Journal of Abnormal Psychology
2014, Vol. 123, No. 1, 231–236

0021-843X/14/$12.00 DOI: 10.1037/a0035637

BRIEF REPORT

An Investigation of Differential Item Functioning Across Gender of BPD Criteria

Carla Sharp
University of Houston and The Menninger Clinic, Houston, Texas

Jared Michonski
Dialectical Behavior Therapy Center of Seattle and Seattle Children's Hospital, Seattle, Washington

Lynne Steinberg
University of Houston

J. Christopher Fowler
The Menninger Clinic, Houston, Texas and Baylor College of Medicine

B. Christopher Frueh
The Menninger Clinic, Houston, Texas and University of Hawaii

John M. Oldham
The Menninger Clinic, Houston, Texas and Baylor College of Medicine

Gender differences in prevalence rates of Borderline Personality Disorder (BPD) may reflect true differences between groups or may reflect some form of gender bias in diagnostic criteria. The detection of differential item functioning (DIF) using item response theory methods provides a powerful method of evaluating whether gender differences in prevalence rates of BPD reflect true mean differences or criterion bias. The aim of the current study was to evaluate gender-based DIF in DSM BPD criteria. The Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM–IV*) Axis II Personality Disorders (SCID-II: First, Spitzer, Gibbon, Williams, & Benjamin, 1994) was administered to 747 adult inpatients. Results indicated DIF for 2 BPD criteria (impulsivity and uncontrolled anger), such that it was easier for these items to be endorsed for men compared with women at the same level of latent trait. At the level of the test, men were expected to be rated slightly higher than women on the SCID-II at the same level of latent BPD liability. Implications of these results for research and clinical assessment are discussed.

Keywords: Borderline Personality Disorder criteria, gender differences, item response theory, differential item functioning

The topic of gender differences in Borderline Personality Disorder (BPD) has been the subject of much interest over the years. All versions since the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) have indicated that BPD is unequivocally more common in women than men (Sansone & Sansone, 2011). The *DSM-5* (American Psychiatric Association, 2013), for instance, reports a 3:1 female to male gender ratio. However, data from epidemiological studies paint a more complex picture. Whereas the National Epidemiological Survey on Alcohol and Related Conditions found no differences between men and women for BPD (Grant et al., 2008), other studies found higher rates of BPD in women (Kringlen, Torgersen, & Cramer, 2001; Swartz, Blazer, George, & Winfield, 1990), and in the United Kingdom higher rates were reported for men (Coid, Yang, Tyrer, Roberts, & Ullrich, 2006). In clinical samples, a meta-analytic review (Widiger & Trull, 1993) revealed that 76% of borderline patients were women.

Gender differences in prevalence rates may reflect true group differences or may reflect some form of gender bias in diagnostic criteria—that is, criteria may assume unfairly that stereotypical female characteristics (e.g., emotionality) are pathological. Indeed, Jane, Oltmanns, South, and Turkheimer (2007) suggested that greater support has emerged for the argument that assessment instruments may contain gender bias. Whether gender differences in prevalence rates of BPD reflect true mean differences or item

Carla Sharp, Department of Psychology, University of Houston and The Menninger Clinic, Houston, Texas; Jared Michonski, Dialectical Behavior Therapy Center of Seattle and Seattle Children's Hospital, Seattle, Washington; Lynne Steinberg, Department of Psychology, University of Houston; J. Christopher Fowler, The Menninger Clinic, Houston and Department of Psychiatry, Baylor College of Medicine; B. Christopher Frueh, The Menninger Clinic and Department of Psychology, University of Hawaii; John M. Oldham, The Menninger Clinic and Department of Psychiatry, Baylor College of Medicine.

