' aspects of impulsivity that are unrelated to the liability to BPD. However, these non-specific sources decline in importance with aging so that impulsivity becomes a better indicator of BPD. Our results are consistent with one prior study that examined symptom change with age in a clinical and predominantly female group of BPD patients (Stevenson *et al*. 2003). Of the four major symptom dimensions examined, only impulsivity correlated significantly (negatively) with age (Stevenson *et al*. 2003). The BPD impulsivity criterion has also been found to have a relatively non-specific factor-loading pattern in multivariate latent variable modeling of all the Axis II 10 personality disorder criteria (Røysamb *et al*., unpublished observations).

Our results can be further interpreted by comparison with one prior reported study of the BPD criteria using item response modeling (Feske *et al*. 2007). Both studies found that a single-factor solution fit the data well. At the criterion characteristic level, agreement was less consistent. For example, in our sample the threshold for the 'identify disturbance' criterion was located at the highest end of the BPD continuum compared to the other criteria whereas the corresponding criterion in the Feske *et al*. study was 'avoidance of abandonment'. Given that Feske *et al*. used DSM-III-R criteria in a clinical sample whereas we applied DSM-IV criteria to a community sample, such differences may be expected but are still important to note. MI was not examined in Feske *et al*. (2007) but the authors note the importance of doing so. Our findings also departed from the findings of Jane *et al*. (2007), who reported no evidence for gender bias in BPD DSM-IV criteria. However, their sample size was smaller ($n = 599$) and differed in ascertainment, coming from both college students and Air Force recruits who screened positive for personality disorder symptoms on a self-report measure.

Implications

Measurement invariance is an important property for diagnostic criteria to display. If MI holds, comparisons between rates of a disorder in different populations or subgroups within a population can be attributed to valid substantive differences on the construct (Borsboom, 2006). In the absence of MI, however, the interpretation of such differences becomes more

https://doi.org/10.1017/S0033291709005807 Published online by Cambridge University Press

difficult. Prevalence differences and relationships with putative risk factors could reflect 'true' population features or may be confounded with differential functioning of diagnostic criteria. Efforts to explore risk factors or measure treatment response could be seriously compromised because of problems of measurement if the individual criteria do not have the same meaning in different subpopulations.

The findings from this study suggest that caution is advised when comparing BPD diagnoses in groups that differ by age and sex. Our MI analyses identified two of the nine DSM-IV criteria displaying differential age and sex moderation, one of which was particularly egregious. Indeed, these results suggest that the simple exclusion of the impulsivity criterion would eliminate the failure of MI within the BPD criteria. Given the centrality of impulsivity, removing it may lack theoretical justification.

This study also has heuristic value in showing that the failure of MI altered a factor-level result that could be attributed to differential functioning in the impulsivity criterion. The strong differential sensitivity of the BPD impulsivity criterion to sex and age relative to the other BPD criteria may also be of clinical interest. Psychometrics has hitherto not played much of a role in the development and evaluation of diagnostic criteria for psychiatric disorders. Given that both DSM and ICD psychiatric diagnostic manuals are now undergoing revision, it is timely to reconsider this position.

Finally, identifying and describing how the diagnostic criteria set may differentially relate to the disorder phenotype for key covariates also seems to have potential for adding to our nosological, etiological and clinical understanding of BPD. Although typically viewed as threats to valid measurement and group comparisons, differential moderation effects may also represent forms of disorder expression that have substantive significance. The strong differential age and age by sex interaction moderation effects for impulsivity may suggest a more complex organizational relationship between BPD and the symptomatology used to describe it. For example, it may be that differential age moderating effects for a criterion reflect developmental features associated with changes in BPD liability.

Strengths and limitations

This study has several strengths. Our sample is relatively large and epidemiologic. All subjects were evaluated for all BPD criteria without 'skip-outs'. However, the findings should be interpreted in the context of three potential limitations. First, the sample is restricted to young adult Norwegian twins. Second, our sample has undergone attrition and it is possible that this subsample may be unrepresentative. We have explored this question empirically in some detail (Harris *et al.*, unpublished results) and found little evidence that cooperation is predicted by psychopathology. Third, the age range of this sample was relatively restricted. Fourth, age is confounded with cohort in this research sample. Although it seems more likely that changes associated with date of birth are due to age than to social trends over this period, the latter cannot be ruled out.

Acknowledgments

This research was supported in part by grants MH-068643 and MH-65322 from the National Institutes of Health. The twin program of research at the Norwegian Institute of Public Health is supported by grants from the Norwegian Research Council and the Norwegian Foundation for Health and Rehabilitation and by the European Commission under the program 'Quality of Life and Management of the Living Resources' of the 5th Framework Program (no. QLG2-CT-2002-01254).

Declaration of Interest

None

References

Akaike H (1981). Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3–14.

Akaike H (1987). Factor analysis and AIC. *Psychometrika* **52**, 317–332.

APA (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC.

Balsis S, Gleason ME, Woods CM, Oltmanns TF (2007). An item response theory analysis of DSM-IV personality disorder criteria across younger and older age groups. *Psychology and Aging* **22**, 171–185.

Bentler PM (1990). Comparative fit indexes in structural models. *Psychological Bulletin* **107**, 238–246.

Bentler PM, Bonett DG (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* **88**, 588–606.

Blais MA, Hilsenroth MJ, Castlebury FD (1997). Content validity of the DSM-IV borderline and narcissistic personality disorder criteria sets. *Comprehensive Psychiatry* **38**, 31–37.

Bock RD, Aitkin M (1981). Marginal maximum-likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* **46**, 443–459.

