Journal of Abnormal Psychology
2007, Vol. 116, No. 1, 166–175

0021-843X/07/$12.00 DOI: 10.1037/0021-843X.116.1.166

Gender Bias in Diagnostic Criteria for Personality Disorders: An Item Response Theory Analysis

J. Serrita Jane
Yale University

Thomas F. Oltmanns
Washington University

Susan C. South and Eric Turkheimer
University of Virginia

The authors examined gender bias in the diagnostic criteria for *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; American Psychiatric Association, 2000) personality disorders. Participants ($N = 599$) were selected from 2 large, nonclinical samples on the basis of information from self-report questionnaires and peer nominations that suggested the presence of personality pathology. All were interviewed with the Structured Interview for *DSM–IV* Personality (B. Pfohl, N. Blum, & M. Zimmerman, 1997). Using item response theory methods, the authors compared data from 315 men and 284 women, searching for evidence of differential item functioning in the diagnostic features of 10 personality disorder categories. Results indicated significant but moderate measurement bias pertaining to gender for 6 specific criteria. In other words, men and women with equivalent levels of pathology endorsed the items at different rates. For 1 paranoid personality disorder criterion and 3 antisocial criteria, men were more likely to endorse the biased items. For 2 schizoid personality disorder criteria, women were more likely to endorse the biased items.

Keywords: gender bias, personality disorders, diagnosis, item response theory

The issue of gender bias with regard to *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) personality disorder criteria has been controversial and widely debated. The current *DSM* (4th ed., text revision; *DSM–IV–TR*; American Psychiatric Association, 2000) makes no explicit statement regarding gender bias among the personality disorders (PDs), but it does suggest that six disorders (antisocial, narcissistic, obsessive-compulsive, paranoid, schizotypal, schizoid) are more frequently found in men. Three others (borderline, histrionic, dependent) are presumably more frequent in women. There are many ways to interpret differential prevalence rates as a function of gender (Corbitt & Widiger, 1995). Some critics have argued that they are an artifact of gender bias (Caplan, 1995; Kaplan, 1983; Walker, 1994). In other words, the PD criteria assume unfairly that stereotypical female characteristics are pathological.

Widiger (1998) described six ways in which gender bias may be related to differential prevalence rates: biases in diagnostic constructs, diagnostic thresholds, sampling of the population, application of the diagnostic criteria, assessment instruments, and diagnostic criteria. Some evidence does support a link between diagnostic constructs and gender stereotypes (Rienzi & Scrams, 1991; Slavney, 1984), but diagnostic thresholds for PDs do not seem to be biased (Funtowicz & Widiger, 1995, 1999). Greater support has emerged for the argument that assessment instruments may contain gender bias, and clinicians may behave in a biased way when they apply PD criteria to men and women. Lindsay and Widiger (1995) found that items in several of the most widely used PD instruments are biased in the sense that they are endorsed more easily by men than by women. Further, research has shown that clinicians do not apply certain PD diagnoses (i.e., histrionic and antisocial) equally to men and women (Garb, 1997).

The evidence is mixed with regard to diagnostic bias in the criteria themselves. Widiger (1998) defined criteria bias as the likelihood that men and women may exhibit the disorder differently because PD criteria include gender-related symptomatology. Anderson, Sankis, and Widiger (2001) conducted a study of potential sources of gender bias in the diagnostic criteria using two samples of professional clinicians from Division 12 of the American Psychological Association. In the first study, 181 clinical psychologists completed measures of statistical infrequency (i.e., how unlikely it would be to find that criterion in a male or female patient) and pathology (i.e., to what extent the presence of that criterion would indicate maladaptive behavior). Measures were collected for *DSM–IV–TR* antisocial, narcissistic, histrionic, and borderline criteria. Significant differences were found for the infrequency ratings with regard to the antisocial, histrionic, and narcissistic criteria sets (in sum, 14 of the 32 diagnostic criteria). Results were not significant for pathology ratings of the four criteria sets (and were significant for only three of the 32 criteria at $\alpha = .05$).

J. Serrita Jane, Department of Psychiatry, Yale University; Thomas F. Oltmanns, Department of Psychology, Washington University; Susan C. South and Eric Turkheimer, Department of Psychology, University of Virginia.

Correspondence concerning this article should be addressed to Thomas F. Oltmanns, Department of Psychology, Washington University, One Brookings Drive, Box 1125, St. Louis, MO 63130-4899. E-mail: toltmann@artsci.wustl.edu

The second study was almost identical, except that criteria from the antisocial, histrionic, narcissistic, and dependent disorders were examined. Results were largely consistent with those of the first study. Infrequency was statistically different for the antisocial, narcissistic, and dependent criteria, but there were no differences by gender for pathology ratings on any of the four criteria sets. The authors concluded that the diagnostic criteria from these five disorders "have differential sex prevalence rates. . . . but professional clinicians who apply these diagnostic criteria to men and women do not perceive the diagnostic criteria as having different implications for maladaptivity or impairment" (Anderson et al., 2001, p. 667).

Only one study has examined differential prevalence ratings on all *DSM–IV–TR* criteria as reported by participants using questionnaires (Morey, Warner, & Boggs, 2002). Gender differences reached significance for 9 of the 79 criteria. As in previous studies (Anderson et al., 2001; Sprock, Crosby, & Nielsen, 2001), participants were also asked to provide pathology ratings (e.g., "a man with this characteristic would have much more trouble functioning than a woman with this characteristic" and vice versa). Results were largely nonsignificant, but when a criterion was viewed as more problematic for one gender, it also tended to be more prevalent in that gender. Morey et al. (2002) concluded that extremes of sex-typed behaviors are viewed by others as most problematic, and "personality problems simply tend to manifest differently in men and women" (p. 62).

Gender bias in the diagnostic criteria for borderline, schizotypal, avoidant, and obsessive-compulsive PDs has also been studied by Boggs et al. (2005). Using data from the Collaborative Longitudinal Personality Disorders Study, these investigators examined relations among diagnostic criteria (measured using semistructured diagnostic interviews) and levels of functional impairment in male and female patients. The data indicated relatively little evidence of gender bias. In other words, specific diagnostic criteria were associated with equivalent levels of impairment in men and women.