Correspondence concerning this article should be addressed to Carla Sharp, Department of Psychology, University of Houston, 126 Heyne Building Psychiatry, Houston, TX 77204. E-mail: csharp2@uh.edu

bias can effectively be evaluated through the application of item response theory (IRT) methods designed to detect differential item functioning (DIF). DIF occurs when individuals who have the same standing on the latent trait do not have the same probability of item endorsement (Thissen, Steinberg, & Wainer, 1993; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). Because DIF is only identified after controlling for group differences on the latent trait, it can be used as a powerful method to disentangle true group differences from bias at the level of item (or in this case, DSM criterion). Doing so using IRT has become a popular method in the field of educational psychology and psychiatry (e.g., Michonski, Sharp, Steinberg, & Zanarini, 2013; Sharp, Goodyer, & Croudace, 2006).

Despite this potential, DIF analyses using IRT have been applied in only one study of BPD in adults, and there has been a call for more studies examining criterion-level gender differences (Boggs et al., 2009; see, however, Aggen, Neale, Roysamb, Reichborn-Kjennerud, & Kendler, 2009 and Boggs et al., 2009 for confirmatory non-IRT factor analytic studies on this topic). In the only IRT study thus far conducted to examine gender DIF in adult BPD, Jane, Oltmanns, South, and Turkheimer (2007) evaluated gender-based DIF in a nonclinical sample. They found no evidence for gender-based DIF for BPD. Although this study is important, conclusions from these results are tempered by the fact that the study combined participants recruited from a military sample (n = 433) and a college student sample (n = 166) which evidenced very low base rates of BPD (12 and 3 individuals in each sample, respectively). As yet, DIF has not been evaluated in a clinical sample using IRT. Against this background, the aim of the current study was to evaluate gender-based DIF in DSM BPD criteria using IRT in a large clinical sample of adult psychiatric inpatients.

Method

Participants

The sample consisted of 747 inpatient adults (376 males and 371 females) consecutively admitted from October 2011 to March 2013 to a private psychiatric hospital that specializes in treatment refractory patients. All patients were engaged in a six- to eight-week intensive multimodal treatment. Descriptions of the setting, treatment, and extant measures are available in detail elsewhere (Allen et al., 2009). Patients were included in the study regardless of symptom severity or comorbid diagnoses. The majority of the sample were Caucasian (91%), with small percentages identifying as multiracial (5%), American Indian (.7%), Asian (1.5%), and Black/African American (.7%). Average age at admission was 33.46 years (SD = 14.35). Of the 747 total sample, 123 patients (16.5%) met full criteria for *DSM–IV* BPD.

Measures

Demographic variables were assessed using a standardized patient information survey (Allen et al., 2009). Borderline personality disorder criteria were assessed using the research version of the Structured Clinical Interview for *DSM–IV* Axis II Personality Disorders (SCID-II; First et al., 1994).

Procedures

The project was approved by the relevant institutional review boards. Master's level and trained researchers administered SCID-II interviews to all patients admitted to the adult programs at the hospital. The SCID screening questionnaire was not administered and interviewers thoroughly assessed and coded each criterion for BPD and did not utilize any skip-out rules.

Data Analytic Plan

The IRT model fitting and the computation of the test statistics were performed using IRTPRO (Cai, du Toit, & Thissen, 2011). Goodness of fit of the models was evaluated using the M_2 statistics and its associated RMSEA value (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005, 2006; Thissen, 2009), as well as the standardized local dependence (LD) chi-square indices (based on the LD index proposed by Chen & Thissen, 1997). The M_2 statistics represents a suitable proxy for G^2 when the table of item response patterns becomes too sparse to compute the likelihood ratio goodness-of-fit chi-square statistic. Local dependence indicates that the observed covariation among responses to the items in an item-pair exceeds that predicted by the model. The LD indices are standardized chi-square values; values 10 or greater are considered noteworthy (Thissen, 2009) and thus challenge the assumption of unidimensionality.

The 2PL model was fitted to the nine dichotomously scored BPD criteria. For the purposes of the present study, the 2PL model represents the probability of being rated as present for a given criterion as a function of the underlying construct of BPD liability. For each item, two types of parameters are estimated—*discrimination* (or slope) and *threshold* (Embretson & Reise, 2000). The discrimination parameter represents the degree of association between the item response and the underlying construct. The threshold reflects the level of latent BPD liability needed for a given symptom (or personality trait) to be rated as present with .50 probability.

