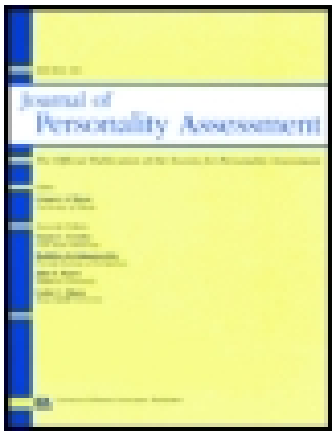

Journal of Personality Assessment
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hjpa20

Comparing the Personality Disorder Interview for DSM-IV (PDI-IV) and SCID-II Borderline Personality Disorder Scales: An Item-Response Theory Analysis

Steven K. Huprich[a], Amy V. Paggeot[a] & Douglas B. Samuel[b]
[a] Department of Psychology, Eastern Michigan University
[b] Department of Psychological Sciences, Purdue University
Published online: 09 Sep 2014.

To cite this article: Steven K. Huprich, Amy V. Paggeot & Douglas B. Samuel (2014): Comparing the Personality Disorder Interview for DSM-IV (PDI-IV) and SCID-II Borderline Personality Disorder Scales: An Item-Response Theory Analysis, Journal of Personality Assessment

To link to this article: http://dx.doi.org/10.1080/00223891.2014.946606

Journal of Personality Assessment, 1–9, 2014

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2014.946606

Routledge
Taylor & Francis Group

Comparing the Personality Disorder Interview for *DSM–IV* (PDI–IV) and SCID–II Borderline Personality Disorder Scales: An Item–Response Theory Analysis

STEVEN K. HUPRICH,[1] AMY V. PAGGEOT,[1] AND DOUGLAS B. SAMUEL[2]

[1]*Department of Psychology, Eastern Michigan University*
[2]*Department of Psychological Sciences, Purdue University*

One-hundred sixty-nine psychiatric outpatients and 171undergraduate students were assessed with the Personality Disorder Interview–IV (PDI–IV; Widiger, Mangine, Corbitt, Ellis, & Thomas, 1995) and the Structured Clinical Interview for *DSM–IV* Axis II disorders (SCID–II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997) for borderline personality disorder (BPD). Eighty individuals met PDI–IV BPD criteria, whereas 34 met SCID–II BPD criteria. Dimensional ratings of both measures were highly intercorrelated (*r*s = .78, .75), and item-level interrater reliability fell in the good to excellent range. An item–response theory analysis was performed to investigate whether properties of the items from each interview could help understand these differences. The limited agreement seemed to be explained by differences in the response options across the two interviews. We found that suicidal behavior was among the most discriminating criteria on both instruments, whereas dissociation and difficulty controlling anger had the 2 lowest alpha parameter values. Finally, those meeting BPD criteria on both interviews had higher levels of anxiety, depression, and more impairments in object relations than those meeting criteria on just the PDI–IV. These findings suggest that the choice of measure has a notable effect on the obtained diagnostic prevalence and the level of BPD severity that is detected.

There are presently a number of diagnostic interviews available for the assessment of borderline personality disorder (BPD). Arguably, some of the more popular of these are the Diagnostic Interview for Borderline Personality (DIB; Zanarini, Gunderson, Frankenburg, & Chauncey, 1989), the Structured Interview for *DSM–IV* Personality Disorders (SCID–II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997), and the Structured Interview for *DSM–IV* Personality (SIDP–IV; Pfohl, Blum, & Zimmerman, 1995). Each of these has extensively documented reliability and validity, but there is one measure that has not been evaluated extensively for its ability to assess BPD–the Personality Disorder Interview for *DSM–IV* (PDI–IV; Widiger, Mangine, Corbitt, Ellis, & Thomas, 1995). The PDI–IV is a semistructured interview for the assessment of BPD according to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [*DSM–IV*]; American Psychiatric Association, 1994) criteria and is a revision of earlier instruments that assessed *Diagnostic and Statistical Manual of Mental Disorders* [3rd ed. [*DSM–III;* American Psychiatric Association, 1980) and *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., rev. [*DSM–III–R*]; American Psychiatric Association, 1987) personality disorders. Although anchored in the polythetic, *DSM* format, the PDI–IV has two to six items for each *DSM–IV* BPD criterion, thereby offering a more expansive set of questions with which to assess each BPD criterion, which is scored by the interviewer on a scale ranging from 0 to 2. As noted in the test manual, "A rating of 1 is virtually equivalent to and usually synonymous with the presence of the respective DSM–IV criterion," and a rating of 2 indicates that the symptom is "present to a more severe or substantial degree" (p. 3; Widiger et al., 1995), thus allowing clinicians to report more extreme levels of each criterion within a somewhat dimensional framework. Thus, an individual receiving a score of 1 or 2 on the set of questions related to that criterion meets the threshold for that criterion. Although an individual might possess a higher or more extreme manifestation of a given criterion with a score of 2 versus 1, the PDI–IV does not differentiate at the subthreshold level. A BPD diagnosis is thus assigned when five of the nine BPD criteria are met, with either a score of 1 or 2 for the criterion. Furthermore, the PDI–IV specifies that individuals answer items about qualities that have been present since their young adulthood and throughout much of their adult life, whereas this is not explicit with other semistructured interviews.

Some psychometric data have been reported for the PDI–IV. Trull, Widiger, Lynam, and Costa (2003) administered the PDI–IV BPD scale to 52 psychiatric outpatients and obtained strong interrater reliability ($\kappa = .84$), as well as moderate levels of convergence with related scales (*r*s ranging between .41–.53). Yang et al. (2000) found that the PDI–IV was only modestly correlated ($r = .11$) with the self-report Personality Diagnostic Questionnaire (PDQ–4; Hyler, 1994). However, a follow-up by the same research team found that the PDI–IV borderline scale was significantly correlated with all of the NEO Revised Personality Inventory (NEO PI–R) facet scales that were hypothesized (Yang et al., 2002). Widiger and Boyd (2009) reported additional convergent validity data for earlier versions of the PDI BPD scale.

By contrast, one of the most popular measures to assess personality disorders has been the SCID–II (First et al., 1997).

Received December 12, 2013; Revised April 29, 2014.

Steven K. Huprich is now at Wichita State University.