In the current study, we investigated whether gender bias is associated with diagnostic criteria for PDs using differential item functioning (DIF), a psychometric method for evaluating whether a construct is expressed equivalently across different groups. Because it is assessed within an item response theory (IRT) framework, DIF can determine whether a person's response on an item (in this case, a PD criterion) depends on both his or her trait level (level of personality pathology) and group membership (gender). In other words, given the same level of personality pathology, are men more likely than women to endorse a PD criterion (or vice versa)?

DIF Using IRT

DIF occurs when two individuals with the same trait level but different group membership do not have the same probability of endorsing a test item. In the current study, the detection of DIF in the *DSM–IV–TR* PD criteria was calculated using IRT.[1] In IRT, an individual's trait level, θ, is estimated from responses to items, and the model specifies how both the trait level and item properties are related to an individual's item responses (Embretson & Reise, 2000). IRT relates the characteristics of the items (in this case, PD diagnostic criteria) and the characteristics of individuals (here, gender) to the probability of endorsing the individual items (Zickar, 1998). Although a full discussion of IRT and DIF is beyond the scope of this article, excellent reviews and articles pertaining to this methodology are available elsewhere (see Embretson & Reise, 2000; Nunnally & Bernstein, 1994; Smith, 2002; Smith & Reise, 1998).

IRT modeling estimates an item characteristic curve (ICC), which is the nonlinear regression trace line that indicates the probability of endorsing the criterion as a function of a person's latent trait level, θ. There are several IRT models, but the most commonly used are the one-, two- and three-parameter models. In the two-parameter model, used in the current study, the ICC is determined by properties of the items, namely discrimination capacity and difficulty. The item discrimination parameter, α, is similar to a factor loading and measures how related the item is to the trait level (i.e., a low value would mean that the probability of endorsing that item is about the same at high and low levels of the latent trait). The difficulty parameter, β, assesses the probability of endorsement given the trait level (i.e., the point on the latent trait at which the probability of endorsing an item becomes .50). The equation for the ICC is

$$P(\theta) = (\exp[\alpha(\theta - \beta)])/(1 + \exp[\alpha(\theta - \beta)]), \quad (1)$$

where α = discriminating function (steepness of slope) and β = difficulty function (trait level necessary to respond above threshold with .50 probability).

The α parameter determines the slope of the ICC, and the β parameter determines the area where the slope of the ICC is the most steep. The larger the α (discrimination) parameter, the more it is associated with a high latent trait. Conversely, the smaller the α parameter, the less it is associated with the latent trait. The ICC represents β (difficulty) as a maximum point of inflection. DIF is present when the ICCs for different groups do not overlap perfectly and an item's α or β parameters are different across groups. When an item has DIF, individuals with the same level of a trait will have different endorsement probabilities (i.e., different ICCs) and therefore different mean raw scores.

IRT has become a popular psychometric method for educational assessment because of its utility for test creation, but it has only recently been applied to the study of personality and psychopathology. For example, Santor, Ramsay, and Zuroff (1994) used IRT methods to examine gender differences on the Beck Depression Inventory. The use of IRT methods is still largely confined to studies of cross-cultural or racial differences in personality and attitude (Cooke, Kosson, & Michie, 2001; Ellis, Becker, & Kimmel, 1993; Ellis, Minsel, & Becker, 1989; Huang, Church, & Katigbak, 1997). Several features of IRT analysis make it particularly appropriate for analyzing personality data. First, it is not necessary to have representative samples of each of the PDs to obtain unbiased estimates of item characteristics (Embretson, 1996). Also, a common metric for θ is used to compare groups. By *anchoring*, or constraining, items to have identical parameters, it is possible to compare directly the trait levels and item parameters for the nonanchor items.

[1] DIF can also be detected by testing for measurement invariance in structural equation models (see Reise, Widaman, & Pugh, 1993).

Current Study

In the current study, we applied the two-parameter logistic IRT model to the responses of two different samples of participants who completed the Structured Interview for *DSM–IV* Personality (SIDP–IV; Pfohl, Blum, & Zimmerman, 1997). Our main goal was to examine whether the difficulty parameters for all items (PD criteria) were gender dependent. In the context of this study, if a criterion displays DIF, then the probability of endorsing that criterion depends on both the individual's level of pathology (PD) *and* his or her gender. That is, DIF occurs when individuals with the same level of pathology, but who differ in gender, do not have the same probability of endorsing the criterion. For instance, if a man and woman had the same level of paranoid pathology, but the man was more likely to endorse a specific paranoid PD criterion, we would say that item displays DIF. To date, there are no published reports of IRT analyses of gender differences of the PD criteria. Therefore, the present study was designed to provide an evaluation of whether the *DSM–IV–TR* PD criteria exhibit DIF for men versus for women.

When DIF does occur on a scale, it is often difficult to compare different groups meaningfully. If endorsement of a particular PD criterion depends on gender, should that criterion be altered to capture the different ways that men and women exhibit personality pathology? Widiger (1998) suggested that the criteria for PDs should minimize the possibility of making false positive and false negative errors in diagnosis. One way to accomplish this is to investigate whether the individual criteria that constitute the PDs are more likely to be endorsed by men or women at the same level of pathology.

If we are to continue to refer to PD syndromes as they are currently defined, mental health professionals must examine whether the defining criteria apply equally to both men and women. In theory, the probability of endorsing a criterion should depend only on the individual's latent level of pathology and not on gender. If DIF occurs within a PD diagnosis, then it becomes difficult to compare men and women on the same disorder, and the utility of applying the diagnostic label is suspect.