The presence of DIF was investigated using the approach advanced by Thissen, Steinberg, and Wainer (1993). Differences in parameter estimates between groups are evaluated using model comparison tests. To implement this approach, a subset of items ("anchor items") is identified as a means to "link" the groups (allowing for an estimated population group mean difference in the underlying construct). Edelen et al. (2006) recommend identifying anchor items by using an exploratory, iterative process whereby each item is initially tested for DIF by using all other items as the anchor set. Items not showing DIF at this step are regarded as anchor items; the remaining items, referred to as the studied items, are then evaluated for DIF. Wald tests based on the procedure proposed by Lord (1977), providing separate χ^2 statistics for the discrimination and threshold parameters for each studied item, are used to evaluate the presence of DIF. When DIF is detected, effect sizes for the threshold and/or slope parameters will aid the description and interpretation of the group differences (Steinberg & Thissen, 2006).

Results

Frequency Differences by Gender

Before IRT analyses were conducted, gender differences were evaluated in cell assignment to a positive rating for each BPD

criterion. Given the categorical nature of cell assignment, the gender distribution was compared by using the chi square statistic. These results are summarized in Table 1 and show significantly higher frequencies for women for all BPD criteria except impulsivity, chronic feelings of emptiness, and paranoid ideation/dissociation, which showed no differences in frequency of ratings. Women were also significantly more likely to meet criteria for the disorder as a whole. Table 1 also summarizes the frequency by gender for all other disorders in the sample. Bar major depressive disorder and eating disorder, rates of other psychiatric disorders were equivalent across gender. The results of an independent sample t test also demonstrated equivalence in age among men (mean age = 34.71; SD = 14.26) and women (mean age = 36.21; SD = 14.42; $t = -1.31$; $p = .19$).

Item Response Theory Analyses

Unidimensionality. In separate analyses of the item response data for males and females, the 2PL unidimensional IRT model showed satisfactory fit: for males, M_2 (27) = 59.26, $p < .001$; $RMSEA$ = 0.06; for females, M_2 (27) = 58.19, $p < .001$; $RMSEA$ = 0.06. The significant M_2 statistic indicates some model error; however, the RMSEA indicates acceptable fit of the model. None of the standardized chi-square indices of LD approached the value of 10.0 for either males or females. For males, the largest LD value was observed between identity disturbance and chronic feelings of emptiness ($LD\ \chi^2 = 3.6$); likewise, for females, the highest value was observed between identity disturbance and chronic feelings of emptiness ($LD\ \chi^2 = 3.1$). These findings with respect to unidimensionality and local independence offered justification for proceeding with unidimensional IRT analyses.

Detection of DIF. The first step in conducting the DIF analyses was to identify a set of anchor items for linking the male and female subgroups. To do so, each item was initially tested for DIF using all the other items as a tentative anchor. Seven items emerged as not exhibiting DIF, as evidenced by nonsignificant Wald test (χ^2) statistics ($p > .05$). The impulsivity ($p = .02$) and uncontrolled anger ($p = .01$) items were significant. In a separate analysis, the remaining seven items were evaluated for DIF to confirm their appropriateness in serving as anchor items. None of the Wald statistics approached significance, indicating a suitable anchor set. The remaining items (impulsivity and uncontrolled anger) constituted the two studied items and were evaluated for DIF using this seven-item anchor.

For evaluating the Wald tests for the two studied items, Type I error rate was controlled using the Benjamini-Hochberg (B-H, 1995) multiple comparisons procedure. Both of the studied items exhibited DIF. In each case, DIF was concentrated in the threshold (b) parameter, as evidenced by a significant Wald test statistic: for impulsivity, $\chi^2(1) = 8.8$, $p < .01$; and for uncontrolled anger, $\chi^2(1) = 9.0$, $p < .01$. Wald test statistics for slope parameters for both items were nonsignificant: for impulsivity, $\chi^2(1) = 0.0$, $p > .05$.; and for uncontrolled anger, $\chi^2(1) = 0.8$, $p > .05$. A final calibration of item parameters was performed by fitting a model in which the slope and threshold parameters for the anchor items were constrained to be equal across gender and the slope parameters for impulsivity and uncontrolled anger were constrained to be equal across gender, with the threshold parameters freely estimated for impulsivity and uncontrolled anger (see Table 2). Goodness of fit for this model was acceptable: M_2 (68) = 118.71, $p < .001$; $RMSEA$ = 0.03. The population distribution mean BPD liability for females was 0.45 standard units higher than the population distribution mean for males.