Address correspondence to Steven K. Huprich, Department of Psychology, Wichita State University, 1845 Fairmount, Box 34, Wichita, KS 67260; E-mail: steven.huprich@wichita.edu

This measure's popularity is likely due to the fact that it is based exclusively on the *DSM–IV* diagnostic criteria for each PD, with its wording being identical to what is listed in the *DSM–IV*. Rogers (2001) evaluated early studies with the SCID–II and reported that obsessive–compulsive personality disorder had a low rate of diagnostic agreement ($\kappa = 0.24$) and that temporal consistency and self–other agreement were marginally acceptable. Many have found the SCID–II to be useful, in that it can be used in conjunction with the SCID–II Questionnaire (First et al., 1997) to screen individuals who meet the threshold level of personality disorder criteria on the self-report who can then be assessed via the interview for the presence or absence of a personality disorder. It does not appear many psychometric studies for the SCID–II have been reported since then (see http://www.scid4.org/psychometric/scidII_reliability.html, searched April 22, 2014), although Lobbestael, Leurgans, and Arntz (2010) found that intraclass correlation coefficient values for total scores ranged between .60 (Schizotypal) and .95 (Borderline).

Assessment psychologists recognize that two measures of the same construct that use the same methodology (e.g., semi-structured diagnostic interviews) are expected to correlate strongly with each other. However, when they do not correlate as strongly as anticipated, this could be the result of several factors, such as the manner by which the construct is defined and assessed within a given instrument (e.g., Skodol, Oldham, Rosnick, Kellman, & Hyler, 1991; Skodol, Rosnick, Kellman, Oldham, & Hyler, 1988). More generally speaking, any measure evaluated within the framework of classical test theory recognizes that respondent and scale characteristics are interrelated, thereby making it challenging to interpret the findings beyond that of the sample being evaluated (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985).

Within the past few decades, however, there has been an interest in measure development from the perspective of item–response theory (IRT). IRT (also known as latent trait theory) and the associated analyses differ markedly from classical test theory in that IRT focuses on properties of items, rather than tests (Embretson & Reise, 2000). IRT analyses proceed by aligning items on a latent dimensional trait and estimating how much psychometric information an item provides about the trait using two parameters: alpha and beta. *Alpha,* referred to as the slope or discrimination parameter, corresponds to the item's ability to differentiate among individuals at a given level of the latent trait and can indicate how strong an indicator that item is for assessing the underlying trait. *Beta* corresponds to the level of the latent trait that is required for an individual to endorse a given response with a 50% probability. Within intellectual assessment, beta is often analogized as the item's difficulty, but within personality and psychopathology assessment it might more accurately be referred to as extremity or severity (Simms et al., 2011).

An important product of IRT analyses is the ability to compare items in terms of their provision of information along the latent trait. For instance, Feske, Kirisci, Tarter, and Pilkonis (2007) used IRT to examine the diagnostic criteria for *DSM–III–R* BPD that were assigned using a conference of judges with available interview information. They found that the criteria had comparable alpha parameters, suggesting they were relatively equivalent in terms of their ability to discriminate among levels of the BPD construct. However, they also found that the items displayed more variation in terms of where they provided that information. For example, whereas the criterion assessing affective instability provided information at moderate levels of the construct, the suicidal behavior criterion was notably more extreme. More recently, Samuel, Carroll, Rounsaville, and Ball (2013) used IRT to determine whether the diagnostic criteria for BPD and Five-factor model (FFM) Neuroticism could be fit along a single latent dimension. They found that borderline criteria assessed the shared latent trait with Neuroticism, but at a level that was more extreme ($d = 1.11$).

In this study, we sought to evaluate the psychometric properties of the PDI–IV in a clinical sample within the framework of both classical test theory and IRT. We report the reliability of the measure, as well as its correlation with another semi-structured diagnostic interview of BPD, the SCID–II (First et al., 1997), arguably considered the gold standard for assessing *DSM* personality disorders. We then report the diagnostic frequencies between measures and perform an IRT analysis of both the PDI–IV and SCID–II to investigate the item parameters from both measures.

Method

Participants and Procedures

Participants were recruited from two populations across five locations. Nonclinical, undergraduate students were recruited from psychology courses at a Midwestern university between 2006 and 2008. For their participation, they received extra course credit. Psychiatric outpatients were recruited from four locations between 2007 and 2009: a university-based psychology clinic, a hospital-based outpatient behavioral health treatment facility, and two offices of a community mental health center. The authors' institutional review board, as well as the review boards at each of the clinical locations from which patients were recruited, approved this study.

After signing up to participate, undergraduates were contacted by one of the interviewers and arranged a time to participate that was mutually convenient. All interviews and data collection of undergraduate participants occurred in the psychology department of the university. At the clinical sites, therapists were informed about the study and provided an information sheet to give to potential participants. To be included in the study, participants had to be at least 18 years old and could not be actively psychotic or have a primary diagnosis of schizophrenia, schizophreniform, major depression with psychosis, bipolar disorder with psychotic features, or psychotic disorder not otherwise specified. Their mental status had to be intact, according to their therapist, and they could not have an organic or medical condition that accounted for their diagnosis. They also could not be actively using or abusing substances, nor in such acute psychological distress that their therapists believed answering questions about their mood, thoughts, relationships, and personality would evoke a strong negative emotional reaction. In other words, based on the ethical principle of nonmaleficence, participants were ruled out by their therapist if it was believed that asking such questions would destabilize their already fragile mood state. Although 192 patients were referred and participated in the study, a review of the participants identified 23 individuals

who either had an exclusionary clinician-assigned diagnosis ($n = 21$; despite the fact that clinicians were informed of these criteria prior to referring patients to the study), refused to do the interview when asked ($n = 1$), or became upset during the interview and wished to discontinue ($n = 1$), Hence, their data were excluded from these analyses.

After being informed about the study, potential participants contacted the principal investigator by phone or email expressing their interest. They were referred to one of three doctoral student research assistants who did an initial phone screening with the participant and scheduled them at their respective treatment facility or within the department at a time that was mutually convenient. At one of the community mental health center locations, a staff member coordinated patient scheduling, as it was more feasible to manage the scheduling in the center that way. Study participation took approximately 2 hr and involved administration of diagnostic questionnaires, as well as completion of a set of self-report questionnaires. For their participation, clinical participants received $75 cash. Undergraduate participants were granted extra credit by their instructors in an undergraduate psychology course.