Method

Participants

Data for these analyses were collected from 599 individuals who completed semistructured diagnostic interviews as part of a larger study that was concerned with the comparison of self-report questionnaires, semistructured diagnostic interviews, and peer nominations for the assessment of PDs (Oltmanns & Turkheimer, 2006; Thomas, Turkheimer, & Oltmanns, 2003). The study included two nonclinical samples: (a) a military sample consisting of 2,033 United States Air Force recruits (62% male, 38% female) who were tested in groups at the end of 6 weeks of basic military training and (b) a college student sample that included 1,171 undergraduates (36% male, 64% female) tested in dormitory groups after living together for at least 5 months. Both samples included young men and women, but they were not selected specifically for the purpose of studying gender differences. We chose these groups because they provided an opportunity to evaluate the importance of multi-informant reports in comparison with self-report measures. The college and military samples offered an opportunity to collect self-report and peer nomination data within groups composed of people who were relatively well acquainted with each other.

The recruits' ages ranged from 18 to 35 years, with a median age of 19 years. Ninety percent were between the ages of 18 and 25 years. Students' ages ranged from 17 to 27 years, with 98% being either 18 or 19 years of age. The racial composition of the two samples was as follows: military recruits, 65% White, 18% Black, 4% Asian, 4% biracial, 9% other; college students, 66% White, 29% Black, 2% Asian, 3% other. All recruits and students signed informed consent statements and participated on a voluntary basis.

Both samples completed the Multi-Source Assessment of Personality Pathology (Oltmanns & Turkheimer, 2006), a procedure designed to collect both self- and peer-report information about the presence of features of PDs. All participants also completed the Schedule for Nonadaptive and Adaptive Personality (Clark, 1993), a self-report measure that includes diagnostic scales for *DSM–IV–TR* PDs. We selected 433 (256 men, 177 women) from the military sample for interviews. Approximately one third of participants were selected for interviews because their Schedule for Nonadaptive and Adaptive Personality scores suggested that they might exhibit symptoms of PD. Another third of those interviewed were selected because their peers nominated them for exhibiting pathological personality traits. The remaining third of the interviewed participants were selected randomly from the remaining recruits in each flight as controls. In selecting people for interviews, we made an effort to choose equally from among the three clusters of PDs. We selected 166 college students (59 men, 107 women) to complete diagnostic interviews. During the 1st year of the study, students were chosen for interviews at random. During the 2nd year of the study, students were chosen for interview using the same procedure used for the military recruits.

In the military sample, interviews were conducted immediately following screening; in the college student sample, participants were contacted approximately 1 to 3 weeks following the personality screening. Interviewers were kept blind to information regarding scores on all of the screening measures. Twenty-six percent of the military sample and 18% of the college student sample met criteria for at least one PD diagnosis. These rates are comparable to prevalence rates reported in large epidemiological studies (Mattia & Zimmerman, 2001; Torgersen, Kringlen, & Cramer, 2001; Weissman, 1993).

Measures

The SIDP–IV (Pfohl et al., 1997) is a semistructured interview for PDs. The interview covers all criteria for the 10 *DSM–IV–TR* PDs and includes 101 questions that are arranged by themes rather than by disorders (e.g., work style, emotions, interests, and activities). Each PD criterion is assessed by one SIDP–IV question. The interviewer assigns a rating to each criterion using a 4-point scale (0 = *not present*; 1 = *some evidence of the trait*; 2 = *clearly present for most of the last 5 years*; 3 = *strongly present, the criterion is associated with subjective distress*).

Because the SIDP–IV is not accompanied by a formal training or reference manual, we relied extensively on the manual for the Personality Disorder Interview–IV (Widiger, Mangine, Corbitt, Ellis, & Thomas, 1995). During interview training and throughout the data collection process, we often referred to the helpful de-

Table 1
Frequency and Reliability of Personality Disorder Diagnosis and Subthreshold Diagnosis in the Military and College Samples

	Air Force sample								College student sample							
	Men (n = 256)			Women (n = 177)			Reliability		Men (n = 59)			Women (n = 107)			Reliability	
Diagnosis	Freq	%	Sub	Freq	%	Sub	Dx	Continuous	Freq	%	Sub	Freq	%	Sub	Dx	Continuous
Paranoid	8	3.10	16	4	2.30	15	0.57	0.84	0	0.00	2	4	3.74	4	0.35	0.84
Schizoid	0	0.00	6	1	0.56	3	0.01	0.81	1	1.70	2	0	0.00	0	0.00	0.74
Schizotypal	0	0.00	4	0	0.00	4	0.03	0.79	0	0.00	0	1	0.93	1	0.00	0.84
Antisocial	10	3.90	17	3	1.70	3	0.62	0.84	2	3.40	2	1	0.93	1	0.22	0.80
Borderline	10	3.90	12	2	1.10	5	0.60	0.85	1	1.70	1	2	1.87	4	0.78	0.87
Histrionic	3	1.20	6	1	0.56	2	0.55	0.77	1	1.69	1	0	0.00	2	−0.05	0.83
Narcissistic	5	2.00	11	3	1.70	7	0.35	0.82	3	5.08	5	2	1.87	5	0.30	0.83
Avoidant	9	3.50	14	8	4.52	11	0.85	0.93	1	1.69	1	2	1.87	5	0.28	0.85
Dependent	2	0.80	5	2	1.13	3	0.84	0.88	0	0.00	0	0	0.00	0	0.00	0.79
Obsessive-compulsive	24	9.40	40	19	10.70	40	0.55	0.84	1	7.00	7	9	8.40	20	0.35	0.84

Note. Freq = number of people meeting diagnosis; % = percentage of people meeting diagnosis; Dx = diagnosis; Sub = subthreshold (one criterion short of diagnosis).

scriptions of PDs and diagnostic dilemmas that are provided in the Personality Disorder Interview–IV manual. Interview procedures were virtually identical for the two samples, lasting approximately 45 to 90 min. Twelve interviewers (three doctoral level clinical psychologists and nine graduate students in clinical psychology) conducted the 599 interviews. Five of the graduate students had master's level clinical experience. Ten of the interviewers were trained by one of the developers of the SIDP–IV, Nancee Blum; the other two interviewers subsequently received training from one of the doctoral level psychologists.