DIF items. As mentioned, both the impulsivity and uncontrolled anger items showed significant DIF in the threshold parameters across gender. The item parameters are presented in Table 2. In each case, the direction of DIF was such that it was "easier" for males than for females to be rated as exhibiting the criterion. In terms of effect size, women were 0.51 standard units higher in the level of BPD liability required to have a 50–50 chance of being rated "positive" for impulsivity. For uncontrolled anger, women were 0.55 standard units higher in the level of BPD liability necessary to have a 50–50 chance of being rated "positive" for uncontrolled anger. Figure 1 shows the trace lines for both items. As depicted, the lines differ in their right-left locations (i.e., threshold parameters) for males and females.

Table 1
Gender Differences in Criterion Ratings and Comorbid Diagnoses

Item	Males	Females	χ^2	p value
Criterion 1 – Abandonment fears	14.4%	23.2%	9.54	.002
Criterion 2 – Unstable relationships	21.0%	31.5%	10.69	.001
Criterion 3 – Identity disturbance	14.9%	25.1%	12.11	.001
Criterion 4 – Impulsivity	23.1%	22.6%	0.03	.872
Criterion 5 – Suicidal behaviors	18.9%	33.2%	19.78	<.001
Criterion 6 – Affective instability	19.4%	33.2%	18.21	<.001
Criterion 7 – Chronic emptiness	38.0%	44.5%	3.20	.074
Criterion 8 – Uncontrolled anger	17.8%	17.0%	0.09	.763
Criterion 9 – Paranoid ideation	13.0%	17.5%	2.91	.088
BPD diagnosis	12.0%	21.0%	11.14	.001
Comorbid diagnoses				
Major depressive disorder spectrum	22%	56%	11.59	.001
Anxiety spectrum	23%	25%	1.45	.228
Psychotic spectrum	6%	4%	2.25	.133
Bipolar spectrum	9%	8%	1.56	.211
Eating disorders	.02%	9%	34.95	<.001

Table 2
IRT Item Parameter Estimates

Item	Gender	*a*	*b*
Anchor items			
Criterion 1 – Abandonment fears	Both	1.35 (.18)	1.69 (.16)
Criterion 2 – Unstable relationships	Both	1.63 (.20)	1.18 (.11)
Criterion 3 – Identity disturbance	Both	1.91 (.24)	1.41 (.12)
Criterion 5 – Suicidal Behaviors	Both	1.65 (.20)	1.18 (.11)
Criterion 6 – Affective instability	Both	2.53 (.33)	1.04 (.09)
Criterion 7 – Chronic emptiness	Both	1.31 (.15)	0.59 (.09)
Criterion 9 – Paranoid ideation	Both	1.36 (.19)	1.92 (.19)
Threshold DIF			
Criterion 4 – Impulsivity	Males	1.52 (.23)	1.11 (.15)
	Females	1.52 (.23)	1.62 (.16)
Criterion 8 – Uncontrolled anger	Males	1.66 (.21)	1.33 (.16)
	Females	1.66 (.21)	1.88 (.18)

Note. Values enclosed in parentheses represent standard errors of the parameter estimates. *a* represents the discrimination parameter; *b* represent the threshold parameter. For Impulsivity and Uncontrolled anger, the slope parameter estimates have been constrained equal for men and women and the threshold parameter estimates are estimated separately for men and women.