All participants were interviewed using the BPD modules from the PDI–IV (Widiger et al., 1995) and the SCID–II (First et al., 1997). Interviewers included eight graduate students who had completed course work in personality theories and psychopathology. They were trained by the first author over the course of 5 hr and then practiced with each other for another 5 to 10 hr. In this initial part of the training, interviewers assumed the role of both an interviewer and interviewee. Except for those interviewers who were involved in the earliest phases of the interview, all interviewers also listened to tapes of previous interviews individually and collectively and compared their scores with each other, discussing differences when appropriate. Prior to interviewing participants, all interviewers had to obtain at least a 90% agreement on all of the interview items from a previously recorded interview conducted by a different interviewer. To minimize interviewer drift, every 6 months interviewers listened to completed taped interviews, scored the interviews, and compared answers to ensure that they were consistently scoring the participant responses. Discrepancies were resolved by discussing the inconsistently scored items and consulting with the first author as needed. As reported later, reliability values were satisfactory.

Measures

Beck Anxiety Inventory. The Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988) is the most popular and widely used, well-researched inventory measuring the severity of anxiety symptoms that are minimally shared with symptoms of depression. The BAI consists of 21 self-report items on a 4-point Likert scale on which respondents report how much they have been bothered by a list of anxiety symptoms during the past week. The validity and reliability of the BAI has been extensively examined. The BAI has excellent internal consistency ($\alpha = .92$) and solid, 1-week test–retest reliability ($r = .75$), as well as good discrimination from experiences of depression. Cronbach's alpha of the BAI in this study was .94.

Beck Depression Inventory–II. The Beck Depression Inventory (BDI) and BDI–II (Beck, Steer, & Brown, 1996). are the most commonly used measure of the severity of depressive symptoms. The BDI–II assesses multiple aspects of depression, including cognitive, emotional, behavioral, and physical domains. The BDI–II consists of 21 self-report items on a 4-point Likert scale. It has been well examined and found to have good validity and reliability. Cronbach's alpha of the BDI in this study was .95.

Bell Object Relations and Reality Testing Inventory–Form O. The Bell Object Relations and Reality Testing Inventory–Form O (BORRTI; Bell, 1995) is a 45-item true–false questionnaire that is designed to measure four dimensions of an individual's object relations: egocentricity, social introversion, alienation, and insecure attachment. The BORRTI has been found to have excellent reliability (Cronbach's α ranging from .78–.90; split-half reliability ranging from .78–.90; and 26-week test–retest reliability among schizophrenics ranging from .58–.72) and has extensive research supporting its validity (Bell, 1995). This measure was given only to the clinical sample. Because of the complex scoring algorithm, internal consistencies in the sample for this study could not be computed.

PDI–IV. As noted earlier, the PDI–IV (Widiger et al., 1995) Borderline scale is a nine-item structured clinical interview where each of the nine *DSM–IV* criterion is assessed with a set of three to four open-ended questions. Each criterion is scored by the interviewer on a 3-point scale from 0 to 2, where 0 indicates the absence of a criterion, 1 indicates the presence of a criterion, and 2 indicates the respondent exceeds the criterion. In this study, the scale had a Cronbach's alpha value of .85. Fifty-nine interviews were selected for interrater reliability analysis, where five of the original interviewers reviewed the taped interviews and recoded each item. Intraclass correlation values calculated using two-way mixed single measure modeling for each criterion in this sample ranged from .68 to .97 and averaged .82.

SCID–II. The SCID–II (First et al., 1995) Borderline scale is a nine-item structured clinical interview on which the individual is asked a question or series of questions about each item and rated on a scale of 1 to 3 based on their responses, where 1 indicates that the criterion is absent, 2 indicates the criterion is present but is subthreshold, and 3 indicates that the criterion is present at the threshold. Cronbach's alpha was calculated at .85 for this sample. Five of the original interviewers participated in the reliability analysis by reviewing the taped interviews and providing a second set of ratings, which was conducted at the conclusion of the data collection. As earlier, 59 interviews from both samples were randomly selected for interrater reliability analysis and coded at the item level. Intraclass correlation values for each criterion ranged from .75 to .95, and averaged .86.

Data Analysis

A fundamental assumption underlying IRT is that the items form a unidimensional latent construct. Stout (1990) argued that what is required for IRT is not the absence of any subfactors, but the presence of a single, dominant factor. Thus, we sought to demonstrate that the underlying trait was essentially

unidimensional, meaning that a broad, general dimension underlies all BPD items. In this study we conducted a confirmatory factor analysis (CFA) using M*plus* 7.0 (Muthén & Muthén, 1998–2012), employing the weighted least squares with mean and variance adjustment estimator. We used several fit indexes to determine the adequacy of this one-factor solution. These included the comparative fit index (CFI) and Tucker–Lewis index (TLI), with values above .90 and .95 indicating acceptable and excellent fit, respectively (Hu & Bentler, 1999). We also used the root mean square error of approximation (RMSEA) with values lower than .080 and .050 indicating close and reasonable fit, respectively, and the weighted root mean square residual (WRMR) where values below .90 indicate good fit.

We chose to estimate item parameters separately for the PDI–IV and SCID–II because of the concern that a joint analysis would violate the assumption of local independence of items. Nonetheless, to investigate its impact we subsequently reanalyzed the data modeling the 18 combined criteria from PDI–IV and SCID–II scales simultaneously. In doing so, we allowed the error terms for matched criteria to correlate in an attempt to explicitly model this local dependence. The IRT parameter estimates using this method were indistinguishable from those modeling the instruments separately, with only minor variations of a few hundredths of a decimal point (full results are available by contacting the third author) so we elected to retain our original analyses.

All IRT parameters were estimated using Samejima's (1969) graded response model (GRM) because both interviews use 3-point Likert scales. The GRM is an extension of the two-parameter logistic model for polytomous items and is commonly used for IRT analyses of personality disorder criterion sets (e.g., Feske et al., 2007). When using the GRM with three response options, three parameters are estimated. The first is the alpha parameter, which we have described previously, and the second and third parameters are labeled beta1 and beta2. Because both the PDI–IV and SCID–II use three response options, there are two beta values indicating the level of the latent trait necessary to endorse the higher response option over the lower one with a 50% probability. For example, beta2 indicates the level of the latent trait necessary to receive a score of threshold versus subthreshold on the SCID–II, whereas beta1 indicates the level needed to score a subthreshold versus absent. The values for the beta parameters are displayed in terms of theta, but can be analogized to z scores, with higher values indicating that an individual needs a higher level of the latent trait to endorse the item affirmatively. All parameters reported here were estimated using IRTPRO 2.1 (Scientific Software International, 2011).