We did not ask questions from the SIDP–IV that are aimed at the optional research categories (depressive, negativistic, and self-defeating PDs). Questions regarding one criterion for schizoid PD (pertaining to interest or importance of sexual experiences) were not asked at the request of Air Force administrators (because of the "don't ask, don't tell" policy). Questions concerning drug use and sexual orientation were not asked for the same reason. These questions are relevant to one criterion concerning conduct disorder in antisocial PD and one criterion for identity disturbance in borderline PD.

All interviews were recorded on videotape and rated a second time by an independent judge. Reliabilities were computed using intraclass correlations, which are equivalent to kappa coefficients (Fleiss, 1981). Reliabilities were consistently higher for dimensional scores (sum of scores assigned to each criterion), ranging from .77 (histrionic) to .93 (avoidant) in the military sample and from .74 (schizoid) to .87 (borderline) in the college student sample. Consistent with findings from previous studies, the reliabilities for the categorical diagnostic scores were lower (see Table 1) and clearly affected by low frequency (Jane, Pagan, Fiedler, Turkheimer, & Oltmanns, 2006).

Data Analysis

Unidimensionality. In most traditional IRT analyses, it is assumed that one latent dimension underlies the data. The assumption of unidimensionality is often unrealistic, however, because most personality measures are multidimensional. In fact, several factor analyses have found that the *DSM–IV–TR* PDs are not unidimensional constructs (Grilo, 2004; Gude, Hoffart, Hedley, & Ro, 2004). The possible multidimensionality of the PDs may lead to finding DIF where there is none (Drasgow, 1987). Nevertheless, several arguments can be made to support the analyses we conducted. Some experts have concluded that unidimensionality, although highly desirable, is often very difficult to achieve, particularly with personality measurement (Smith, 2002). The best solution may be to show that the personality trait is "sufficiently unidimensional." Others have argued that it is not the unidimensionality of the trait that is of greatest importance but the fact that the variance among the items can ultimately be best explained by one superordinate factor (Stout, 1987). This argument has been supported by research with personality measures that have repeatedly been shown to be composed of multiple latent factors underlying one higher order trait (i.e., the Psychopathy Checklist–Revised; Bolt, Hare, Vitale, & Newman, 2004; Cooke et al., 2001; Cooke, Michie, Hart, & Hare, 1999).

Selection of IRT model. A graded response model was used in these analyses (Samejima, 1970). This is a polytomous IRT model that is used when response items are ordered with increasing valence. This was the most appropriate method for our data because the interviewer using the SIDP–IV is required to rate each criterion on an ordered categorical scale.[2] In this model, the probability of earning a 0, 1, or 2 can be shown by an ICC; as the level of the trait increases, the probability of earning a 2 increases, and the probability of earning a 1 decreases. The shape and position of the curves in relation to the trait are a function of the α (discrimination) and β_1 and β_2 (difficulty) parameters.[3] Recall that the α parameter is equivalent to a factor loading; as the value of α increases, the slope of the curve becomes steeper. The β_1 and β_2 relate to the level of the latent trait needed before scores of 1 and 2 are observed on the item, with negative values indicating

[2] Each SIDP–IV item is rated from 0 to 3, but because the 3 rating was used inconsistently and infrequently, all 3 scores were treated as 2 scores for these analyses.

[3] For an ICC, the number of *b* parameters is equal to the number of item responses minus 1.

relatively easy items and positive values indicating more difficult items; as the values of β_1 and β_2 increase (and the item becomes more difficult), the curve moves further to the right, increasing the amount of the trait at which the item discriminates between people low in the trait versus people high in the trait.

All IRT analyses were performed using MULTILOG-VI (Thissen, 1991). Item parameters were estimated across gender using maximum likelihood methods. As a first step, we created a baseline model in which the mean level of the underlying PD and all item parameters (α, β_1, and β_2) were allowed to vary across gender. We then compared this baseline model with a model in which all item parameters are constrained to be equal across groups. Goodness of fit of the models was compared under generalized likelihood ratio testing theory with the G^2 statistic, with a higher G^2 value indicating worse fit. The difference between the baseline and constrained models, ΔG^2, is distributed as a chi-square with degrees of freedom equal to the number of constraints imposed. If the ΔG^2 is not statistically significant, there is no significant DIF between men and women for the PD in question, and trait levels can be compared directly.

If the ΔG^2 is statistically significant, there is a difference in at least one of the items between men and women, and the data cannot be modeled setting all item parameters equal to each other. This would mean that for a particular PD, men and women are not on the same scale. In this case, it is necessary to establish a common scale between men and women by finding criteria that are invariant across gender. An item-by-item approach was used to find invariant items that would "anchor" the trait level across groups. For each item, a fully constrained model (a model in which the α and βs are constrained) was compared with a model in which α is constrained and the βs are free to vary. If the ΔG^2 was not significant, then a fully constrained model was compared to one in which only the βs are free to vary.

The new model for comparison was chosen by freeing up the β parameters in each of the criteria for the PD individually and examining the ΔG^2. The model including the criterion that produces the largest chi-square difference became the new baseline model from which to compare the remaining criteria. This process continued until there was no significant difference between the models in which the βs of the criteria were freed and the new baseline model. The resulting model included the invariant criteria. At this point, the anchor model should allow for comparison of the PD across men and women, and item bias can be examined.

Results

Overall, 26% of the military sample and 18% of the college sample who were interviewed qualified for a diagnosis of at least one PD. Another 10% of participants from both samples qualified for a "probable" PD, defined as falling one criterion short of the threshold for a diagnosis. Obsessive-compulsive PD was the most frequently diagnosed PD in both samples; the disorders least frequently diagnosed were schizoid, schizotypal, histrionic, and dependent PDs (see Table 1). The proportion of men and women diagnosed with any PD did not differ significantly in either sample. There were no significant gender differences in the distribution of specific PD diagnoses.