Bock RD, Gibbons R, Muraki E (1988). Full-information item factor analysis. *Applied Psychological Measurement* **12**, 261–280.

Borsboom D (2006). When does measurement invariance matter? *Medical Care* **44**, S176–S181.

Borsboom D, Mellenbergh GJ, van Heerden J (2002). Different kinds of DIF: a distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement* **26**, 433–450.

Feske U, Kirisci L, Tarter RE, Pilkonis PA (2007). An application of item response theory to the DSM-III-R criteria for borderline personality disorder. *Journal of Personality Disorders* **21**, 418–433.

Flanagan EH, Blashfield RK (2003). Gender bias in the diagnosis of personality disorders: the roles of base rates and social stereotypes. *Journal of Personality Disorders* **17**, 431–446.

Flora DB, Curran PJ (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods* **9**, 466–491.

Fossati A, Beauchaine TP, Grazioli F, Borroni S, Carretta I, De Vecchi C, Cortinovis F, Danelli E, Maffei C (2006). Confirmatory factor analyses of DSM-IV Cluster C personality disorder criteria. *Journal of Personality Disorders* **20**, 186–203.

Grilo CM, McGlashan TH, Morey LC, Gunderson JG, Skodol AE, Shea MT, Sanislow CA, Zanarini MC, Bender D, Oldham JM, Dyck I, Stout RL (2001). Internal consistency, intercriterion overlap and diagnostic efficiency of criteria sets for DSM-IV schizotypal, borderline, avoidant and obsessive-compulsive personality disorders. *Acta Psychiatrica Scandinavica* **104**, 264–272.

Harris JR, Magnus P, Tambs K (2002). The Norwegian Institute of Public Health Twin Panel: a description of the sample and program of research. *Twin Research* **5**, 415–423.

Haslam N (2003). Categorical versus dimensional models of mental disorder: the taxometric evidence. *Australian and New Zealand Journal of Psychiatry* **37**, 696–704.

Holland PW, Wainer H (1993). *Differential Item Functioning*. Lawrence Erlbaum: Hillsdale, NJ.

Jane JS, Oltmanns TF, South SC, Turkheimer E (2007). Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. *Journal of Abnormal Psychology* **116**, 166–175.

Johansen M, Karterud S, Pedersen G, Gude T, Falkum E (2004). An investigation of the prototype validity of the borderline DSM-IV construct. *Acta Psychiatrica Scandinavica* **109**, 289–298.

Lindsay KA, Sankis LM, Widiger TA (2000). Gender bias in self-report personality disorder inventories. *Journal of Personality Disorders* **14**, 218–232.

Masters GN (1982). A Rasch model for partial credit scoring. *Psychometrika* **47**, 149–174.

McDonald RP (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology* **34**, 100–117.

Meredith W (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika* **58**, 525–543.

Muthen BO, Muthen LK (2004). *Mplus User's Guide*, 3rd edn. Muthen & Muthen: Los Angeles, CA.

Neale MC, Aggen SH, Maes HH, Kubarych TS, Schmitt JE (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors* **31**, 1010–1034.

Neale MC, Boker SM, Xie G, Maes HH (2004). *Mx: Statistical Modeling*, 6th edn. Department of Psychiatry, Medical College of Virginia, Virginia Commonwealth University: Box 980126, Richmond, VA 23298.

Paris J (2007). The nature of borderline personality disorder: multiple dimensions, multiple symptoms, but one category. *Journal of Personality Disorders* **21**, 457–473.

Pfohl B, Blum N, Zimmerman M (1995). *Structured Interview for DSM-IV Personality* (*SIDP-IV*). Department of Psychiatry, University of Iowa: Iowa City.

Rothschild L, Cleland C, Haslam N, Zimmerman M (2003). A taxometric study of borderline personality disorder. *Journal of Abnormal Psychology* **112**, 657–666.

Sanislow CA, Grilo CM, McGlashan TH (2000). Factor analysis of the DSM-III-R borderline personality disorder criteria in psychiatric inpatients. *American Journal of Psychiatry* **157**, 1629–1633.

Sanislow CA, Grilo CM, Morey LC, Bender DS, Skodol AE, Gunderson JG, Shea MT, Stout RL, Zanarini MC, McGlashan TH (2002). Confirmatory factor analysis of DSM-IV criteria for borderline personality disorder. *American Journal of Psychiatry* **159**, 284–290.

Skodol AE, Gunderson JG, Pfohl B, Widiger TA, Livesley WJ, Siever LJ (2002). The borderline diagnosis I: Psychopathology comorbidity, and personality structure. *Biological Psychiatry* **51**, 936–950.

Stevenson J, Meares R, Comerford A (2003). Diminished impulsivity in older patients with borderline personality disorder. *American Journal of Psychiatry* **160**, 165–166.

Taylor J, Reeves M (2007). Structure of borderline personality disorder symptoms in a nonclinical sample. *Journal of Clinical Psychology* **63**, 805–816.

Thissen D (1991). *Multilog Users' Guide: Multiple Categorical Item Analysis and Test Scoring using Item Response Theory*. Scientific Software, Inc.: Chicago, IL.

Torgersen S, Kringlen E, Cramer V (2001). The prevalence of personality disorders in a community sample. *Archives of General Psychiatry* **58**, 590–596.

Trull TJ, Widiger TA, Guthrie P (1990). Categorical versus dimensional status of borderline personality disorder. *Journal of Abnormal Psychology* **99**, 40–48.

Tucker LR, Lewis C (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* **38**, 1–10.

Widiger TA (1998). Invited essay: sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders* **12**, 95–118.