Results

The sample consisted of a total 340 participants, 169 of whom were psychiatric outpatients and 171 of whom were undergraduate students. The combined sample had a mean age of 31.61 ($SD = 15.07$) years, and consisted of 234 female participants and 97 male participants. Nine individuals did not report their gender, all of whom were in the clinical sample. Within the undergraduate sample, participants ranged in age from 18 to 59, with a mean of 21.91 ($SD = 6.23$) years. There were 124 female and 47 male participants. Participants identified themselves as White (62%, $n = 235$), African American (24%, $n = 60$), Asian (6%, $n = 10$), Hispanic (3%, $n = 11$), Middle Eastern (2%, $n = 4$), and other (4%, $n = 16$). One person did not report his or her ethnic heritage. The majority of the undergraduate sample designated their relationship status as single (80%), although 10% were cohabitating, 7% were married, and 4% were divorced or separated.

Within the outpatient subsample, 37.9% were recruited from two community mental health facilities ($n = 64$), 21.3% were recruited from a university psychology clinic ($n = 36$), and 40.8% from an outpatient behavioral health facility associated with a hospital ($n = 69$). Participants ranged in age from 18 to 76, with a mean of 42.08 ($SD = 14.83$) years. There were 110 female and 50 male participants. Participants identified themselves as White (77.7%, $n = 129$), African American (12.0%, $n = 20$), Middle Eastern (0.6%, $n = 1$), Hispanic (3.6%, $n = 6$), and other (6.0%, $n = 10$). Three individuals did not report their ethnic heritage. The majority of the sample designated their relationship status as single (40.2%, $n = 68$), although 24.3% ($n = 41$) were married, 24.3% ($n = 41$) were divorced or separated, 8.3% ($n = 14$) were cohabitating with a partner, and 3.0% ($n = 5$) were widowed. We also obtained clinician-assigned, Axis I and II *DSM–IV* diagnoses, presented in Table 1.

In the clinical sample, total BPD scores on both interviews were highly and significantly correlated ($r = .78$, $p < .001$). Sixty-seven individuals met diagnostic criteria for BPD on the PDI–IV, whereas only 32 met criteria on the SCID–II. All who met the BPD cutoff criterion on the SCID–II were also diagnosed with BPD on the PDI–IV, meaning that the PDI–IV identified 35 individuals as having a BPD diagnosis who the SCID–II did not identify. This yielded an unweighted kappa of .53 between interviews. The frequency of criterion endorsement and percentages of individuals meeting threshold values for each criterion for the SCID–II

Table 1.—Comorbid diagnoses of the clinical sample.

Axis I	
Adjustment disorders	11
ADHD–inattentive	2
Academic problem	1
Alcohol dependence	1
Bipolar disorders	35
Bulimia nervosa	1
Cocaine dependence	1
Depressive/mood disorder NOS	10
Dysthymic disorder	16
Generalized anxiety disorder	3
Impulse control disorders	1
Major depressive disorders	1
Opioid dependence	1
Partner relational problem	1
PTSD	3
Social phobia	2
Axis II	
Antisocial personality disorder	2
Avoidant personality disorder	1
Borderline personality disorder	16
Deferred	15
Dependent personality disorder	3
Not otherwise specified	6
Schizotypal personality disorder	1

Note. $n = 169$. ADHD = attention deficit hyperactivity disorder; NOS = not otherwise specified; PTSD = posttraumatic stress disorder.

TABLE 2.—Frequency and percentage of individuals meeting each scoring criterion and diagnostic thresholds.

	SCID–II				PDI–IV			
Criterion	1	2	3	%Meet	0	1	2	%Meet
Fears of abandonment	251	52	37	10.9	287	39	14	15.6
Unstable relationships	222	48	70	20.6	235	85	20	30.9
Identity problems	265	53	22	6.5	267	45	28	21.5
Impulsivity	200	80	60	17.6	218	72	50	35.9
Suicidal behavior	244	38	58	17.1	245	57	38	27.9
Affective instability	212	46	82	24.1	202	109	29	40.6
Chronic emptiness	230	32	78	22.9	232	68	40	31.8
Difficulty controlling anger	229	68	43	12.6	236	77	27	30.6
Dissociation	247	56	37	10.9	242	71	27	28.8
Meet diagnostic threshold	34 (10%)				80 (20.6%)			

Note. These values are computed from the entire sample of 340 participants. SCID–II = Structured Clinical Interview for *DSM–IV* Axis II Disorders; PDI–IV = Personality Disorder Interview for *DSM–IV*.

and PDI–IV are presented in Table 2. Sixteen individuals were assigned a BPD diagnosis by their clinician (see Table 1); of these, all met the diagnostic threshold criteria for a BPD diagnosis on the SCID–II and PDI–IV.

In the nonclinical sample, total BPD scores on both interviews were highly and significantly correlated ($r = .75$, $p < .001$). Thirteen individuals met diagnostic criteria for BPD on the PDI–IV, whereas only 2 met criteria on the SCID–II. As with the clinical sample, all who met the BPD criterion threshold on the SCID–II were also diagnosed with BPD on the PDI–IV, meaning that the PDI–IV identified 11 individuals as having a BPD diagnosis that the SCID–II did not. This yielded an unweighted kappa coefficient of .25 between interviews (see Table 2).

Collectively, the total BPD scores on both interviews for all participants were highly and significantly correlated ($r = .82$, $p < .001$). Kappa values among the 59 interviews selected for reliability yielded values of .72 (PDI–IV) and .74 (SCID–II).

Unidimensionality

We conducted two CFAs to determine if the data from each measure met the assumption of essential unidimensionality and were appropriate for IRT analyses. The nine borderline items from the PDI–IV were fit to a one-factor model within a separate analysis. The resulting fit indexes were $\chi^2(27) = 48.2$, CFI = .99, TLI = .98, RMSEA = .048, and WRMR = .74. The nine items from the SCID–II BPD scale were fit to a one-factor model, and the resulting fit indexes were $\chi^2(27) = 34.7$, CFI = .99, TLI = .99, RMSEA = .029, and WRMR = .59. These values were quite comparable to those from the PDI–IV and again suggested excellent fit for a one-factor model, supporting the assumption of essential unidimensionality.