Anchor items. The 2PL results showed that all symptom criteria were found to be adequately discriminating. Discrimination (slope) parameters are analogous to factor loadings in traditional or confirmatory factor analysis. In fact, discrimination parameters can be translated into factor loadings (see McLeod, Swygert, & Thissen, 2001, p. 199). Values that are 1.0 (corresponding to a factor loading of 0.50) or greater are considered substantial. The discrimination parameters ranged from 1.31 (chronic emptiness) to 2.53 (affective instability). Threshold parameters were all located above the mean, ranging from 0.59 (chronic emptiness) to 1.92 (paranoid ideation).

Overall test curves. The impact of DIF on the test as a whole can be evaluated by considering the test characteristic curve. The test characteristic curve models the expected summed score (i.e., the expected sum of the 9 BPD criterion scores) as a function of one's standing on the latent construct. Figure 2 (upper) shows that for values of the construct ranging from approximately + 0.60 to + 2.30 standard units, men are expected to be rated slightly higher than women on the SCID-II at the same level of latent BPD liability. For males, the expected test score approaches 5 (the number of criteria required to be rated as positive for BPD) at a value of the latent construct of approximately + 1.35 standard units, whereas for females the expected test score approaches 5 at a value of the latent construct of approximately + 1.45 standard units. This is not considered a large effect. Figure 2 (lower) displays the test information curves for both genders for the SCID-II as a whole. These curves indicate where along the continuum of the underlying construct measurement is most precise. As depicted in Figure 2, the BPD score is most informative at the positive end of the continuum, primarily within the range of + 0.75 to + 2.0.

Discussion

The main aim of the current study was to examine the measurement equivalence (or presence of DIF) of each DSM-based BPD criterion across gender using IRT. The rationale for this study lies in the dearth of studies that investigate whether noted gender differences in BPD reflect true group differences or gender bias, and recent calls for further study in this area (Boggs et al., 2009). Widiger and Spitzer (1991) suggested that bias can function at two levels: assessment bias (a biased application of diagnostic criteria), and criterion bias (bias within the defining criteria for the disorder). Consistent with most prevalence studies of BPD in clinical

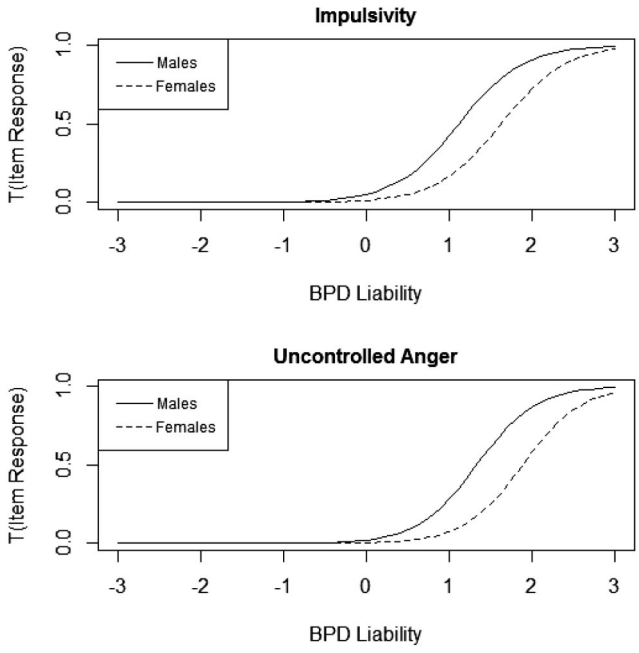

Figure 1. Item characteristic curves (from top to bottom) for impulsivity and uncontrolled anger. These curves depict differential item functioning with respect to thresholds across gender.