To evaluate the amount of multidimensionality present in each of the 10 PDs, we conducted a principal-component factor analysis separately by sample (college students and military). Table 2

Table 2
Eigenvalues From Principal-Component Factor Analysis of SIDP–IV Criteria

Diagnosis	1	2	3	4	5	6	7	8	9
				Military					
Paranoid	2.69	0.96	0.82	0.70	0.66	0.62	0.55		
Schizoid	1.97	1.11	0.91	0.81	0.75	0.45			
Schizotypal	2.16	1.37	1.18	0.98	0.95	0.81	0.61	0.56	0.39
Antisocial	3.31	1.04	0.76	0.69	0.65	0.61	0.48	0.45	
Borderline	3.09	1.23	0.91	0.79	0.73	0.64	0.60	0.53	0.47
Histrionic	2.37	1.14	0.92	0.88	0.82	0.70	0.63	0.54	
Narcissistic	2.82	1.14	0.89	0.87	0.79	0.74	0.61	0.58	0.55
Avoidant	3.12	0.90	0.79	0.70	0.60	0.50	0.38		
Dependent	2.42	1.02	0.99	0.91	0.78	0.74	0.61	0.54	
Obsessive–compulsive	2.14	1.07	0.96	0.92	0.86	0.77	0.68	0.61	
				College student					
Paranoid	3.01	0.97	0.84	0.69	0.57	0.49	0.42		
Schizoid	1.69	1.10	1.01	0.99	0.79	0.43			
Schizotypal	1.97	1.48	1.37	0.97	0.86	0.73	0.70	0.48	0.45
Antisocial	2.98	1.36	0.92	0.77	0.73	0.55	0.45	0.23	
Borderline	3.21	1.16	0.87	0.83	0.74	0.72	0.61	0.52	0.34
Histrionic	2.31	1.19	1.10	0.94	0.74	0.67	0.61	0.44	
Narcissistic	3.40	1.21	0.96	0.76	0.71	0.65	0.50	0.45	0.36
Avoidant	2.58	1.17	0.98	0.79	0.66	0.52	0.30		
Dependent	1.95	1.31	1.19	0.93	0.81	0.68	0.64	0.48	
Obsessive–compulsive	1.96	1.19	1.07	0.96	0.82	0.77	0.66	0.58	

Note. $N = 599$. SIDP–IV = Structured Interview for DSM–IV–Personality.

presents the eigenvalues for each of the 10 PDs. There was a strong first-order factor for each of the 10 PDs in both samples. This first factor was often greater than the second by a ratio of more than 2:1. There was also a strong second factor for many of the PDs but only rarely evidence of a third factor. A notable exception was schizotypal PD in the college student sample, which appeared to be accounted for by a three-factor solution. In general, the presence of these dominant first dimensions suggests that an IRT analysis is appropriate despite the confirmation of some multidimensionality in these personality traits (Smith & Reise, 1998).

DIF was analyzed for each PD criterion, with the 10 PDs as latent traits in separate models. As a first step, for each PD, analyses were conducted to determine whether the items (criteria) had the same relationship to the underlying trait (personality scale) for men and women; in classical test theory, this would be the item-to-scale correlation, but here it is a measure of whether the items were equally discriminating. Comparing the similarity of discrimination parameters in IRT is equivalent to constraining factor loadings between groups in structural equation modeling. This was done so that in subsequent analyses, we could constrain the discrimination parameters to be equal across gender and focus on the difficulty parameters, β_1 and β_2. To evaluate whether the discrimination parameters were equivalent across gender, we compared the unconstrained two-parameter model with a model in which the α parameters were constrained to be equal and the β parameters were free to vary. The resulting ΔG^2 was assessed for significance. We did not expect to find many differences in the slopes, and our results largely supported this belief.

Next, we tested a series of models to find items that were invariant across gender that could act as anchors and establish a common metric for θ. To discover the potentially biased criteria, we permitted each criterion's β parameters to vary and compared them individually with the fully constrained model. When the ΔG^2 indicated an item was significantly different across gender, the new model for comparing the remaining items thus became one in which each subsequent item's α and β parameters were constrained while the β parameters from the item with DIF were allowed to vary. MULTILOG-VI estimated 79 item-discrimination parameters and 316 item-difficulty parameters (158 for each gender). Table 3 presents the final item parameter estimates for items in the criterion sets for paranoid, schizoid, and antisocial PDs. Biased items are listed in bold. None of the items in the other seven criterion sets showed significant gender bias.

Differences in the β parameters (i.e., thresholds) were found for several items. Four *DSM–IV–TR* items appeared to contain gender bias in the sense that men were more likely than women to endorse the items when they possessed the same level of the latent trait. These items included one criterion for paranoid PD (see Table 4): "Perceives attacks on his or her character or reputation that are not apparent to others and is quick to react angrily or counterattack." They also included three criteria for antisocial PD (see Table 5): "failure to conform to social norms with respect to lawful behaviors as indicated by repeatedly performing acts that are grounds for arrest;" "irritability and aggressiveness, as indicated by repeated physical fights or assaults;" and "reckless disregard for safety of self or others."

Table 3
Results of Differential Item Functioning Analysis for Men and Women Using MULTILOG-VI

Criterion	Discrimination *a*	Adjusted b_1		Adjusted b_2	
		M	W	M	W
Paranoid PD					
1. Suspects others	1.56	1.32	1.32	2.51	2.51
2. Doubts loyalty of friends	2.29	1.19	1.19	1.88	1.88
3. Reluctant to confide	1.60	0.88	0.88	1.45	1.45
4. Reads demeaning meanings	2.03	0.77	0.77	1.62	1.62
5. Bears grudges	1.47	1.43	1.43	2.08	2.08
6. Perceives attacks, reacts with anger	**2.11**	**1.18**	**1.70**	**1.83**	**2.75**
7. Suspects sexual partner	1.14	2.02	2.02	2.79	2.79
Schizoid PD					
1. Doesn't enjoy relationships	3.42	2.45	2.45	2.86	2.86
2. Chooses solitary activities	**1.42**	**2.16**	**1.35**	**2.44**	**1.68**
3. Enjoys few activities	1.37	3.22	3.22	4.17	4.17
4. Lacks close friends	**12.11**	**1.74**	**1.48**	**2.44**	**1.60**
5. Indifferent to praise	1.02	2.85	2.85	4.49	4.49
6. Emotional coldness	0.69	3.06	3.06	4.77	4.77
Antisocial PD					
1. Performs illegal acts	**2.80**	**1.10**	**1.41**	**1.50**	**1.84**
2. Deceitfulness	1.89	1.50	1.50	2.32	2.32
3. Impulsivity/failure to plan	1.49	1.64	1.64	2.44	2.44
4. Repeated physical fights	**1.51**	**1.67**	**2.20**	**2.19**	**2.47**
5. Reckless with self/others	**1.65**	**1.27**	**2.18**	**2.13**	**2.71**
6. Inconsistent work behavior	1.40	1.98	1.98	3.16	3.16
7. Lack of remorse	3.40	1.43	1.43	1.75	1.75
8. Conduct disorder	5.04	1.52	1.52	1.62	1.62