IRT Parameter Estimates

Table 3 presents the alpha and beta parameters for all BPD criteria from the PDI–IV and SCID–II items, respectively. The alpha parameter values for the PDI–IV criteria ranged considerably. The difficulty controlling anger criterion again obtained the lowest value (1.53), whereas the suicidal behavior criterion had the highest (2.41), suggesting it was the best item for discriminating among individuals. The beta1 parameters from the PDI–IV ranged from a low of .35 (affective instability) to 1.34 (fear of abandonment), with a mean of .75. The beta2 values ranged from 1.41 (impulsivity) to 2.32 (fear of abandonment) with a mean of 1.86.

For the SCID–II criteria, the alpha values were all above 1.0, suggesting that each provided substantial information about the latent trait. However, these values ranged considerably, with affective instability and the suicidal behavior criteria obtaining the highest values, suggesting they were best able to differentiate individuals with different levels of the latent trait and could be said to be the best indicators of the BPD construct. The criterion concerning difficulty controlling anger was least able to discriminate among individuals. The SCID–II beta1 parameters ranged from 0.32 (impulsivity) to 1.06 (identity problems), with a mean value of 0.66. The beta2 parameters ranged from 0.89 to 2.13, with a mean of 1.43. The standard errors for all parameters were relatively low, suggesting reasonably robust estimates.

Cross-Instrument Comparison

Although they were calculated in separate analyses, we also compared these two semistructured diagnostic interviews on the IRT parameters. We first evaluated for consistency across the two instruments in the patterns of parameter estimates. For example, the alpha parameter values were largely consistent in terms of rank-order across the two interviews, as evinced by a high correlation ($r = .71$) between them. The suicidal behavior criterion had among the highest alpha values on both instruments, whereas the difficulty controlling anger criterion was the lowest. We also compared the alpha parameter values across the two instruments using a dependent samples t test and found that they were not significantly different, $t(8) = -1.4$, $p = .19$.

There were more notable differences, however, for the parameter estimates for beta1 and beta2 across the two interviews. The rank-order consistency for beta1 criteria was still quite high ($r = .79$), and the identity problems criterion was among the most extreme item on both measures. In addition, the affective instability and impulsivity criteria had the lowest beta value on both instruments. Similarly, a dependent samples t test indicated that the beta1 values for the two BPD interviews were not significantly different, $t(8) = -1.4$, $p = .19$. Nonetheless, it is important to note that although these beta1 values were similar across the instruments, the anchors for these response options are not. Whereas the lowest response option on both the SCID–II and PDI–IV indicates

TABLE 3.—Means, standard deviations, intraclass correlations, and item–response theory parameter estimates for borderline criteria from the PDI–IV and SCID–II.

	SCID–II									PDI–IV								
BPD criterion	*M*	*SD*	ICC	Alpha	*SE*	b1	*SE*	b2	*SE*	*M*	*SD*	ICC	Alpha	*SE*	b1	*SE*	b2	*SE*
Fears of abandonment	1.37	.67	.77	1.98	.28	.83	.10	1.64	.16	.20	.49	.73	2.06	.33	1.34	.14	2.32	.25
Unstable relationships	1.55	.81	.80	1.99	.27	.52	.09	1.10	.12	.37	.59	.84	1.86	.26	.67	.10	2.15	.22
Identity problems	1.29	.58	.96	1.84	.27	1.06	.12	2.13	.22	.30	.61	.94	2.10	.31	1.02	.11	1.80	.18
Impulsivity	1.59	.77	.92	1.72	.22	.32	.09	1.32	.14	.51	.74	.79	1.90	.26	.48	.09	1.41	.15
Suicidal behavior	1.45	.77	.95	2.08	.29	.75	.10	1.24	.13	.39	.68	1.0	2.41	.34	.74	.09	1.51	.14
Affective instability	1.62	.85	.87	2.11	.29	.39	.09	.89	.11	.49	.65	.64	1.85	.24	.35	.09	1.90	.18
Chronic emptiness	1.55	.84	.93	1.95	.28	.61	.09	.99	.11	.44	.70	.98	2.11	.29	.62	.09	1.53	.15
Difficulty controlling anger	1.45	.71	.70	1.42	.20	.68	.11	1.78	.20	.39	.63	.68	1.53	.22	.73	.11	2.11	.24
Dissociation	1.38	.68	.79	1.65	.23	.82	.11	1.76	.18	.37	.63	.64	1.72	.24	.79	.11	2.00	.21

Note. These values are computed from the entire sample of 340 participants. PDI–IV = Personality Disorder Interview for *DSM–IV*; SCID–II = Structured Clinical Interview for *DSM–IV* Axis II Disorders; BPD = borderline personality disorder; ICC = intraclass correlation coefficient.

absence of the criterion, the middle option differs. On the SCID–II this score indicates that the individual possesses the criterion at the subthreshold level, whereas the PDI–IV anchor indicates the criterion is present according to the *DSM–IV* definition of the item (i.e., over the threshold). In this way, the highest response for the SCID–II criteria (i.e., threshold *or* true) is verbally most similar to the middle response option from the PDI–IV. The PDI–IV's highest response option (present to a more severe degree) does not have an equivalent on the SCID–II. For instance, the SCID–II item assessing BPD impulsivity is phrased, "Have you often done things impulsively?" This initial question, if answered affirmatively by the interviewee, is followed up with "What kinds of things?" At this point, the interviewer is provided a number of prompts to provide to the interviewee to help clarify scoring. A further note states that a 3 is only to be scored if "several examples indicate a pattern of impulsive behavior (not necessarily limited to examples above)."