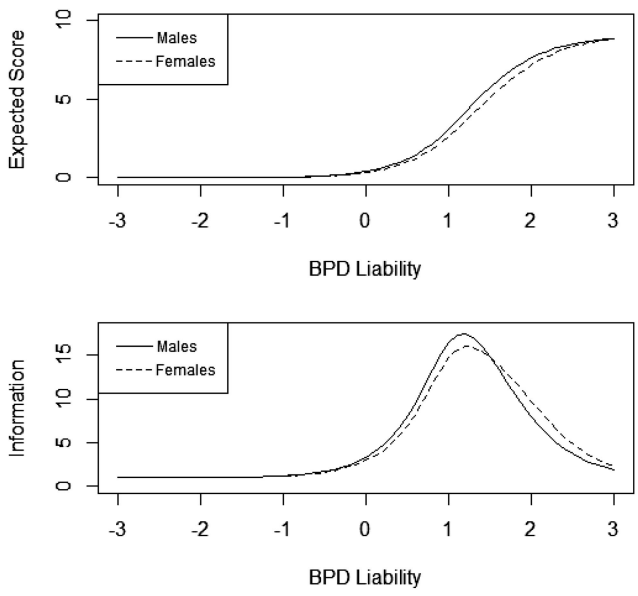

Figure 2. Test characteristic and test information curves (from top to bottom) by gender, with the impact of DIF noticeable in the range .60 to 2.3.

populations, we found significantly higher frequencies of BPD diagnosis for women. In addition, we found higher frequencies of all BPD criteria for women except for the following: impulsivity, chronic feelings of emptiness, and paranoid ideation/dissociation, which showed no differences in frequency of ratings. The results of DIF analyses suggested gender invariance for seven of nine BPD criteria. Uncontrolled anger and impulsivity, however, functioned differently across gender, suggesting that SCID formulations of these DSM criteria make it easier for clinicians to assign positive ratings to men. In other words, these criteria require a lower level of BPD liability in order to be rated as present in men, while women would be rated as positive with higher levels of BPD liability.

To our knowledge, this is the first study to use IRT to evaluate DIF in a clinical sample, and the findings by and large replicate those from the only other IRT study of which we are aware (Jane et al., 2007). Specifically, Jane et al. reported gender invariance for all BPD criteria, whereas we report gender invariance for seven of nine criteria. Our findings do, however, diverge from a recent study by Boggs et al. (2009), which used a regression model of bias to identify bias as differences in slopes or intercepts between men and women in the relationship between each diagnostic criterion and level of impairment. While Boggs et al. utilized a method distinct from the current study in that they examined intercepts and slopes in relation to functional impairment, it is worth noting that bias for all BPD criteria *except impulsivity was demonstrated*. Finally, although not conducted from an IRT perspective per se, one additional study used a latent-trait-based approach to examine the functioning of the BPD criteria at the item level. In this study, Aggen, Neale, Roysamb, Reichborn-Kjennerud, and Kendler (2009) used a categorical CFA approach in a population-based sample and included age, gender, and age by gender covariates to investigate the presence of DIF. Similar to the present study, they found that impulsivity was more "difficult" to endorse among older females. The immediate implication of our findings alongside those of Jane et al. (2007) and Aggen et al. (2009) is that the higher prevalence rate in women is unlikely to reflect a gender bias, given bias was not evident in seven of nine criteria. Also, the observed bias for impulsivity and uncontrolled anger would lead to an overdiagnosis of those particular BPD criteria in men relative to women. The IRT approach to investigating differential item functioning allows us to separate group differences in the distribution of the latent variable from the differences in the probability of endorsement of an item. The evaluation of DIF is "corrected for" the .45 standard unit difference (higher mean for women relative to men) in the population distribution mean BPD liability.

When considering possible explanations for DIF, it is important to bear in mind that DIF may result from "bias" in the traditional sense, wherein there is a problem in the wording of a given test item such that it favors members of a particular subgroup—that is, a measurement artifact (Michonski, 2011; Wicherts & Dolan, 2010). Applied to the SCID II, this form of DIF may occur because the question prompts used to guide an interviewer's rating of a given BPD criterion represent a sample of behaviors that are more stereotypically masculine (or feminine) manifestations of the trait/symptom under consideration. On face value, the SCID questions for the two questions that showed DIF in the current study do not contain gender bias (*Do you have temper outbursts or get so angry you lose control?* and *Have you often done things impulsively?*). However, when clinician interviewers interpret these questions, they are guided by the following DSM criteria: "impulsivity in at least two areas that are potentially self-damaging (e.g., spending, sex, substance abuse, reckless driving, binge eating)" and "inappropriate, intense anger or difficulty controlling anger (e.g., frequent displays of temper, constant anger, recurrent physical fights)" – both of which may be considered stereotypically more masculine behaviors (with the exception of spending and binge eating). That biases lie at the level of the assessment tool per se (and not the criteria) also points to the possibility that the slight divergence in our findings from Jane et al.'s findings may be explained by the fact that Jane et al. used the Structured Interview for *DSM–IV* Personality Disorders (SIPD-IV; Pfohl, Blum, & Zimmerman, 1997). It also points to the importance of specificity in the examples that clinical interviewers use to operationalize DSM criteria.