Note. $N = 599$. *a* is the slope of the item characteristic curve at the point of inflection, b_1 and b_2 are thresholds. Items displaying differential item functioning are in bold. Item descriptions were shortened for ease of reading. Results are available for the full criteria set for all 10 personality disorders upon request from Thomas F. Oltmanns. M = men; W = women; PD = personality disorder.

Table 4
Item Bias in Paranoid Personality Disorder

Base model		Item 6 free	
Constr[a] = 839.4	$\Delta\chi^2$	Item 6 = 823.3	$\Delta\chi^2$
Item 1 = 837.8	1.6	Item 1 = 820.1	3.2
Item 2 = 837.0	2.4	Item 2 = 822.8	0.5
Item 3 = 837.4	2.0	Item 3 = 818.9	4.4
Item 4 = 834.6	4.8	Item 4 = 821.4	1.9
Item 5 = 836.9	2.5	Item 5 = 822.0	1.3
Item 6 = 823.3	**16.0**	Item 7 = 820.0	3.3
Item 7 = 836.2	3.2		

Note. $N = 599$. For all χ^2, $df = 2$. If $\Delta\chi^2 > 5.99$, then $p < .05$. Boldface type indicates an item that does not fit the model, which means the item is biased.
[a] Constr = all parameters are constrained.

Two criteria appeared to show gender bias in the opposite direction. In other words, for people who possess the same level of the latent trait, women are more likely than men to endorse the item. The following were both criteria for schizoid PD (see Table 6): "lacks close friends or confidants other than first-degree relatives" and "almost always chooses solitary activities."

With metric equivalence established, we were able to compare the overall means on the latent trait. Men had, on average, higher scores on schizoid, antisocial, narcissistic, and avoidant PDs. Women, on average, had higher scores for paranoid, schizotypal, borderline, histrionic, dependent, and obsessive-compulsive scales.

Discussion

Overall, our results suggest there may be relatively little systematic gender bias in the diagnostic criteria for PDs. This finding is generally consistent with conclusions previously reported by Boggs et al. (2005), who studied four types of PD. Among the 79 criteria that define the 10 disorders listed in *DSM–IV–TR*, only 6 showed evidence of differential item functioning. Of course, this finding cannot be used to infer that these diagnoses are never misused or applied as labels in a biased manner (see Garb, 1997). Most of the PD criteria seem to be equally useful in diagnostic decision-making for both genders. We did find that six items performed differently for men and women, and these results suggest a need for further investigation. Our results may be most interesting with regard to the diagnosis of antisocial PD. Three of the items that we found to have potential gender bias were from the criterion set for antisocial PD. These data are consistent with emerging evidence that the antisocial PD and psychopathic personality may present quite differently in men and women (Moffitt, Caspi, Rutter, & Silva, 2001).

DIF of PD Diagnostic Criteria

The following three *DSM–IV–TR* criteria for antisocial PD were more likely to be endorsed by men than women, even when both genders were at the same level of antisocial pathology: "failure to conform to social norms with respect to lawful behaviors, as indicated by repeatedly performing acts that are grounds for arrest;" "irritability and aggressiveness, as indicated by repeated physical fights or assaults;" and "reckless disregard for safety of self or others." The behavioral focus of these criteria reflects changes that were introduced when the criteria for antisocial PD were altered for the third edition of the *DSM* (*DSM–III*; American Psychiatric Association, 1980). The definition of antisocial PD had previously focused more exclusively on criteria that tapped emotional deficits and personality traits such as superficial charm and deceitfulness (Cleckley, 1988). The introduction of more easily measured behavioral criteria improved diagnostic reliability, but these revisions also led to other problems. These included a long and cumbersome set of diagnostic criteria, failure to capture the meaning of the psychopathy construct, and an overdiagnosis of antisocial PD in criminal offenders (Hare, 2003). Experts argued that antisocial PD should be distinguished from criminality; every criminal does not have antisocial PD, and every person with antisocial PD is not a criminal (Cleckley, 1988; Hare, 1993; Lykken, 1995). Although the *DSM–IV–TR* workgroup considered diagnostic changes that would have returned the emphasis of the disorder to a more personality-based conceptualization, the revisions to *DSM–IV–TR* ultimately focused on simplification of already existing criteria.

Do the behavioral features that were added to the definition of antisocial PD truly reflect the original meaning of the antisocial

Table 5
Item Bias in Antisocial Personality Disorder

Base model		Item 5 free		Item 4 free		Item 1 free	
Constr[a] = 835.7	$\Delta\chi^2$	Item 5 = 818.0	$\Delta\chi^2$	Item 4 = 812.1	$\Delta\chi^2$	Item 1 = 806.0	$\Delta\chi^2$
Item 1 = 833.2	2.5	Item 1 = 813.3	4.7	**Item 1 = 806.0**	**6.1**	Item 2 = 805.7	0.3
Item 2 = 835.3	0.4	Item 2 = 818.0	0.0	Item 2 = 812.1	0.0	Item 3 = 804.2	1.8
Item 3 = 833.5	2.2	Item 3 = 816.2	1.8	Item 3 = 810.0	2.1	Item 6 = 804.5	1.5
Item 4 = 831.0	4.7	**Item 4 = 812.1**	**5.9**	Item 6 = 810.0	2.1	Item 7 = 804.6	1.4
Item 5 = 818.0	**18.0**	Item 6 = 815.5	2.5	Item 7 = 811.3	0.8	Item 8 = 801.6	4.4
Item 6 = 818.3	17.0	Item 7 = 817.2	0.8	Item 8 = 810.2	1.9		
Item 7 = 834.4	1.3	Item 8 = 816.7	1.3				
Item 8 = 835.2	0.5						

Note. $N = 599$. For all χ^2, $df = 2$. If $\Delta\chi^2 > 5.99$, then $p < .05$. Boldface type indicates an item that does not fit the model, which means the item is biased.
[a] Constr = all parameters are constrained.