However, the PDI–IV opens with an item stating, "Ever spend so much money that you had trouble paying it off?" This is followed up with four other questions asked in series: "Ever go on a drinking or eating binge? Have you ever taken any major chances or risks with drugs? Ever do anything impulsive that was risky or dangerous? Have you ever become sexually involved with someone in a risky or dangerous way?" These items are then assessed on a scale from 0 to 2, with 0 indicating an absence of this criterion; 1 being scored if the interviewer judges there is "Impulsivity in at least two areas that are potentially self-damaging" and (like the SCID–II) restricting this to exclude self-harm or suicidal behavior; and 2 if the interviewer judges that the interviewee has "Impulsivity in at least three areas, at least one of which has been physically self-damaging." Here, a higher rating requires greater impulsivity than on the SCID–II, where a higher rating simply means meeting the full criteria listed in the *DSM–IV–TR.*

This might help to explain why the beta2 values were less similar across the two interviews. These parameter estimates evinced lower rank-order consistency ($r = .31$), with notable variation in specific criteria. For example, the identity problems criterion had the highest beta2 values for the SCID–II, but was lower than all but three of the criteria on the PDI–IV. The beta2 values also differed significantly in terms of their mean value according to dependent samples *t* test. The mean of 1.86 ($SD = .32$) for the PDI–IV was higher than the value of 1.43 ($SD = .42$) for the SCID–II, $t(8) = -2.9$, $p = .02$, $d = -.97$. Finally, because the response anchors for the highest response option on the SCID–II and the middle coding on the PDI–IV were most equivalent (i.e., threshold), we also compared SCID–II beta2 with PDI–IV beta1 ($M = .75$, $SD = .29$) via a dependent samples *t* test. The means for these two parameters were significantly different, $t(8) = 6.5$, $p < .01$, $d = 2.17$. They also correlated significantly ($r = .67$).

Given that the diagnostic prevalence was notably different between measures, we reasoned that, if the intermediate value on the SCID–II (a subthreshold rating of 1) were considered as sufficient for meeting each BPD criterion, prevalence rates might increase. Subsequently we could evaluate the extent to which the measures converged and consider to what extent an intermediate level rating is the implicit metric used to increase interrater agreement. When recoded this way, 93 individuals met criteria for a BPD diagnosis on the SCID–II, compared to only 34 meeting criteria when scoring the SCID–II in the standard fashion. Thus, 59 individuals met criteria using the subthreshold rating system that did not meet the criteria using full threshold values. This yielded an unweighted kappa of .46 between methods of scoring the SCID–II. When comparing this revised SCID–II scoring to the PDI–IV scoring, the kappa was computed at .65.

Because of this substantial difference in diagnostic rates between instruments, we also compared those diagnosed with BPD from only the PDI–IV to those who were diagnosed with BPD on both the PDI–IV and SCID–II with regard to their scores on three other measures that were not part of the original design for this study–the BDI–II, BAI, BORRTI, and Global Assessment of Functioning (GAF) scores for those clinical patients for whom we had this rating. Those who met criteria for BPD on both measures had significantly higher scores on the BDI–II, BAI, BORRTI–Alienation scale, Insecure Attachment scale, and BORRTI–Egocentricity scale. Interestingly, they did not differ on GAF scores. Complete results are presented in Table 4.

DISCUSSION

One of the objectives of this study was to evaluate the reliability and convergent validity of the PDI–IV BPD scale. Overall, the scale was found to have strong interrater reliability and internal consistency that is adequate for use with basic and applied settings (Nunnally & Bernstein, 1994). In

TABLE 4.—Means and standard deviations comparing those diagnosed with BPD on the PDI–IV and SCID–II to those diagnosed with BPD on the PDI–IV only.

	BPD on PDI–IV and SCID–II		BPD on the PDI–IV only				
Measure	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
BDI–II	32.77	13.87	23.97	11.19	3.06	< .01	0.71
BAI	28.06	13.90	14.68	9.70	4.63	< .01	1.27
Alienation	68.34	7.34	62.84	8.81	2.87	< .01	0.67
Insecure attachment	66.85	7.10	58.86	7.57	4.68	< .01	1.09
Egocentricity	64.97	7.46	58.91	8.41	3.34	< .01	0.76
Social incompetence	57.85	8.13	55.65	8.45	1.16	.25	0.26
Global assessment of functioning	49.67	7.09	49.00	6.64	0.39	.70	0.10

Note. These values are computed from the entire sample of 340 participants. Alienation, Insecure Attachment, Egocentricity, and Social incompetence are all subscales of the Bell Object Relations and Reality Testing Inventory–Form O. BPD = borderline personality disorder; PDI–IV = Personality Disorder Interview for *DSM–IV*; SCID–II = Structured Clinical Interview for *DSM–IV* Axis II Disorders; BDI–II = Beck Depression Inventory (2nd ed.); BAI = Beck Anxiety Inventory.

addition, we found that both interviews (independently and collectively) evinced strong unidimensionality on the basis of their fit to a one-factor CFA model. Nonetheless, despite the ostensible similarity among the measures, we noted that the PDI–IV identified more than double the number of individuals meeting BPD diagnostic criteria than did the SCID–II (80 vs. 34). This discrepancy was quite striking and yielded kappa values that were lower than we anticipated. Although past research has routinely indicated relatively poor agreement between independent personality disorder interviews (e.g., Skodol et al., 1991), the categorical agreement was notably low, particularly considering how highly their total dimensional scores correlated with each other (i.e., $r > .75$).

As one means by which to probe this difference, we evaluated the levels of symptoms (i.e., anxiety and depression) and object relations endorsed by those who met criteria by each interview. Given that everyone who was diagnosed by the SCID–II was also diagnosed by the PDI–IV, it is perhaps not surprising we found that those meeting criteria on both interviews were more depressed, more anxious, and had more disturbances in object relatedness than those just meeting criteria on the PDI–IV. This suggests the diagnostic threshold on the SCID–II is associated with greater levels of negative affect and disturbances in their capacity to be interpersonally related than those diagnosed by the PID–IV. Interestingly, GAF scores between groups did not significantly differ. As GAF scores were assigned at the beginning of patients' treatments, it is possible that they no longer accurately reflected patients' level of functioning at the time of the interviews. Thus, although we cannot determine which interview was more accurate (absent any gold standard) this does raise important questions about the variation in diagnostic threshold across two different semistructured interviews. Namely, what is clear from these findings is that it is relatively more difficult to meet criteria on the SCID–II than the PDI–IV. This suggests that the choice of interview measure would likely have a substantial effect on the overall prevalence of diagnosis within a sample and the impairment associated with the diagnosis.