Additionally, measurement artifacts may occur for reasons of social desirability (Michonski, 2011). An example of this would be if women were less likely to give good examples of, for instance, physical aggression, because they believe it is socially undesirable for women to acknowledge this. Alternatively, DIF may simply occur because, in addition to the common factor that is being measured, a given item also taps a specific factor that really does differ across subgroups (Michonski, 2011; Wicherts & Dolan, 2010). For example, in addition to measuring "BPD liability" (common factor), the impulsivity and uncontrolled anger criteria may also reflect an externalizing tendency (specific factor), more typically observed in psychopathology among men (Hicks et al., 2007).

Several limitations in the current study are of note. It was not possible to calculate interrater reliability on the SCID videos given feasibility issues. Of particular value for future research would be reliability differences across gender for each *DSM–IV* criterion. Moreover, the sample in the current study was not diverse and future research should include patients from more diverse socioeconomic and racial and ethnic backgrounds.

In all, the findings of the current study, in the context of the mixed findings discussed above, caution against a reformulation of BPD criteria of impulsivity and uncontrolled anger in the absence of further research. These considerations also point to the fact that a variety of quantitative methods can and should be used across a variety of samples to clarify mixed findings. Boggs et al. (2009) provocatively quoted Widiger (1998), who noted that the purpose of the DSM system is to provide an accurate classification of psychopathology and not a system that would democratically diagnose as many women as men. The DSM 5 helpfully also adds that "Although these differences in prevalence probably reflect real gender differences in the presence of such patterns, clinicians must be cautious not to overdiagnose or underdiagnose certain personality disorders in females or in males because of social stereotypes about typical gender roles and behaviors." (p. 648). Consistent with this sentiment, we call for more studies that examine whether diagnostic criteria and associated measures optimize the detection of true gender differences. The method of DIF detection using IRT offers a valuable tool in this pursuit.

References

Aggen, S. H., Neale, M. C., Roysamb, E., Reichborn-Kjennerud, T., & Kendler, K. S. (2009). A psychometric evaluation of the *DSM–IV* borderline personality disorder criteria: Age and sex moderation of criterion functioning. *Psychological Medicine, 39,* 1967–1978. doi:10.1017/S0033291709005807

Allen, J. G., Frueh, B. C., Ellis, T. E., Latini, D. M., Mahoney, J. S., Oldham, J. M. . . . Wallin, L. (2009). Integrating outcomes assessment and research into clinical care in inpatient adult psychiatric treatment. *Bulletin of the Menninger Clinic, 73,* 259–295. doi:10.1521/bumc.2009.73.4.25910.1521/bumc.2009.73.4.259[pii]

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th *Edition*). Washington, DC: Author.

Boggs, C. D., Morey, L. C., Skodol, A. E., Shea, M. T., Sanislow, C. A., & Grilo, E. A. (2009). Differential impairment as an indicator of sex bias in *DSM–IV* criteria for four personality disorders. *Personality Disorders: Theory, Research, and Treatment, 17,* 61–68.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible professional item response theory modeling for patient reported outcomes [Computer software]. Chicago, IL: SSI International.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology, 59,* 173–194. doi:10.1348/000711005X66419

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Coid, J., Yang, M., Tyrer, P., Roberts, A., & Ullrich, S. (2006). Prevalence and correlates of personality disorder in Great Britain. *The British Journal of Psychiatry, 188,* 423–431. doi:10.1192/bjp.188.5.423

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach - Application to the Mini-Mental State Examination. *Medical Care, 44,* S134–S142. doi:10.1097/01.mlr.0000245251.83359.8c

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London, UK: Erlbaum.