Table 6
Item Bias in Schizoid Personality Disorder

Base model		Item 4 free		Item 2 free	
Constr[a] = 253.9	$\Delta\chi^2$	Item 4 = 239.2	$\Delta\chi^2$	Item 2 = 223.8	$\Delta\chi^2$
Item 1 = 249.9	4.0	Item 1 = 239.1	0.1	Item 1 = 223.7	0.1
Item 2 = 246.3	7.6	**Item 2 = 223.8**	**15.4**	Item 3 = 223.4	0.4
Item 3 = 253.0	0.9	Item 3 = 239.0	0.2	Item 5 = 219.9	3.9
Item 4 = 239.2	**15.0**	Item 5 = 233.6	5.6	Item 6 = 221.4	2.4
Item 5 = 242.9	11.0	Item 6 = 236.2	3.0		
Item 6 = 248.3	5.6				

Note. $N = 599$. For all χ^2, $df = 2$. If $\Delta\chi^2 > 5.99$, then $p < .05$. Boldface type indicates an item that does not fit the model, which means the item is biased.
[a] Constr = all parameters are constrained.

construct, or do they actually serve as a proxy for criminality or a socially deviant lifestyle? Several IRT analyses have been conducted with the Psychopathy Checklist–Revised (Hare, 2003), which has often been found to contain two factors, one more personality-based (Factor 1) and one defined by antisocial behavior (Factor 2). These recent multigroup IRT analyses of the Psychopathy Checklist–Revised (Bolt et al., 2004; Hare, 2003) generally found that items from Factor 1 are more discriminating, whereas items from Factor 2 are the primary contributors of DIF. The criteria that showed evidence of gender bias in the present study are obviously socially unacceptable behaviors that can result in incarceration. In light of emerging research that suggests antisocial PD and psychopathy may have different developmental courses, behavioral expression, and mental health correlates in women (Cale & Lilienfeld, 2002; Moffitt et al., 2001), our results reinforce the possibility that the current antisocial criteria do not adequately reflect how the construct is expressed in women.

One paranoid PD criterion, "perceives attacks on his or her character or reputation that are not apparent to others and is quick to react angrily or counterattack," also showed differential performance by gender. At similar trait levels, the criterion is more likely to be endorsed by men. Several explanations might account for this finding. One is the possibility of gender differences in the overt expression of anger. Another is that the difference may be more specifically related to social conventions regarding what the person is willing to admit during an interview. Men and women may both experience feelings of anger when they believe that they have been attacked, and women may be just as likely to counterattack, but they may also be less willing than men either to admit their angry feelings or describe their angry behavior. It will take skillful interviewers and collateral informants to untangle these options and to determine why we observed gender differences with respect to this criterion.

Summary

Four of the criteria that showed evidence of gender bias were easier for men than for women to endorse (given similar levels of the latent trait). When we consider the content of these items, they seem to make intuitive sense. The literature indicates that boys (a) exhibit more aggressive behavior than do girls, (b) are more likely to approach than to withdraw, (c) are more assertive, and (d) are higher on agentic or instrumental traits (Bakan, 1966; Feingold, 1994; Shiner, 1998). For the men in our study, biased items were those that reflect extremes of sex-typed behavior (e.g., reacting angrily, physicality, recklessness; Morey et al., 2002).

Two of the criteria that showed evidence of gender bias were easier for women than for men to endorse (given similar levels of the latent trait). Their content is more surprising. They are both criteria for schizoid PD, a disorder that is often considered to be more common in men than in women. Furthermore, their content does not reflect extreme forms of female-stereotyped behavior. We see no clear explanation for this finding, and we believe that this aspect of our results should be interpreted with special caution.

It is notable that there was no indication of gender-biased criteria in the borderline, histrionic, and dependent PDs. This is in contrast with what is predicted by critics of these disorders, who suggest they are biased against women. It is possible, however, that other sources of bias, including assessment and clinical bias (Widiger, 1998), are still at work in relation to these disorders. The results do show that the group means are higher in women than in men, an expected result considering the higher prevalence rate of these disorders for women.

Limitations of Current Study

The methods that we used to detect DIF are powerful, but the present study also has some limitations, especially with regard to the samples that we employed. We combined data from two different populations—college students and military recruits—who might be expected to differ in many ways, including personality styles. We are reasonably confident that this mixture of populations did not adversely affect the results because IRT parameters are invariant with respect to samples. Another possible limitation of this study is in the low rate of diagnosis for some of the PDs in our nonclinical samples. However, as long as each item is endorsed at each level, IRT analyzes the pattern of responses in the determination of DIF, and thus the low prevalence rate should not matter. Some may argue that these samples are quite unique in their composition, and the types of personality pathology found in them may not be comparable with general community or clinical samples. For instance, it is possible that both types of samples may pull for individuals who are higher in obsessive-compulsive personality traits. On the other hand, the prevalence rates found in our samples are comparable to results reported from epidemiological surveys in community samples. Further, we would argue that the

type of analysis conducted here is particularly suited to samples that may not be "ideal." IRT is a method that can be applied with unequal samples sizes and when pathology may not reach the level of diagnosis.

The results presented in this article need to be replicated in other, more diverse samples. One benefit to the current study was the large sample size of 599, which provided for greater statistical power in the Graded Response Model (Ankenmann, Witt, & Dunbar, 1999). This was made possible, however, by combining data from two populations that, although sharing some characteristics (age, race), were quite different in other regards (educational background, normal personality traits). Future research should involve community and clinical populations. For example, it would be useful to replicate this study with an older sample, for whom PDs have become more stable and perhaps led to greater social impairment.