We also used IRT analyses to more thoroughly compare the two interviews and understand the reasons for the vast differences in frequency. It appears that a primary reason is that the individual items and criterion ratings on the two interviews are scaled differently. Although both have three response options, with the lowest indicating an absence of that symptom, the middle option represents subthreshold on the SCID–II but at threshold on the PDI–IV; the highest option represents threshold on the SCID–II and above threshold on the PDI–IV. In essence, the SCID–II provides finer psychometric precision in the subthreshold range, whereas the PDI–IV is able to differentiate among individuals who are above the diagnostic threshold. To consider the effect of removing subthreshold criteria, we recoded the SCID–II such that subthreshold was considered indicative of the diagnosis and found a sizable increase in the number of individuals who met criteria (93), a number much closer to that observed (80) on the PDI–IV. A comparison of the beta (difficulty) parameters also revealed the same pattern of results. Whereas the beta1 rank order correlation was high between measures, the beta2 rank order correlation was not (*r*s of .79 and .31, respectively). Specifically, it required a similar level of the latent trait to receive the middle response option on the SCID–II (i.e., subthreshold) as it did to receive the middle score on the PDI–IV (i.e., threshold), even though those scores are intended to have different meanings. Stated differently, the SCID–II subthreshold ratings (1 vs. 2) generally corresponded to the PDI–IV criterion ratings (0 vs. 1). Such findings confirm what we identified earlier–namely, that the SCID–II has a higher threshold for diagnosing BPD, and is better suited at identifying more severe levels of BPD. The PDI–IV, however, is better suited at identifying less severe, although still clinically significant, levels of BPD.

It should be noted here that both interviews yielded acceptable and comparable levels of internal consistency and interrater agreement. Moreover, interviewers attended regular meetings in which they discussed the interviews and difficulties they were experiencing. Thus, it does not appear that interviewers ignored the provided response anchors and merely rated items on both the same implicit 3-point scale. Specifically, the beta2 parameters on the PDI–IV were significantly higher than those for the SCID–II, which is consistent with the coding instructions for the highest response option on the PDI–IV indicating a greater level of severity above threshold. One possible explanation for this discrepancy in diagnostic frequency is a difference in the number of probes for each criterion across the two interviews. Whereas the PDI–IV asks patients two to six questions per diagnostic criterion, the SCID–II includes one item per criterion, with most allowing the option of follow-up questions if needed to confirm the presence of the diagnostic criterion. Thus, we hypothesize that the PDI–IV has an advantage at detecting the presence or absence of each criterion by providing more opportunities to uncover relevant behaviors that push the individual into the diagnostic range.

In any event, our results make clear that even when two interviews achieve total scores that agree highly, their categorical frequencies can vary quite substantially. This has important implications for the assessment of personality disorders as it suggests the instrument chosen can have notable, and perhaps unanticipated, consequences on the prevalence and cooccurrence of personality disorders. Further research would be quite helpful in understanding the causes of these discrepancies. One potential manipulation of interest would be to vary the response anchors or Likert-type scales within the same interview to determine what impact these have on each item's difficulty within an IRT framework. A second intriguing possibility would be to compare the number of specific probes within a criteria to test our hypothesis that asking more questions leads to higher rates of diagnosis. However, given that these interviews are semistructured and interviewers are allowed to ask follow-up questions as needed to make a decision, the number of follow-up probes is variable and not predictable, thus making this option not viable. Although such research is challenging based on the difficulty of collecting extensive interview data, the retention of the existing diagnostic categories in *DSM–5* creates a pressing need to improve the diagnostic options available.

Comparison of Individual Diagnostic Criterion

In addition to the comparison of the psychometric properties of the two BPD interviews, our data also provide a unique opportunity to examine some of the IRT parameters associated with individual diagnostic criterion of BPD. In general, suicidal behavior was the criterion that across both measures most strongly differentiated levels of the latent construct, whereas dissociation and difficulty controlling anger had generally lower alpha values, suggesting they were least effective for discriminating among levels of BPD. In some ways, it is not surprising that suicidal ideation and behavior is the strongest indicator of BPD, although this does not suggest that suicidal ideation or behavior is specific to the identification of BPD (e.g., Joyce, Light, Rowe, Cloninger, & Kennedy, 2010).

The collective evaluation of the alpha and beta values for each interview criterion set and the revised diagnostic thresholds with the SCID–II highlight the fact that each diagnostic criterion does not have equal ability to detect the latent BPD construct, and in fact that these criteria might vary in their ability to detect BPD across diagnostic interviews. For example, on the SCID–II, affective instability had the highest alpha value, whereas it was seventh in the rank ordering of PDI–IV symptoms, suggesting a lack of clarity in regard to how effective this criterion is at discriminating levels of BPD. Nonetheless, this criterion obtained the lowest beta1 and beta2 values on both measures. This indicates that affective stability consistently requires a lower level of the BPD latent construct to endorse across both measures. This could suggest that it is the least "severe" of the BPD criteria, or perhaps that other criteria are built on the top of a general tendency toward affective dysregulation. As noted by Cooper, Balsis, and Zimmerman (2010), these unequal beta values might suggest that certain combinations of criteria are more severe than others. Two individuals might meet the same number of criteria required for a BPD diagnosis, but if one individual is meeting criteria with lower beta values (e.g. affective instability) whereas the other meets the same number of criteria but with higher beta values, simply looking at the number of criteria met might be misleading with regard to the overall severity of the disorder within each individual.

The remainder of the items (unstable relationships, fears of abandonment, emptiness, identity problems, impulsivity) tended to fall in the midrange in terms of their rank order, although there was not much consistency in ranking relative to each other. In the absence of significance testing for alpha and beta values, it is difficult to know how to describe these findings. There have been two prior studies that have examined BPD criteria using IRT analyses (Feske et al., 2007; Samuel et al., 2013) that can provide some context for understanding the results reported here. Nonetheless, methodological differences across BPD IRT studies make a comparison of findings challenging. For instance, this study assessed both clinical and nonclinical undergraduate samples, whereas Samuel et al. (2013) evaluated patients being treated specifically for substance abuse disorders. Samuel and colleagues (2013) also dichotomized SCID–II responses to conform to a yes–no format consistent with the *DSM–IV*'s dichotomous approach to criterion assessment. Finally, they did not include the unstable relationships and impulsivity BPD criteria, in part to evaluate the BPD criteria relative to the NEO Five-Factor Inventory (Costa & McCrae, 1992) items that were most directly associated with the BPD construct. Hence, the constructs being assessed in the Samuel and colleagues study were different than the ones we assessed in this study. These notable differences notwithstanding, they found that beta values for the SCID–II ranged between .30 (chronic emptiness) and 1.84 (recurrent suicidality). Similarly, in a study of *DSM–III–R* BPD criteria with a methodology more similar to the present effort, Feske et al. (2007) found that the suicidal behavior criterion was among the most extreme beta values (only surpassed by abandonment). Why suicidality appeared more severe in these samples compared to ours is difficult to determine. With the sampling, scaling, and assessment tool differences across studies, it might not be possible to know with much certainty. Clearly, future IRT studies that wish to compare across studies will need to consider these issues more carefully to more collectively integrate findings.