First, M., Spitzer, R., Gibbon, M., Williams, J., & Benjamin, L. (1994). *Structured Clinical Interview for DSM–IV Axis II personality disorders (SCID II)*. New York, NY: Biometric Research Department.

Grant, B. F., Chou, S. P., Goldstein, R. B., Huang, B., Stinson, F. S., Saha, T. D. . . . Ruan, W. J. (2008). Prevalence, correlates, disability, and comorbidity of *DSM–IV* borderline personality disorder: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Clinical Psychiatry, 69,* 533–545.

Hicks, B. M., Blonigen, D. M., Kramer, M. D., Krueger, R. F., Patrick, C. J., Iacono, W. G., & Mcgue, M. (2007). Gender differences and developmental change in externalizing disorders from late adolescence to early adulthood: A longitudinal twin study. *Journal of Abnormal Psychology, 116,* 433–447. doi:10.1037/0021-843X.116.3.433

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for personality disorders: An item response theory analysis. *Journal of Abnormal Psychology, 116,* 166–175. doi:10.1037/0021-843X.116.1.166

Kringlen, E., Torgersen, S., & Cramer, V. (2001). A Norwegian psychiatric epidemiological study. *The American Journal of Psychiatry, 158,* 1091–1098. doi:10.1176/appi.ajp.158.7.1091

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Portinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, The Netherlands: Swets and Zeitlinger.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association, 100,* 1009–1020. doi:10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713–732. doi:10.1007/s11336-005-1295-9

McLeod, L. D., Swygert, K., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum.

Michonski, J. D. (2011). Borderline personality disorder criteria in a population-based sample of 11 to 12-year-old children: An item-level analysis. Ph.D. dissertation, University of Houston, Houston, TX.

Michonski, J. D., Sharp, C., Steinberg, L., & Zanarini, M. C. (2013). An item response theory analysis of the *DSM–IV* Borderline Personality Disorder criteria in a population-based sample of 11- to 12-year-old children. *Personality Disorders: Theory, Research, and Treatment, 4,* 15–22. doi:10.1037/a0027948

Pfohl, B., Blum, N. & Zimmerman, M. (1997). *Structured Interview for DSMÎV Personality Disorders (SIPD)*. New York, NY: American Psychiatric Press.

Sansone, R. A., & Sansone, L. A. (2011). Gender patterns in borderline personality disorder. *Innovations in Clinical Neuroscience, 8,* 16–20.

Sharp, C., Goodyer, I. M., & Croudace, T. J. (2006). The Short Mood and Feelings Questionnaire (SMFQ): A unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology, 34,* 365–377. doi:10.1007/s10802-006-9027-x

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402–415. doi:10.1037/1082-989X.11.4.402

Swartz, M. S., Blazer, D., George, L., & Winfield, I. (1990). Estimating the prevalence of borderline personality disorder in the community. *Journal of Personality Disorders, 4,* 257–272. doi:10.1521/pedi.1990.4.3.257

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models (pp. 67–113). In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (2009). The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system—IRT software. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement, 29,* 39–47. doi:10.1111/j.1745-3992.2010.00182.x

Widiger, T. A. (1998). Invited essay: Sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders, 12,* 95–118. doi:10.1521/pedi.1998.12.2.95

Widiger, T., & Spitzer, R. (1991). Sex bias in the diagnosis of personality disorders. *Clinical Psychology Review, 11,* 1–22.

Widiger, T. A., & Trull, T. (1993). Borderline and narcissistic personality disorders. In P. Sutker & H. Adams (Eds.), *Comprehensive textbook of psychopathology* (pp. 371–394). New York, NY: Plenum Press.

Received August 29, 2013
Revision received December 2, 2013
Accepted December 3, 2013 ■