References

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Anderson, K. G., Sankis, L., & Widiger, T. A. (2001). Pathology versus statistical infrequency: Potential sources of gender bias in personality disorder criteria. *The Journal of Nervous and Mental Disease, 189,* 661–668.

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.

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion.* Chicago: Rand McNally.

Boggs, C. D., Morey, L. C., Skodol, A. E., Shea, M. T., Sanislow, C. A., Grilo, C. M., et al. (2005). Differential impairment as an indicator of sex bias in DSM–IV criteria for four personality disorders. *Psychological Assessment, 17,* 492–496.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist—Revised. *Psychological Assessment, 16,* 155–168.

Cale, E. M., & Lilienfeld, S. O. (2002). Sex differences in psychopathy and antisocial personality disorder: A review and integration. *Clinical Psychology Review, 22,* 1179–1207.

Caplan, P. J. (1995). *They say you're crazy: How the world's most powerful psychiatrists decide who's normal.* Reading, MA: Addison Wesley.

Clark, L. A. (1993). *Manual for the Schedule for Nonadaptive and Adaptive Personality.* Minneapolis: University of Minnesota Press.

Cleckley, H. M. (1988). *The mask of sanity* (5th ed.). Augusta, GA: Emily S. Cleckley.

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., Hart, S. D., & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist—Revised (PCL: SV): An item response theory analysis. *Psychological Assessment, 11,* 3–13.

Corbitt, E. M., & Widiger, T. A. (1995). Sex differences among the personality disorders: An exploration of the data. *Clinical Psychology: Science and Practice, 2,* 225–238.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24,* 133–148.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology, 24,* 665–684.

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.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Funtowicz, M. N., & Widiger, T. A. (1995). Sex bias in the diagnosis of personality disorders: A different approach. *Journal of Psychopathology and Behavioral Assessment, 17,* 145–165.

Funtowicz, M. N., & Widiger, T. A. (1999). Sex bias in the diagnosis of personality disorders: An evaluation of the *DSM–IV* criteria. *Journal of Abnormal Psychology, 108,* 195–201.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice, 4,* 99–120.

Grilo, C. M. (2004). Factor structure of DSM–IV criteria for obsessive-compulsive personality disorder in patients with binge eating disorder. *Acta Psychiatrica Scandinavica, 109,* 64–69.

Gude, T., Hoffart, A., Hedley, L., & Ro, O. (2004). The dimensionality of dependent personality disorder. *Journal of Personality Disorders, 18,* 604–610.

Hare, R. D. (1993). *Without conscience.* New York: Pocket Books.

Hare, R. D. (2003). *Hare Psychopathy Checklist–Revised (PCL-R): Technical manual* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.

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

Jane, J. S., Pagan, J. L., Fiedler, E. R., Turkheimer, E., & Oltmanns, T. F. (2006). The interrater reliability of the Structured Interview for DSM–IV Personality. *Comprehensive Psychiatry, 47,* 368–375.

Kaplan, M. (1983). A woman's view of *DSM–III*. *American Psychologist, 38,* 786–792.

Lindsay, K., & Widiger, T. A. (1995). Sex and gender bias in self-report personality disorder inventories: Items analyses of the MCMI-II, MMPI, and PDQ-R. *Journal of Personality Assessment, 65,* 1–20.

Lykken, D. T. (1995). *The antisocial personalities.* Hillsdale, NJ: Erlbaum.

Mattia, J. I., & Zimmerman, M. (2001). Epidemiology. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (pp. 107–123). New York: Guilford Press.

Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). *Sex differences in antisocial behaviour: Conduct disorder, delinquency and violence in the Dunedin Longitudinal Study.* Cambridge, England: Cambridge University Press.

Morey, L. C., Warner, M. B., & Boggs, C. D. (2002). Gender bias in the personality disorders criteria: An investigation of five bias indicators. *Journal of Psychopathology and Behavioral Assessment, 24,* 55–65.

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

Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R. F. Krueger & J. Tackett (Eds.), *Personality and psychopathology* (pp. 71–111). New York: Guilford Press.

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured Interview for DSM–IV Personality (SIDP–IV).* Washington, DC: American Psychiatric Press.

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

Rienzi, B. M., & Scrams, D. J. (1991). Gender stereotypes for paranoid, antisocial, compulsive, dependent, and histrionic personality disorders. *Psychological Reports, 69,* 976–978.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 35,* 139.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255–270.

Shiner, R. L. (1998). How shall we speak of children's personalities in middle childhood? A preliminary taxonomy. *Psychological Bulletin, 124,* 308–332.

Slavney, P. R. (1984). Histrionic personality and antisocial personality: Caricatures of stereotypes? *Comprehensive Psychiatry, 25,* 129–141.

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28,* 754–763.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75,* 1350–1362.

Sprock, J., Crosby, J. P., & Nielsen, B. A. (2001). Effects of sex and sex roles on the perceived maladaptiveness of DSM–IV personality disorder symptoms. *Journal of Personality Disorders, 15,* 41–59.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–618.

Thissen, D. (1991). MULTILOG (Version 6) user's guide [Computer software and manual]. Mooresville, IN: Scientific Software.

Thomas, C., Turkheimer, E., & Oltmanns, T. F. (2003). Factorial structure of pathological personality as evaluated by peers. *Journal of Abnormal Psychology, 112,* 81–91.

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

Walker, L. E. A. (1994). Are personality disorders gender biased? In S. A. Kirk & S. D. Einbinder (Eds.), *Controversial issues in mental health* (pp. 22–29). New York: Allyn & Bacon.

Weissman, M. M. (1993). The epidemiology of personality disorders: A 1990 update. *Journal of Personality Disorders, 7,* 44–62.

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

Widiger, T. A., Mangine, S., Corbitt, E. M., Ellis, C. G., & Thomas, G. V. (1995). *Personality Disorder Interview–IV: A semistructured interview for the assessment of personality disorders.* Odessa, FL: Psychological Assessment Resources.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science, 7,* 104–109.

Received May 17, 2005
Revision received July 28, 2006
Accepted August 2, 2006 ■