This study has several strengths. It included both nonpatients and psychiatric outpatients at several locations. All participants were assessed by trained interviewers with reliable and valid instruments, thus enhancing the ecological validity of these findings. We also were able to compare two measures of BPD, with particular interest in the psychometric properties of a relatively less studied measure, the PDI–IV. In doing so, we identified features of each measure that are associated with the assessment of BPD, and provided psychometric evidence of the reliability and convergent validity of the PDI–IV.

There also are notable limitations to our study. First, our sample was relatively small compared to what is desired for IRT analyses (Embretson & Reise, 2000); however, given the challenges of collecting interview data in clinical settings, along with our inclusion of nonpatients, this sample size is relatively favorable. Second, there are many other BPD interviews, some of which are more widely used than the PDI–IV. Thus, we do not know to what extent our findings are generalizable to other measures. Third, because we asked therapists not to identify patients who were actively using or abusing

substances, the prevalence of impulsivity in this sample might be lower than what could be found in other samples of BPD patients. Fourth, it is possible that the interviewers did not differentially apply the scoring metrics between measures, thus leading to a considerably higher rate of diagnosis on the PDI–IV than the SCID–II. Although this is indeed a possibility, such interviewer behavior would be inconsistent with the systematic training on interview administration they received and thus would seem unlikely to happen so pervasively and systemically across interviewers.

Acknowledgments

Portions of this article were presented at the 2011 Annual Midwinter Meeting of the Society for Personality Assessment.

Funding

This study was funded by grants from the American Psychoanalytic Association and the International Psychoanalytic Association provided to Steven K. Huprich.

References

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

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

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

Beck, A. T., Epstein, N., Brown, G., & Steer, R. (1988). An inventory for measuring clinical anxiety. *Journal of Consulting and Clinical Psychology*, *56*, 893–897.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory II manual*. San Antonio, TX: Psychological Corporation.

Bell, M. D. (1995). *Bell Object Relations and Reality Testing Inventory–Form O*. Los Angeles, CA: Western Psychological Service.

Cooper, L. D., Balsis, S., & Zimmerman, M. (2010). Challenges associated with a polythetic diagnostic system: Criteria combinations in the personality disorders. *Journal of Abnormal Psychology*, *119*, 886–895.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO–FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

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

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

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). *Structured Clinical Interview for* DSM–IV *Axis II Personality Disorders Self-Report*. Washington, DC: American Psychiatric Association.

Hambleton, R. K., & Swaminathan, H. (1985). *Item-response theory: Principles and applications*. New York, NY: Guilford.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hyler, S. E. (1994). *Personality Diagnostic Questionnaire–4 (PDQ–4)*. New York, NY: New York State Psychiatric Institute.

Joyce, P. R., Light, K. J., Rowe, S. L., Cloninger, C. R., & Kennedy, M. A. (2010). Self-mutilation and suicide attempts: Relationships to bipolar disorder, borderline personality disorder, temperament, and character. *Australian and New Zealand Journal of Psychiatry*, *44*, 250–257.

Lobbestael, J., Leurgans, M., & Arntz, A. (2010). Inter-rater reliability of the Structured Clinical Interview for *DSM–IV* Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clinical Psychology and Psychotherapy*, *18*, 75–79.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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

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

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs*, *34* (Suppl. 4).

Samuel, D. B., Carroll, K. M., Rounsaville, B. J., & Ball, S. A. (2013). Personality disorders as maladaptive, extreme variants of normal personality: Borderline personality disorder and neuroticism in a substance using sample. *Journal of Personality Disorders*, *27*, 625–635.

Scientific Software International. (2011). *Item-response theory pro (IRT–PRO)* [version 2.1]. Skokie, IL: Author.

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of Personality Assessment*, *93*, 380–389.

Skodol, A. E., Oldham, J. M., Rosnick, L., Kellman, H. D., & Hyler, S. E. (1991). Diagnosis of *DSM–III–R* personality disorders: A comparison of two structured interviews. *International Journal of Methods in Psychiatric Research*, *1*, 13–26.

Skodol, A. E., Rosnick, L., Kellman, D., Oldham, J., & Hyler, S. (1988). Validating structured *DSM–III–R* personality disorder assessments with longitudinal data. *American Journal of Psychiatry*, *145*, 1297–1299.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Trull, T. J., Widiger, T. A., Lynam, D. R., & Costa, P. T., Jr. (2003). Borderline personality disorder from the perspective of general personality functioning. *Journal of Abnormal Psychology*, *112*, 193–202.

Widiger, T. A., & Boyd, S. (2009). Assessing personality disorders. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (3rd ed., pp. 336–363). New York, NY: Oxford University Press.

Widiger, T. A., Mangine, S., Corbitt, E. M., Ellis, C. G., & Thomas, G. V. (1995). *Personality Disorder Interview–IV*. Odessa, FL: Psychological Assessment Resources.

Yang, J., Dai, X., Yao, S., Cai, T., Gao, B., McCrae, R., & Costa, P. T. (2002). Personality disorders and the five-factor model of personality in Chinese psychiatric patients. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 215–222). Washington, DC: American Psychological Association.

Yang, J., McCrae, R. R., Costa, P. T., Yao, S., Dai, X., Cai, T., & Gao, B. (2000). The cross-cultural generalizability of Axis II constructs: An evaluation of two personality disorder assessment instruments in the People's Republic of China. *Journal of Personality Disorders*, *14*, 249–263.

Zanarini, M. C., Gunderson, J. G., Frankenburg, F. R., & Chauncey, D. L. (1989). The revised Diagnostic Interview for Borderlines: Discriminating BPD from other Axis II disorders. *Journal of Personality Disorders*, *3*, 10–18.